HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Predicting Compound Melting Temperature from Computationally Derived Properties via Machine Learning

Amy Lin[1] and Mark Eastburn[#]

[1]Princeton High School, USA
[#]Advisor

ABSTRACT

Melting temperature is a fundamental material property used in a wide variety of scientific disciplines. To eliminate the need for experimentally measuring melting temperature, often costly and potentially dangerous, we developed a new computational approach that predicts a compound's melting temperature from its material properties, computed by Density-functional theory or other computational methods, using machine learning models. The proposed machine learning models achieved promising results, with a Mean Percentage Absolute Error of 15%-20%. We also identified important material properties in predicting melting temperature across various groups of compounds containing metals, transition metals, post transition metals, alkaline earth metals, halogens, and chalcogens.

Introduction

Melting temperature is a fundamental material property used in a wide variety of scientific disciplines. Structural materials used in machines, buildings, heat shields, and more must have a sufficiently high melting temperature to ensure safe operation, while molten salt used in thermal energy storage needs a sufficiently low melting temperature to be useful. Measuring the melting temperature is an important task in material science. Compounds with extreme melting temperatures are time-consuming to measure experimentally [1] and such experiments can be dangerous. Although simulations exist for calculating melting temperatures, they are computationally expensive. Only 10% of around 200,000 known inorganic compounds have determined melting temperatures. Efficient methods for determining melting temperature are an active research topic in material science.

Research work has been done in developing effective methods to determine compound melting temperatures. Existing approaches include using thermodynamic calculations and using empirical calculations represented by the CALculation of PHAse Diagrams (CALPHAD) method [3]. These calculations are extremely time-consuming. Hong et al. proposed a graph neural network to predict a compound's melting temperature using its chemical formula [1]. However, when new materials are discovered, the material properties in their approach have to be determined through potentially costly and dangerous experiments. We are interested in using computationally determined material properties to predict a compound's melting temperature. With a computational approach, when new materials are discovered, their properties can be easily determined through computational models, thus avoiding expensive experiments.

Recently, researchers have turned to ab-nitio material databases that host a large amount of information about the properties of chemical compounds [4, 5, 6]. The Materials Project is one such database. Most of the material properties in the database are calculated based on Density-functional theory (DFT) [7]. DFT uses an arrangement of atoms, i.e. the atomic structure, and computes the material's zero-kelvin energetic stability [2].

From this relationship, DFT then deduces other material properties. However, because DFT calculations predict properties at zero-Kelvin temperature, it cannot be used to predict temperature-dependent properties such as melting temperature.

In this work, we propose to use machine learning models to predict compound melting temperatures from their DFT-derived properties and other computationally determined properties. We further propose to identify the material properties that are primary determinants of melting temperature by analyzing the feature importance presented in the machine learning models.

Methodology

This work uses a computational approach to determine the melting temperature of various compounds. Specifically, we build machine learning models to estimate the melting temperature from DFT-derived and computationally determined material properties. We use datasets published by the Materials Project and Proceedings of the National Academy of Sciences (PNAS) [1], which provide material properties and the compounds' melting temperatures. We then train Random Forest Regression models, a type of machine learning model, to predict melting temperature from material properties. We evaluate the prediction accuracy of these machine learning models and analyze the importance of different material properties in computationally determining melting temperature.

Data Acquisition

Three data sources are used in this work. The first database is published on PNAS [1]. It provides the chemical compounds, their chemical space groups, and their melting temperatures. The second database is the Materials Project database [7]. It provides a large number of DFT-derived compound properties. The third database comes from Matminer [9], where element-based properties are provided [10]. The information from the three databases is used to create our curated dataset, which contains 1408 compounds in total.

Materials Project Database

The Materials Project is an open web-based database of computed information on known and predicted materials [7]. The Materials Project REST API allows a user to retrieve information from its database [8]. In this work, we use the Materials Project Legacy API to retrieve DFT-derived properties.

The Python Materials Genomics (Pymatgen) package [8] is integrated with the Materials Project REST API. The Materials Project ID (MPID) is the unique identification label for a chemical compound. We used Pymatgen to find the MPID for each compound based on their space group and chemical formula. Compounds whose MPID's were not found in the Materials Project database were excluded from our curated dataset.

