FINAL REPORT

Name of the Student	Armağan Dağıstan
Name of the Advisor	Assist. Prof. Dr. Irmak Sargın
Project Title	Prediction of Thermoelectric Material Properties with Machine Learning

Contents

1 INTRODUCTION

At the beginning of 19th century. Thomas J. Seeback noticed how a compass needle deviates when two junctions of distinct metals were kept at different temperatures. This observation led that an electric current in a closed circuit can be driven by the electric potential created by the temperature difference. This is now referred as the Seeback effect. This phenomena was the first discovery of thermoelectric effect and thermoelectric materials. In 1834, Jean C. A. Peltier discovered that an electrical current could produce a temperature difference at a junction in isothermal conditions. In 1838, Lenz demonstrated that heat could be removed or generated in the junction depending on the direction of the current flow. The electrical current is directly proportional to the amount of heat generated or absorbed at the junction. It is called Peltier coefficient. In 1851, William Thomson harmonized the laws of thermodynamics with the thermoelectric. He demonstrated the necessity of a third effect by using a theoretical examination of the link between Thomson effect and Peltier effect. This third effect is the creation or absorption along a current-carrying conductor under a thermal gradient [1], [2], [3].

In the 20th century two theoretical papers by E. Altenkirch regarding the effectiveness of thermopile published. The first proof that a suitable TE material should have a big Seeback Coefficient, low thermal conductivity, and high electrical conductivity was provided by Altenkirch[4].

In this project our goal is to identify and design low thermal conductivity compounds for thermoelectric materials using maching learning. We are focusing on pre-published datas to uncover the most promising materials. Our main question is whether we can use this data and find the effect of different descriptors such as, Ionization energy,ıonic radius, atomic radius, lattice constants, electronegativity, electrical conductivity, to figure out which materials are best at turning heat into electricity. By doing this, we aim to reveal patterns that distinguish the top performing materials.

With a particular focus on thermoelectric figure of merit (ZT) this research aims to overcome the difficulties involved in the creation of new and improved thermoelectric materials by using a data-driven approach. ZT is an important measure that defines a materials thermoelectric performance. It depends on the Seebeck coefficient (S), electrical conductivity (σ), thermal conductivity (κ), and temperature (T) according to the equation below.

$$ZT = \frac{S^2\sigma}{\kappa} T \qquad 1$$

Improvements of thermoelectric materials are necessary because of their uses in many areas, which includes waste heat recovery, solar power generation, and compression-free refrigeration[5]. Although they have advanced, traditional experimental efforts have not found many novel materials with exceptionally high thermoelectric performance. This is mainly because there are not many effective chemical design techniques and it is challenging to optimize thermal transport features. [5] In the improvements of TE materials minimizing k while simultaneously keeping seeback coefficient and electrical conductivity high is significant issue. A few notable findings about discovery of thermoelectric materials are as follows: Venkotosubraomanian et al. suggested that p-type Bi2Te3/Sb2Te3 superlattices could have the highest ZT of roughly 2.4 at 300K [6],[7]. Hsu et al. [6],[8] demonstrated that the material system AgPbmSbTe2+m[Lastm) wit m=10 and 18 suitably doped could show a high ZT of 2.2 at 800K. The large decrease in heat conductivity in these complicated or nanostructured chalcogenides had a major role in acomplishmetn of high ZT. Cobaltitis [6],[9],[10],[11],[12] are attracting more attention recently because of their thermoelectric qualities. In figure 1, common instances of cobalties are displayed. These materials share the following characteristics in common: (i) large seeback coefficient. (ii) low thermal conductivity due to their layered structure.

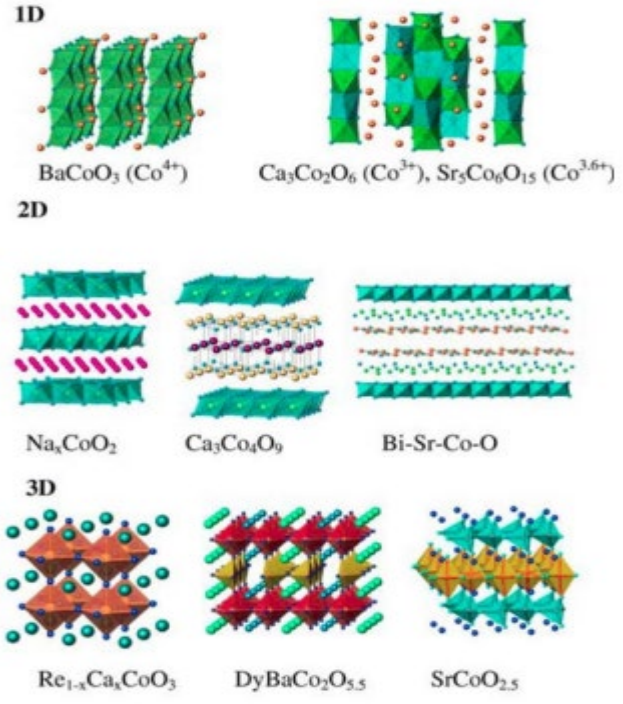


Figure 1. Typical examples of 1-D, 2-D, and 3-D cobaltites

This project's data-driven methodology depends on the idea that cast chemical spaces, made up of thousands of chemicals may be effectively ranked uncomplicated and accurate metrics. The idea of predicting material properties with machine learning has

been presented in earlier research. For example, to optimize a reduced mechanism of 2-butanone, Wong et al. developed a self-adoptive differential evolution method [13],[14]. Peng et al. examined the influence of different parameters on solar evaporation using the random forest technique [13],[15]. Zhang et al. provide a summary of machine learning techniques for determining a compound, composite, or alloy's thermal conductivity.[13],[16]. Wan et al. examined the state of research on the discovery of new materials and the use of materials informatics to predict their properties in heat transport [13],[17]. These techniques have shown that benefits over current ones by improving accuracy and capturing detailed non-linear interactions [13].