Compound Property Retrieval

We define compound properties as the physical and chemical characteristics of a chemical compound. For each compound, we consider the following 5 DFT-calculated compound properties: Voigt-Reuss-Hill (VRH) average bulk modulus, VRH shear modulus, cohesive energy, formation energy, and density. In the following discussions, these compound properties are denoted as k_vrh, g_vrh, e_coh, e_form, and density, respectively. We considered these properties because from general chemistry knowledge, except for density, they are the most influential properties in determining a compound's melting temperature. For a small number of chemical compounds, the values for elastic moduli are not found in the Materials Project database. In those cases, machine learning-predicted elastic moduli values are provided in the database, and our curated dataset uses those values as a proxy for elastic moduli.

Element-Based Property Retrieval with Matminer

We define element-based properties as the statistical parameters of elemental properties. Specifically, the following 17 element-based properties are considered in this work: mean atomic weight, mean column on periodic table, mean row on periodic table, mean atomic number, mean atomic radius, mean electronegativity of elements, average s-suborbital valence electrons, average p-suborbital valence electrons, average d-suborbital valence electrons, average f-suborbital valence electrons, range of atomic number, range of atomic radius, range of electronegativity of elements, fraction of s-suborbital valence electrons, fraction of p-suborbital valence electrons, fraction of d-suborbital valence electrons, and fraction of f-suborbital valence electrons.

For notational convenience, these properties are denoted as follows in presenting the results and figures: mean AtomicWeight, mean Column, mean Row, mean Number, mean AtomicRadius, mean Electronegativity, avg s valence electrons, avg p valence electrons, avg d valence electrons, avg f valence electrons, range Number, range AtomicRadius, range Electronegativity, frac s valence electrons, frac p valence electrons, frac d valence electrons, and frac f valence electrons.

Figure 1 shows the scatterplots of the material properties (VRH bulk modulus, VRH shear modulus, density, mean atomic weight of atoms, mean atomic radius, and average f-suborbital valence electrons) and the corresponding melting temperatures. Each point represents a compound in the scatterplots of property-melting temperature. These features do not appear to have a clear linear relationship with melting temperature. It is not straightforward to find a prediction function with a good fit.

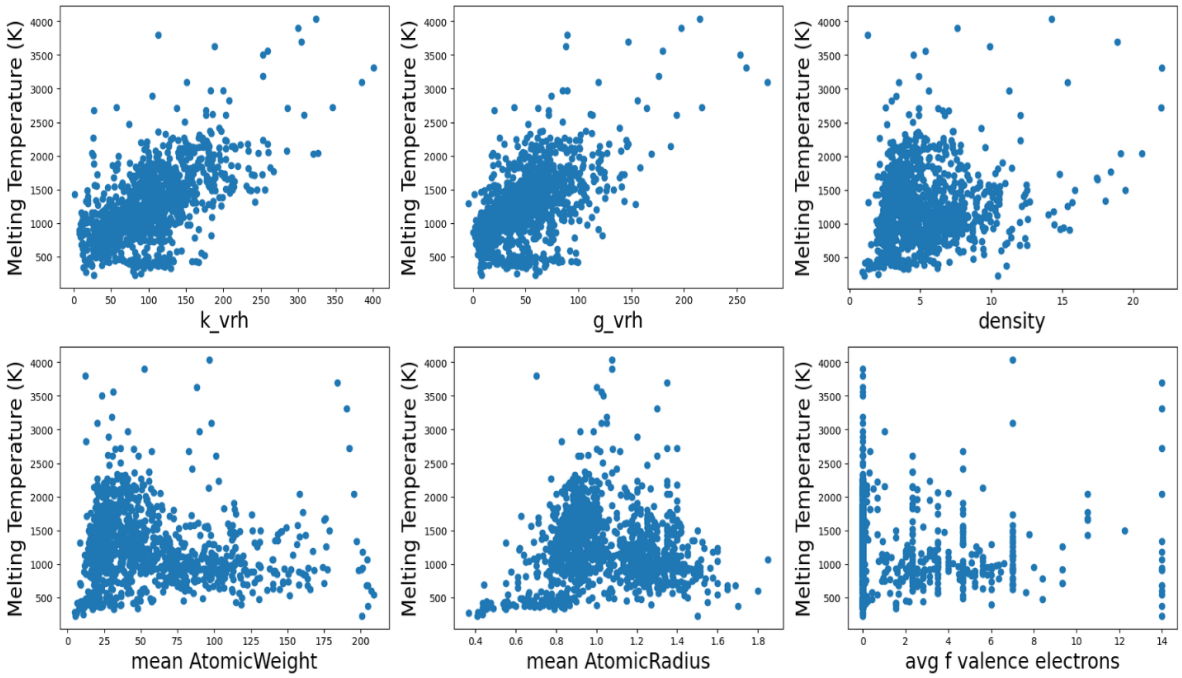

Figure 1. The relationship between some of the material properties (VRH bulk modulus, VRH shear modulus, density, mean atomic weight of atoms, mean atomic radius, and average f-suborbital valence electrons) and the corresponding melting temperatures.