In this project we started with the data collection from the TE design Lab that provides tools for examining unprocessed experimental thermoelectric data and adding them to the publicly accessible database. The datas we have collected from this database are as follows; Seeback coeffiecients, Thermal conductivity, electrical conductivity, and temperature values of these compounds. These data are consisting of both experimental and computational calculations. In addition to this we also gathered information about ionization energies of elements from the atomic spectra database and atomic radiis of elements, we wrote different python codes to calculate the compounds weighted average values of valence electrons, ionization energies, covalent radiuses, atomic radiis, lattice parameters, electronegativity, electrical conductivity, thermal conductivity. To do this, first we have collected elemental values of these data then using python we calculated corresponding weighted average value for each compound. There are about 5000 compound available in our data. From this large database we have begun a careful data analysis process. Our investigation focused on finding relationships and correlations between the key material features in these TE materials. Finally, we used random forest regression to reveal relations between our inputs (valence electrons, ionization energies, covalent radii, atomic radii, lattice parameters, electronegativity, and temperature of each compound. For our targets, we used thermal conductivity, seeback coefficient, electrical conductivity, and the figure of merit (ZT).

2 METHOD

An extensive review of the literature had carried as a part of the first workpackage of our project. Furthermore, this assessment had included and examination of the relevant implementation of machine learning techniques as essential elements to be utilized in the future phases of the project.

Additionally, descriptors also determined for the machine learning model. Significant contributions are made by descriptors to the creation of the reliable predictions about the characteristic of different compounds [18]. Feature representation is used to characterize the material attributes before models are optimized. For these reasons, we have determined descriptors for the materials such as, ionization energy, atomic radius, lattice parameters, ionic radius. For the ionization energies we used the database of NİST (Atomic spectra database). To do this, for every element, valance, first second and third ionization energy values taken as a list one by one. After that, we wrote a python code for the compounds listed in the database of TE design lab and generated a function that extract elements from this compound list. The code initializes empty lists “elements” and “weights” to store the extracted elements and their weights. Then it uses re.findall to match for chemical elements and their weights in each compound. In addition, the weighted average ionization energies of each compound are calculated. This process is also repeated for the atomic radii information of all compounds in our database. In addition, similarliy we calculated the weighted average distribution for ionic radius, lattice parameters, electronegativity, thermal conductivity.

In preprocession phase of our project we used standardization and normalization of data. Standardization and normalization are teqniques that used in machine learning to scale and transform features before feeding them into the model. Using this techniques we ensured that the features are in the similar scale. Also we have used the box plots to compare our features distributions. When dealing with multiple features, box plots enable the comparison of their distributions. This is particularly useful for assessing the relative importance of different features and identifying features that may require further investigation or transformation.

For these thermoelectric materials we have studied how different features are connected by looking at how a material’s structure relates its electrical conductivity, seeback coefficient, thermal conductivity and ZT. After that we divided ZT values of each compound by its corresponding experimental temperature values using python. We have used statistics and visuals to understand these relationships. For the correlation analysis we have used Spearman and Pearson correlation values for between Z and valance ionization, first ionization, second ionization, third ionization energies, atomic radii,

lattice parameters, electronegativity, electrical conductivity, thermal conductivity.

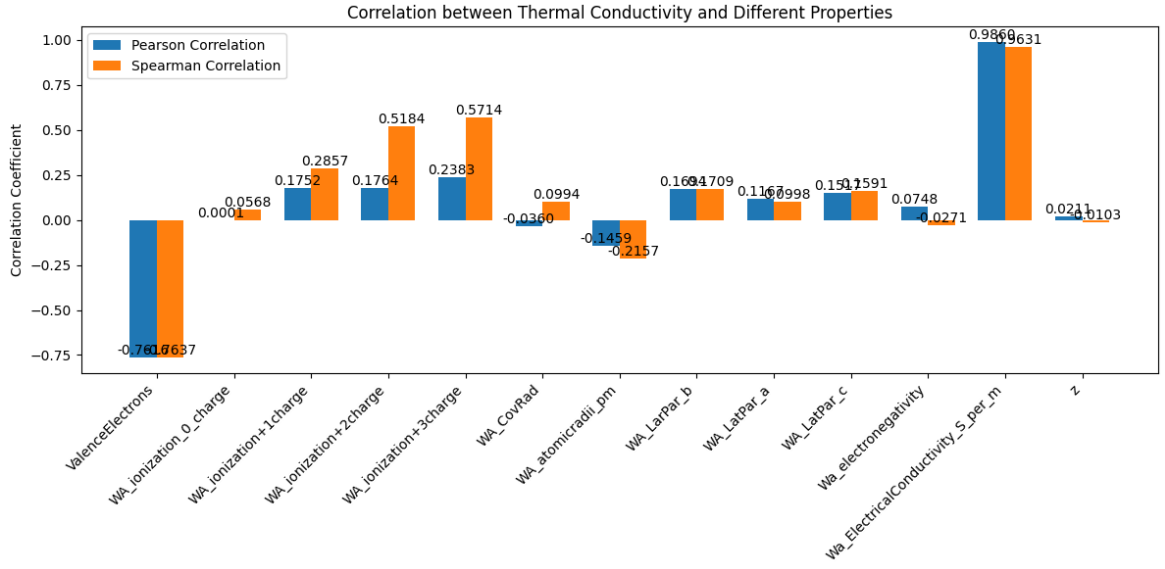
As the final phase of this project, to model complex relations between input features and target thermoelectric properties we used decision tree-based approach.