Data Preparation

Data Augmentation

The material properties listed above are used as input features to machine learning models. In addition to the 5 DFT-derived compound properties and 17 element-based properties, we augment the set of properties by generating pairwise products of these property values. For example, for k_vrh and density, we compute the product of k_vrh x density as a new feature, denoted as k_vrh_x_density. These pairwise products are extended features augmented to our dataset. The purpose of feature augmentation is to study if jointly using material properties can improve the accuracy of machine learning models in predicting melting temperature. From 5 compound properties and 17 element-based properties, we generate 231 new features. In total, we use 253 features to train the machine learning models.

Compound Grouping

Our dataset contains 1408 samples. In addition to using one machine learning model to predict melting temperatures of all compounds, we are also interested in using separate models for different compound groups. We divide the compounds into 6 groups based on the types of elements contained in the compound: metal, transition metal, halogen, chalcogen, alkaline earth metal, and post transition metal. The number of compounds in each group is shown in Table 1. In the Results section, these groups are denoted as metal, transition_metal, halogen, chalcogen, alkaline, and post_transition_metal in the figures presented.

Table 1. Number of compounds per compound group.

Compound Group	Number of Compounds (Samples)
All compounds	1408
Metal	1159
Transition metal	693
Halogen	212
Chalcogen	982
Alkaline earth metal	308
Post transition metal	358

Data Analysis

We built various Random Forest Regression (RFR) models to determine a compound's melting temperature based on its properties and extended features. We also used the feature importance in the RFR model to gain insight into the features most important in achieving the best prediction accuracy.

Random Forest Regression (RFR)

RFR is an ensemble method for regression. It consists of multiple decision trees. Each decision tree is learned to provide a prediction based on a set of input features. An illustrative example of a decision tree is shown in Figure 2.

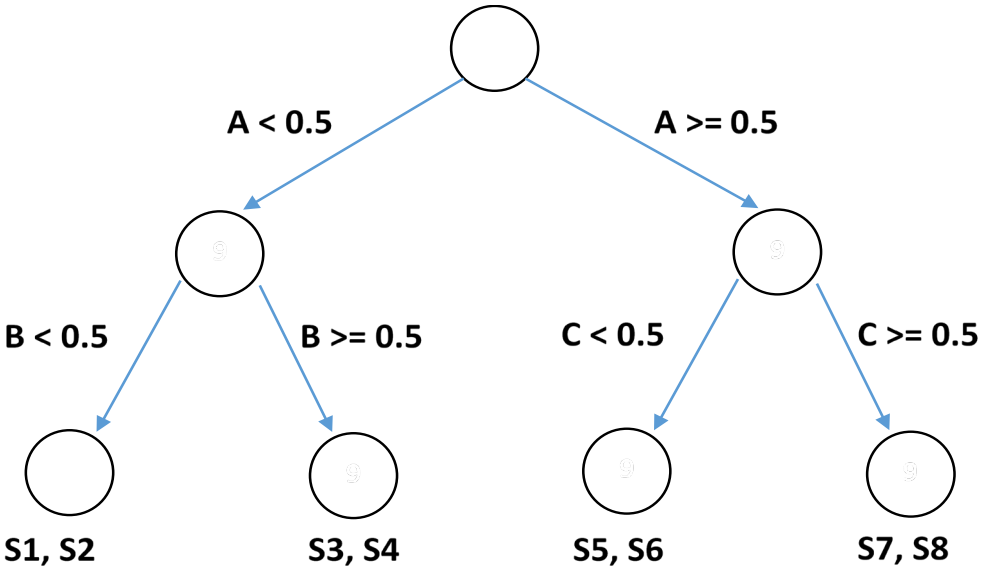

Figure 2. Illustrative example of a decision tree.

There are 8 compounds (S1, S2, ..., S8), each with 3 properties (A, B, C). The decision tree has a depth of 2. At the root node (level 0 of the tree), the data is split based on the value of property A. If A < 0.5, then the compound goes to the left node. If A >= 0.5, then the compound goes to the right node. On level 1 of the tree, compounds at the left node are further split into 2 subsets based on the value of property B. Compounds at the right node are split into 2 subsets based on the value of property C. At the leaf nodes (level 2 of the tree), compounds are split into 4 subsets: {S1, S2}, {S3, S4}, {S5, S6}, and {S7, S8}. The melting temperature of each set is predicted as the average melting temperatures of the compounds in the subset. For example, for a new material with properties A = 0.7, B = 0.3, and C = 0.2, its predicted melting temperature is the average of the melting temperatures of S5 and S6. The final prediction of RFR is the average of the predictions made by all the trees. For RFR models, 2 hyperparameters determine the model's structure: number of decision trees and maximum depth.

Input Features

There are a total of 253 features composed of 22 material properties and 231 extended features. We trained two sets of models, one set using the 22 material properties and the other set using both the 22 material properties and the 231 extended features. The purpose of these experiments was to determine if using extended features can improve the prediction accuracy of these computational models.

Table 2. Number of training, validation, and testing samples for each compound group.

Compound Group	Number of training samples	Number of validation samples	Number of testing samples
All compounds	846	281	281
Metal	697	231	231
Transition metal	417	138	138
Halogen	128	42	42
Chalcogen	590	196	196

Alkaline earth metal	186	61	61
Post transition metal	216	71	71

Training, Validation, and Testing RFR Models

The dataset of compound features was randomly split into training, validation, and testing sets. For each dataset, 60% of the data was used for training, 20% of the data was used for validation, and the remaining 20% of the data was used for testing. Table 2 shows the number of training, validation, and testing samples for each compound group.

In machine learning, to train machine learning models, we use samples in the training set to learn RFR models with different hyperparameter values, i.e. number of decision trees in RFR and maximum depth of each tree. The prediction accuracy is evaluated using samples in the validation set and compared over different choices of hyperparameters. The optimal hyperparameter values are decided as the ones having the highest prediction accuracy over the validation set. With the chosen hyperparameters, model accuracy is measured over the testing set by comparing the predicted melting temperature with the corresponding ground truth value.

Performance Evaluation

Different metrics can be used to evaluate the prediction accuracy of RFR models. The Mean Absolute Error (MAE) is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

where n is the number of samples, y_i is the melting temperature, and $\hat{y}_i$ is the predicted melting temperature. The Mean Percentage Absolute Error (MPAE) is defined as

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

The Root Mean Squared Error (RMSE) is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (3)$$

We chose to use the MPAE because it accounts for the large range of melting temperatures. For example, Tantalum Carbide (TaC) has a melting temperature of 4037 Kelvin, while the compound $H_{22}C_{10}O_3$ in the chalcogen group has a melting temperature of 230 Kelvin. MPAE normalizes the absolute error by the melting temperature and measures prediction errors across different compounds on the same scale.

Results

We have trained and evaluated RFR models on the following subsets of data: all compounds with all features of 22 material properties and 231 extended features; all compounds with 22 material properties; each compound group with all features; and each compound group with 22 material properties. By learning models using material properties with and without extended features, we can evaluate if extended features help to improve prediction accuracy. By comparing one model for all compounds with separate models for different compound groups, we can evaluate if dividing compounds into compound groups and using a model specific to a compound group can improve prediction accuracy.