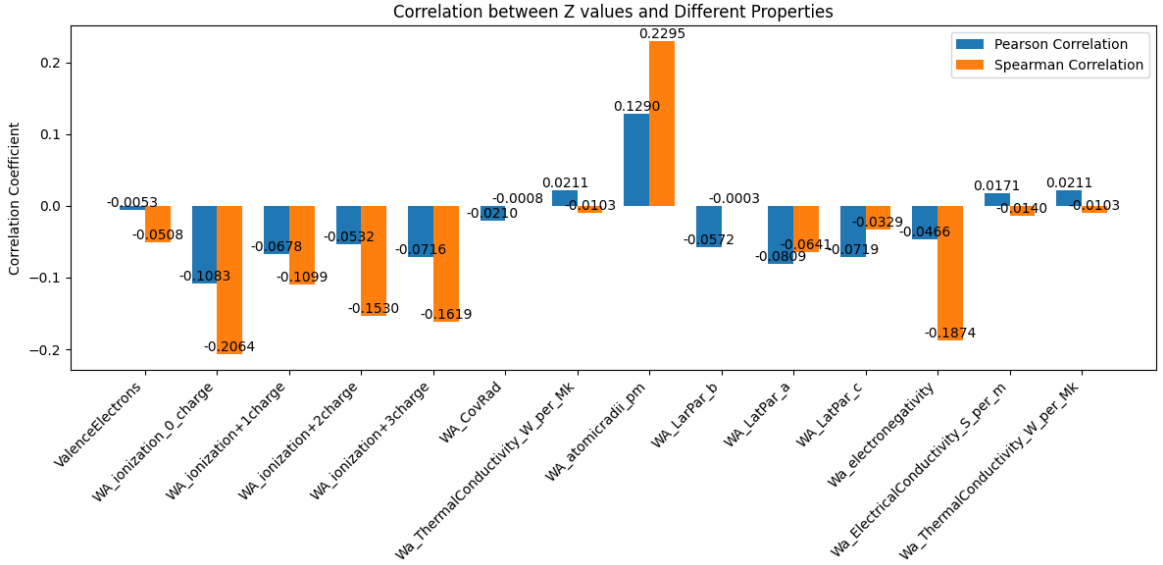
Decision tree approach works by iteratively partitioning the training samples into branches so that information gained at each split is maximized. This approach enables model to naturally capture the non-linear relations in the dataset.

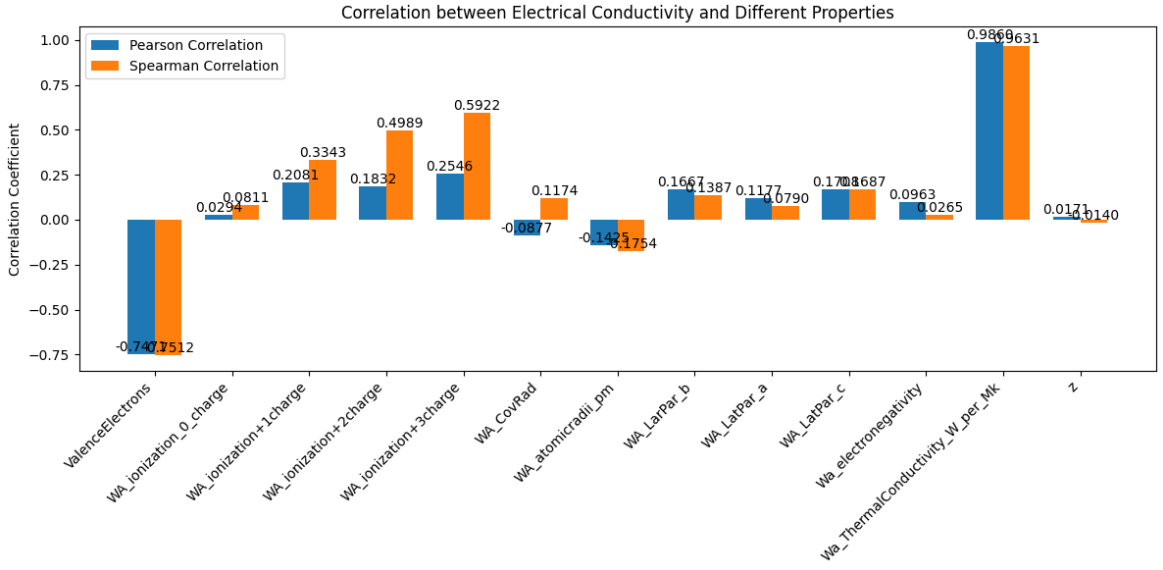
To increase the performance and to decrease the overfitting, the random forest ensemble method is used. In this project, the input features encompassed crucial parameters such as, valance electrons, ionization energies, covalent radii, atomic radii, lattice parameters, electronegativity of compounds. These features are the bases of our training model. For our target data, our model aimed to predict thermal conductivity, seeback coefficient, electrical conductivity and the dimensionless figure of merit(ZT). These target data consist of both experimental and computational data collected from Te design lab. In addition, we divided our data as 80% training set and 20% test set.

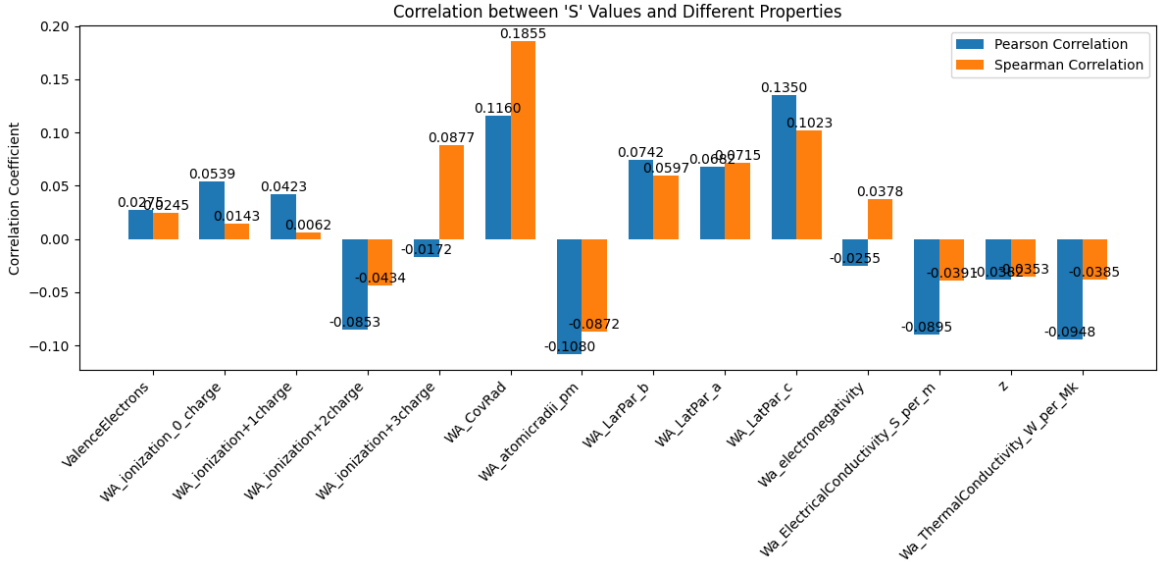
3 RESULTS AND DISCUSSION

As our first result to decide which machine learning model to use in this project we looked for the correlations between Thermal conductivity, Z, electrical conductivity and seeback coefficient and the other values that are calculated as elementally.



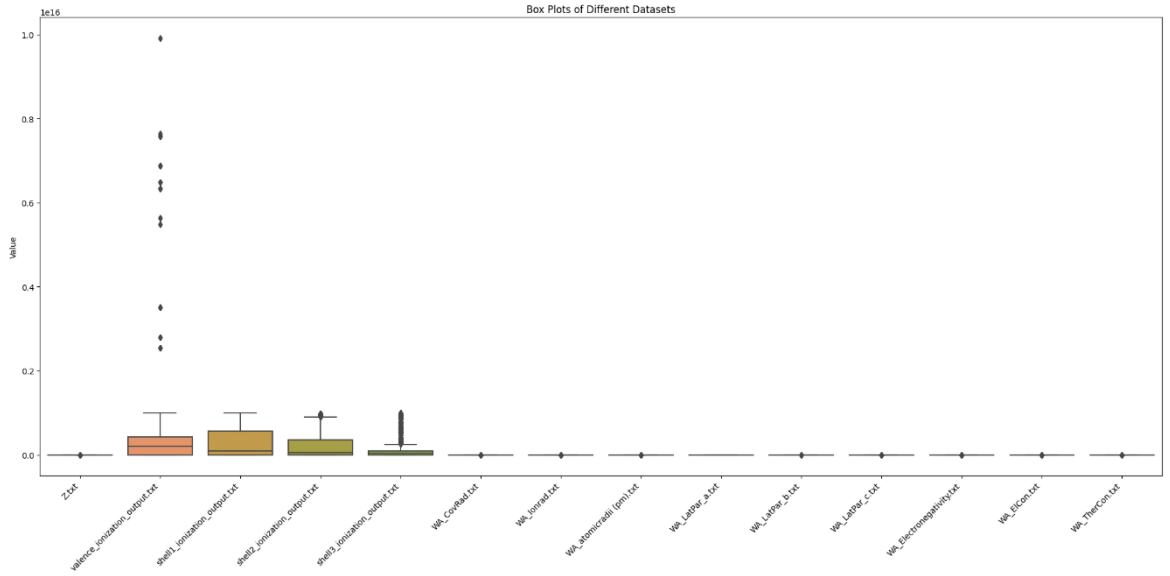




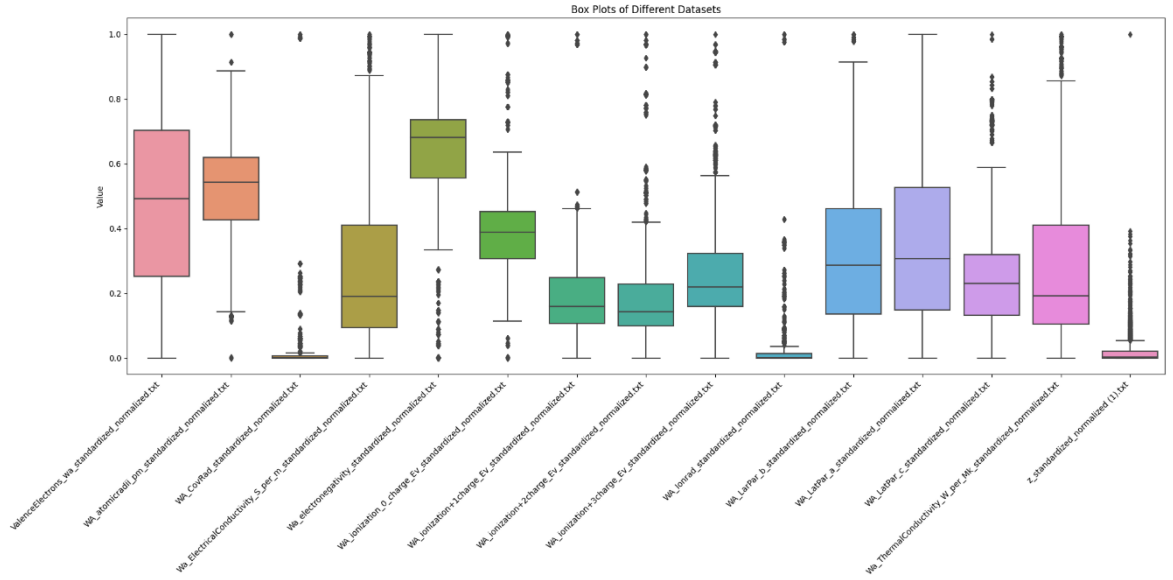


Based on these correlation values we decided to use random forest algorithm. The decision to use Random Forest was driven by its ability to adapt when handling non-linear interactions, its capacity to shed light on feature importance through ensemble averaging and mitigate outliers. The algorithm's ability to manage non-linearity was especially advantageous for characteristics with low correlations.

In the preprocessing phase of our data we also used box plots to see whether we can use this data directly to feed our model or we can should we improve our data.



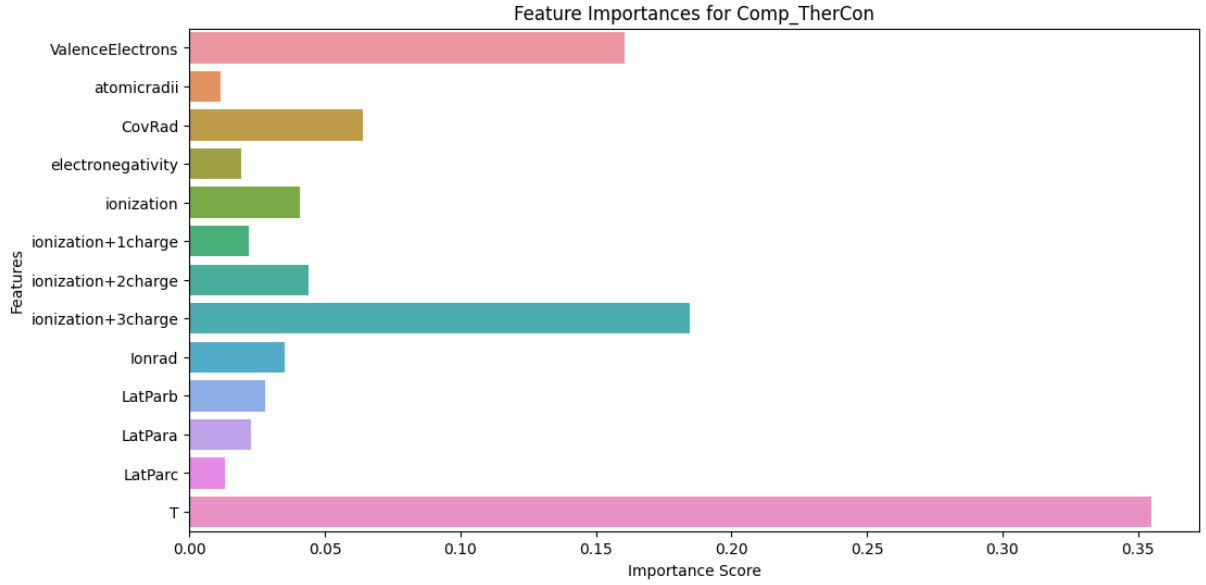
Before using normalization and standardization of our data the distribution of our data can bee seen in this plot. The ranges were too wide and it was hard to train our data with this raw data. So we used standardization and normalization and the distribution of our data can be seen in the second box plot :



In the final phase of our project we build four different model with random forest regression. These are prediction of thermal conductivity, prediction of electrical conductivity, prediction of seeback and prediction of ZT using our elemental values as inputs. The results are as follows:

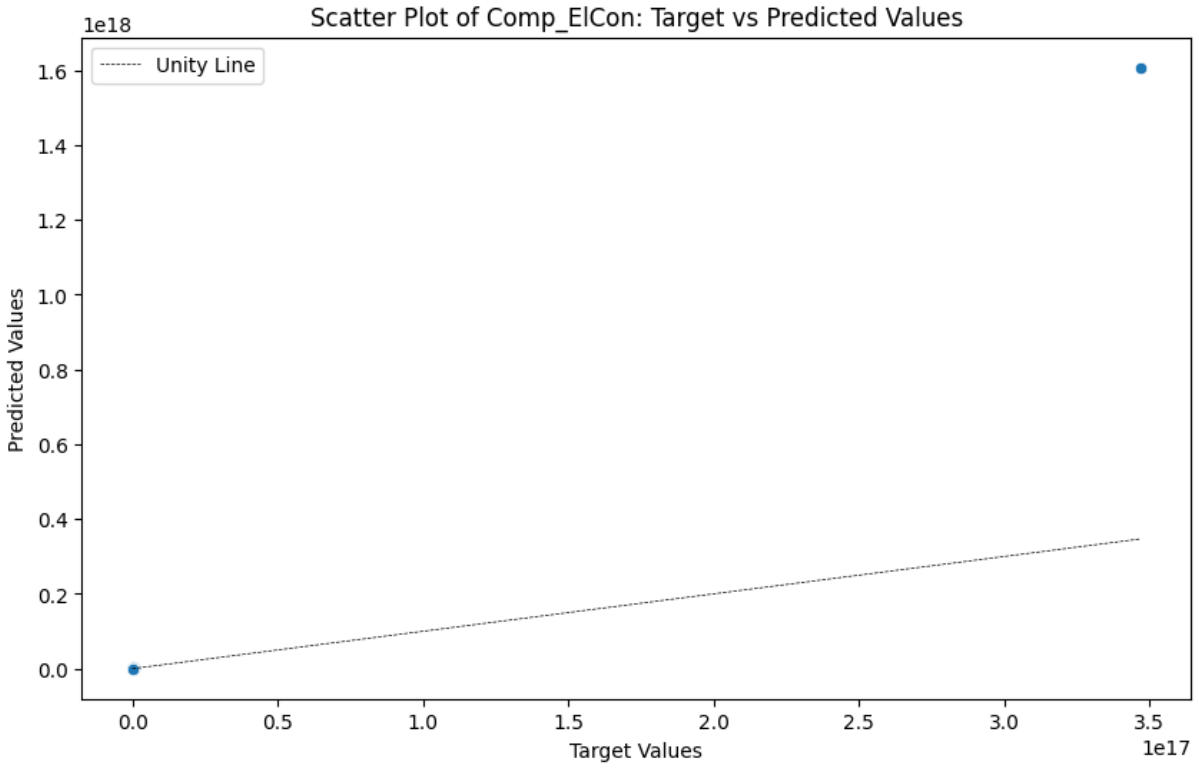
Model for thermal conductivity prediction

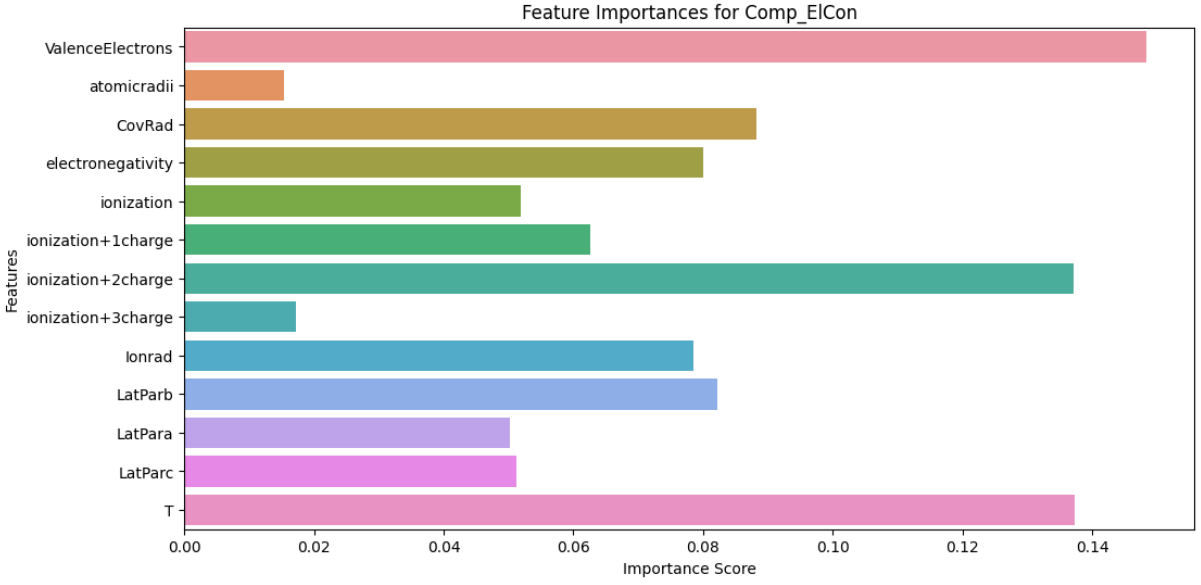
Mean Absolute Error: 45.13411814199221
Target Mean: 41.32686531495059
MAE as Percentage of Target Mean: 109.21%



Model for electrical conductivity prediction

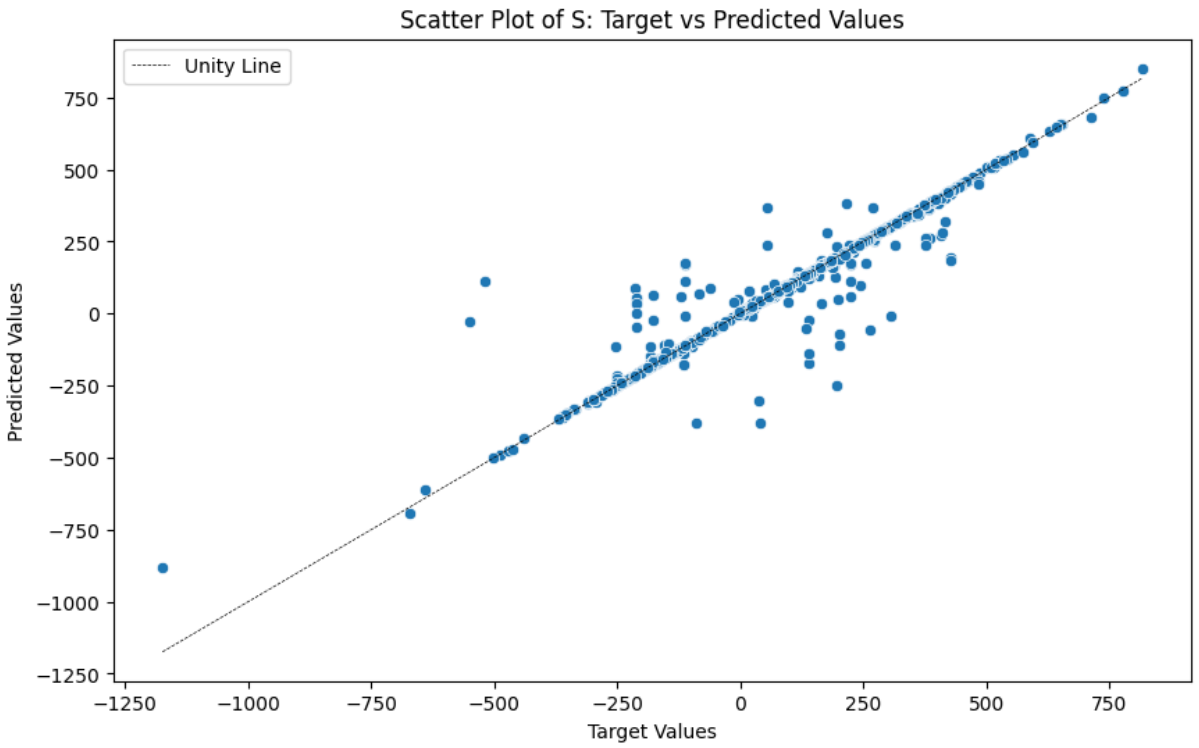
Mean Absolute Error: 1354711356359159.8
Target Mean: 586228037979073.0
MAE as Percentage of Target Mean: 231.09%

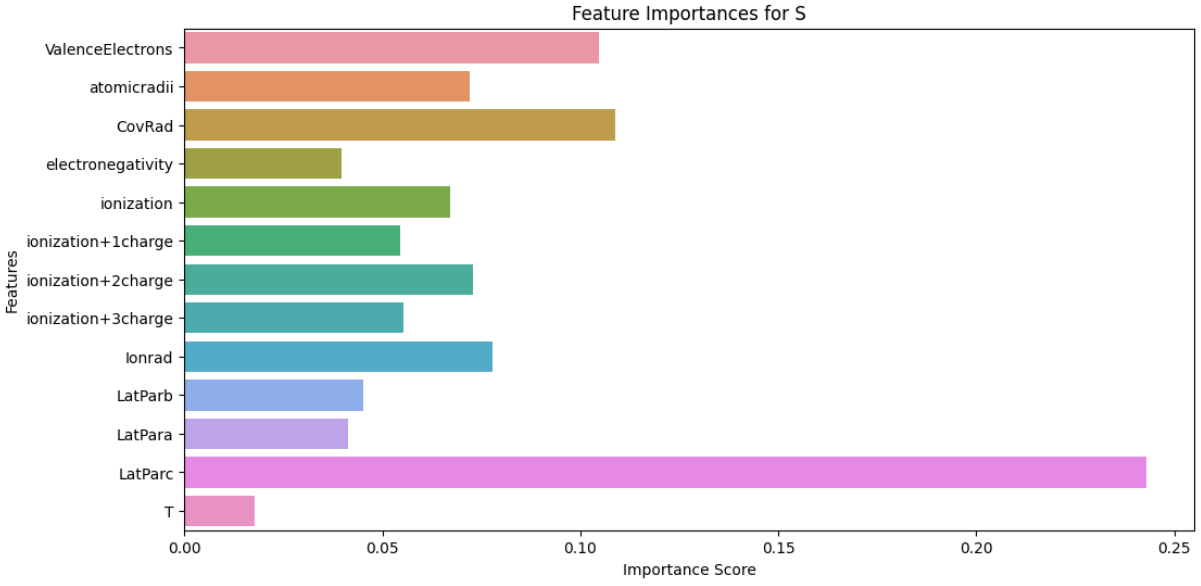




Model for seeback prediction

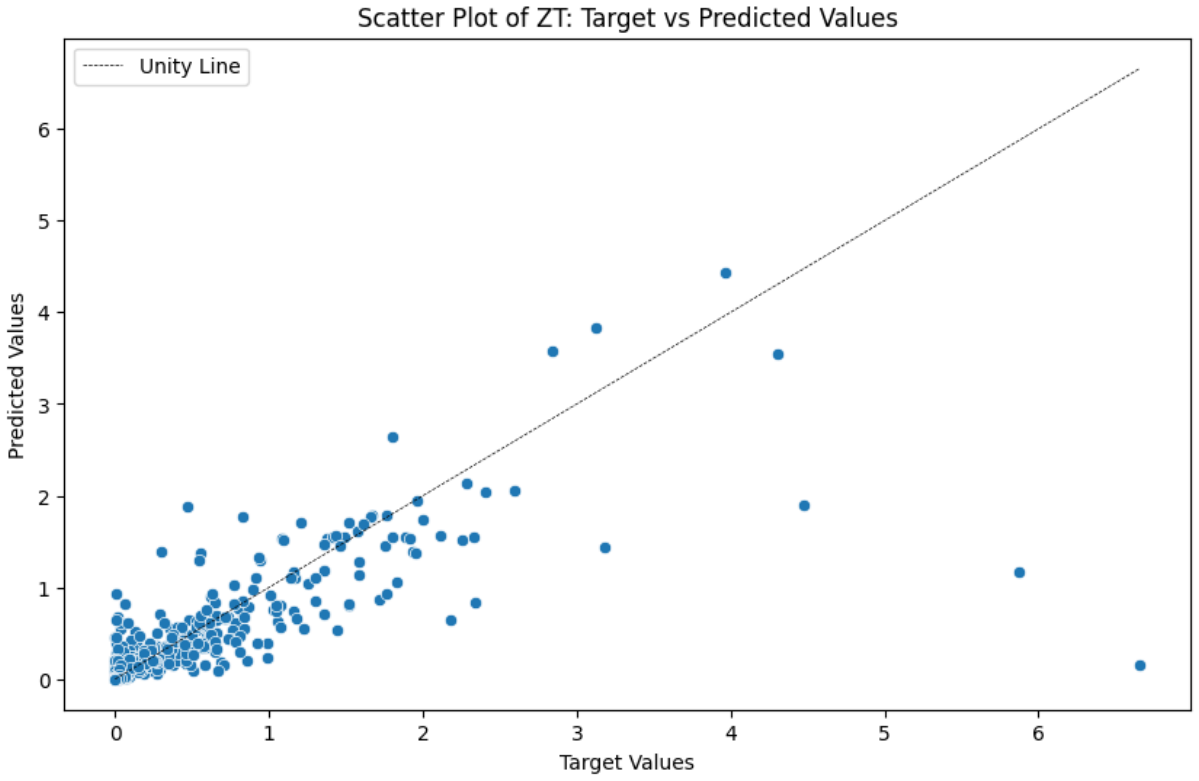
Mean Absolute Error: 16.478286370584257
Target Mean: 74.48528342245989
MAE as Percentage of Target Mean: 22.12%

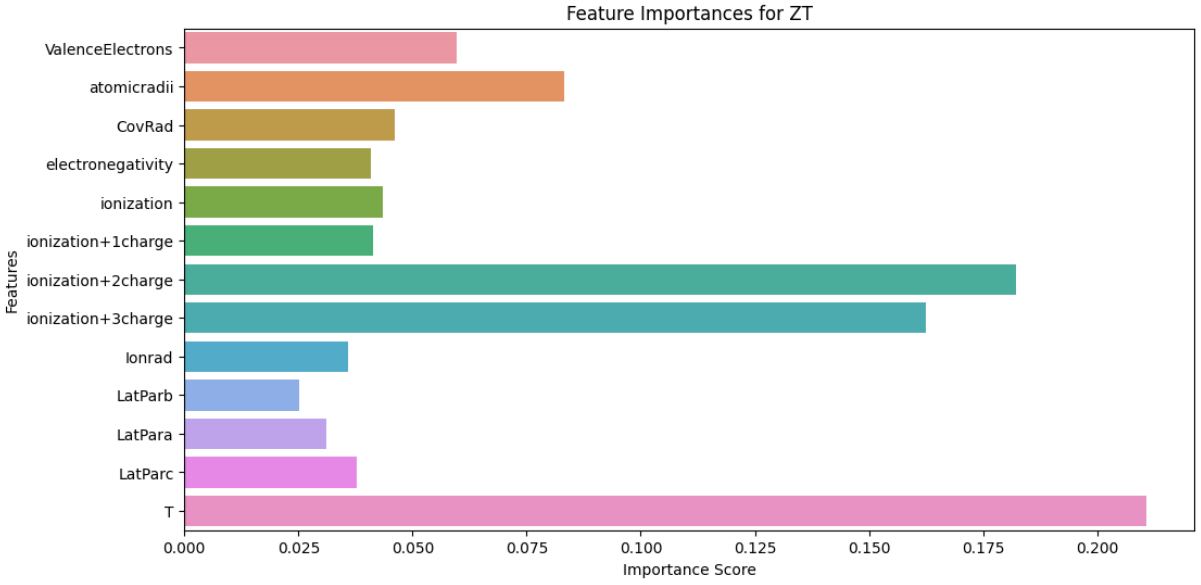




Model for ZT

Mean Absolute Error: 0.10370836495930205
Target Mean: 0.21969207939455226
MAE as Percentage of Target Mean: 47.21%





While evaluating four random forest machine learning model, different performances can be observed. We looked at mean absolute errors od the target data. The errors are all calculated for test values of our machine learning model.

Thermal conductivity model shows a 109.21% mean absolute error of the target value. The reason for this high percentage error might be the potentially complex nature of thermal conductivity of compounds.

In the electrical conductivity model, a notably higher MAE can be seen as a percentage of the target mean with 231.09%. This error also shows that there is a challenge in accurately predicting electrical conductivity from our input data.

In the prediction of ZT model, achives a relatively modest MAE as a percentage of the target mean at 47.21%. This result suggest a reasonable level of accuracy in predicting ZT.

Our most promising model is the prediction of Seeback model with MAE representing 22.12% of the target mean. This lower error percentage indicates a higher precision and proves that it can be used in predicting seeback coefficients for different compounds.

4 CONCLUSİON

To sum up, our project's goal was to find and create thermoelectric materials with low thermal conductivity by applying a data-driven strategy and machine learning approaches. Our aim was to find patterns that differentiate high-performing materials by utilizing a variety of descriptors, including electronegativity, atomic radius, ionization energy, lattice constants, and electrical conductivity, and by concentrating on pre-published data.

Key attributes of materials, such as temperature, thermal conductivity, electrical conductivity, and Seebeck coefficients, were calculated computationally and experimentally in the first exploration, which collected data from the TE design Lab. Furthermore, data about input values are collected and calculated elementally using python. To guarantee the effectiveness of the model, a comprehensive preprocessing step comprised box plot analysis, standardization, and normalizing.

We utilized the random forest regression approach after correlation analysis and it showed complex correlations between thermoelectric characteristics and material variables. Its capacity to manage non-linear interactions, highlight feature significance, and reduce outliers supported this decision. There were variable degrees of success in predicting ZT, thermal conductivity, electrical conductivity, and Seebeck coefficients from the data of four distinct models; the Seebeck prediction model performed the best. The models' mean absolute errors were used to evaluate them, underscoring the difficulties in precisely forecasting some material attributes.

For future outlook of this project, other machine learning models can be investigated

using our input data. In addition, data set can be improved using other thermoelectric data sources. Enhancing the feature set with new descriptors related to defects, electrical band structure, or crystal structure might provide a more thorough knowledge of material attributes. The training set may be further diversified by adding data from experimental studies and other thermoelectric sources to the dataset. This extension might enable the creation of more reliable machine learning models and capture a wider variety of material behaviors. Advanced machine learning approaches, such ensemble methods or deep learning architectures, might open up new modeling opportunities for thermoelectric material complicated interactions.

5 REFERENCES

[1] Th. J. Seebeck "Magnetische Polarisation der Metalle und Erze Durch Temperatur-Differenz"1822-23 in Ostwald's Klassiker der Exakten Wissenshaften Nr. 70 (1895). Seebeck Biography 1. Seebeck Biography 2.

[2] G. Magnus, Poggendorf's Annalen der Physik 83 p469 (1851)

[3] E. Becquerel, Ann. de chim. et phys. (4) 8. (1866)

[4] W. Thomson "On the Dynamical Theory of Heat. Trans." R. Soc. Edinburgh: Earth Sci. 3

[5] TE Design Lab: A virtual laboratory for thermoelectric material design, Prashun Gorai et al., Colorado School of Mines, National Renewable Energy Laboratory, University of Colorado Boulder, Northwestern University, 2015.

[6] J.C.Zheng,"Recent advances on thermoelectric materials" Front. Phys. China,vol.3,no.3,pp.269−279,2008.

[7] R. Venkatasubramanian, E. Siivola, T. Colpitts, and B. O'Quinn, Nature, 2001, 413: 597

[8] K. F. Hsu, S. Loo, F. Guo, W. Chen, J. S. Dyck, C. Uher, T. Hogan, E. K. Polychroniadis, M. G. Kanatzidis, Science, 2004, 303: 816.

[9] J. Hejtmánek, M. Veverka, K. Knížek, H. Fujishiro, S. Hebert, Y. Klein, A. Maignan, C. Bellouard and B. Lenoir, Mater. Res. Soc. Symp. Proc. Vol. 886, 2006 # 0886-F01- 07.1

[10] (a) Y. F. Hu, W. D. Si, E. Sutter, Q. Li, Appl. Phys. Lett., 2005, 86: 082103; (b)Y. F. Hu, E. Sutter, W. D. Si, Q.Li, Appl. Phys. Lett., 2005, 87: 171912.

[11] (a) P. Oleynikov, L. Wu, J.C. Zheng, V.V. Volkov, R.F. Klie, Y. Zhu, H. Inada, K. Nakamura, R. Twestern, “Structural analysis of layered Ca3Co4O9 thermoelectrics using aberration corrected STEM and EELS”, Advanced Electron Microscopy in Materials Physics Workshop, November 7-8, 2007, Brookhaven National Laboratory, USA; (b). P. Oleynikov, J. Hanson, J.C. Zheng, L. Wu, V. Volkov, Q. Jie, Q. Li, and Y. Zhu, “Electron Microscopy Study of Layered thermoelectric Cobalt Oxide [Ca2CoO3]0.62CoO2”, Workshop of “Electronic structure and functionality of thermoelectric materials”, 30 July – 1 August 2007, Reykjavik, Iceland.

[12] Y. Y. Wang, N. S. Rogado, R. J. Cava, and N. P. Ong, Nature, 2003, 423: 425.

[13] Machine Learning Prediction for Bandgaps of Inorganic Materials, Lang Wu et al., 2020.

[14] Y. Wang, S. Liu, J. Cheng, X. Xiao, W. Feng, N. Yang and C. Zuo, ES Mater. Manuf., 2019, 6, 28-37. doi: 10.30919/esmm5f615.

[15] G. Peng, S.W. Sharshir, Y. Wang, M. An, A.E. Kabeel, J. Zang, L. Zhang and N. Yang, arXiv preprint arXiv:1906.08461, 2019

[16] H. Zhang, K. Hippalgaonkar, T. Buonassisi, O.M. Løvvik, E. Sagvolden and D. Ding, arXiv preprint arXiv:1901.05801, 2019.

[17] X. Wan, W. Feng, Y. Wang, H. Wang, X. Zhang, C. Deng and N. Yang, Nano Lett., 2019, 19(6), 3387-3395. doi: 10.1021/acs.nanolett.8b05196.

[18]T.Wang,C.Zhang,H.Snoussi,andG.Zhang,"MachineLearningApproachesforThermoelectricMaterialsResearch,"Adv. Funct. Mater., vol.29, no.33,p.1906041,2019.