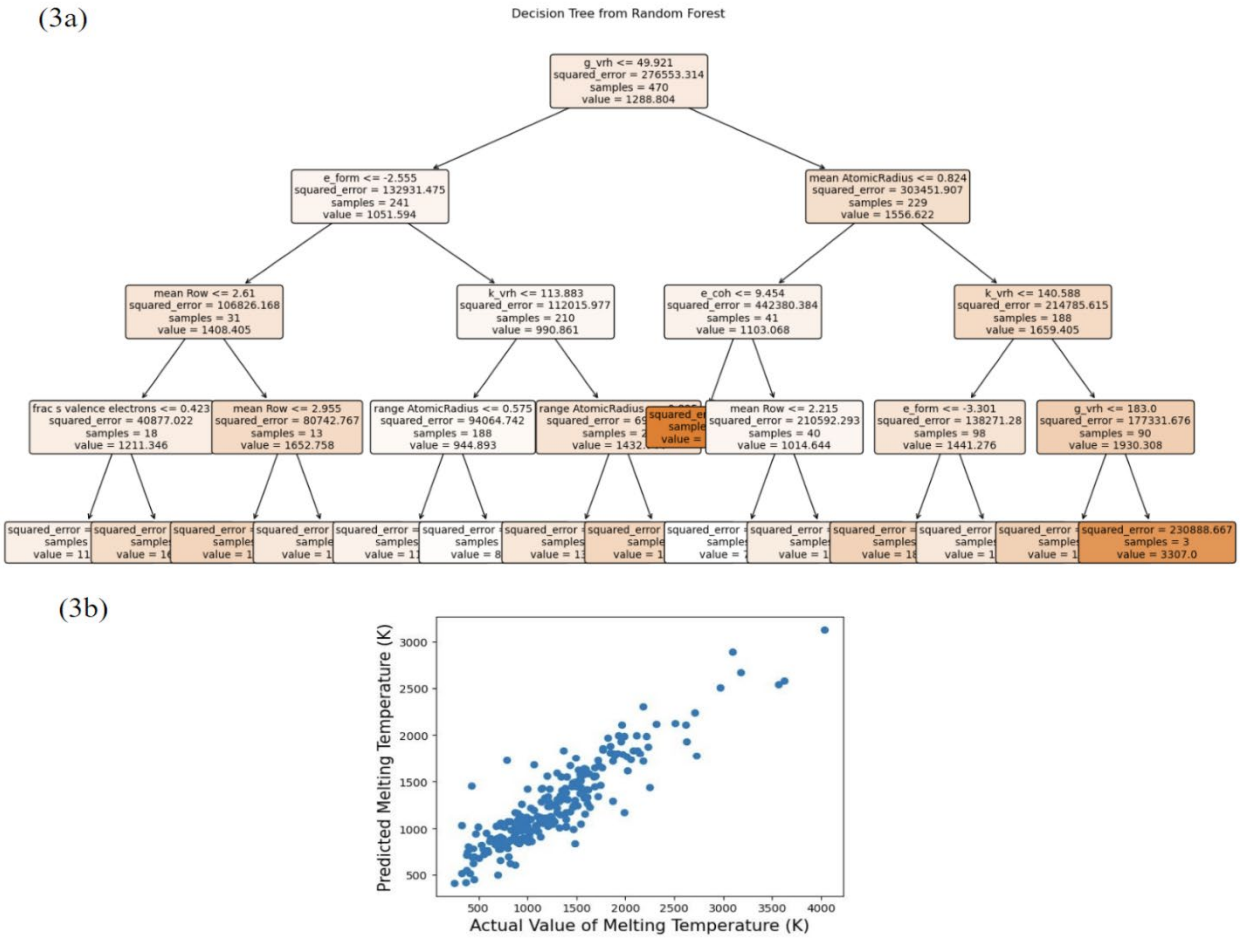

Figure 3. Example of sample decision RFR model and prediction value. (a) One of the decision trees in a RFR model for all compounds using only compound and element-based properties (maximum depth was set to 4). (b) Comparing predicted values with actual values of melting temperature.

One RFR Model for All Compounds

We used all compound samples to train two RFR models to predict melting temperature for all compounds, one model without extended features, and the other model with extended features. In an example to demonstrate the structure of the decision trees learned from material properties, we train a RFR model with 50 decision trees and a maximum depth of 4. Figure 3 shows the structure of the first decision trees in this model. At the root node, g_vrh is selected to split the samples. In level 1 of the tree, e_form is selected to split the samples at the left node and mean AtomicRadius is selected to split the samples at the right node. As the level increases, the predicted value becomes more precise. Using a RFR model of 50 decision trees with maximum depth of 10, the predicted values are displayed and compared with the actual values of melting temperature in Figure 13.

Hyperparameters (Number of Decision Trees, Maximum Depth)

Figure 4 displays the MPAE of the learned RFR model over the testing set with an increasing number of trees, where the maximum depth was set to 10. As the number of trees increases, the prediction error decreases, until

the number of trees reaches around 50 and the prediction error stops decreasing. It suggests that with 50 decision trees, the model reaches optimal performance.

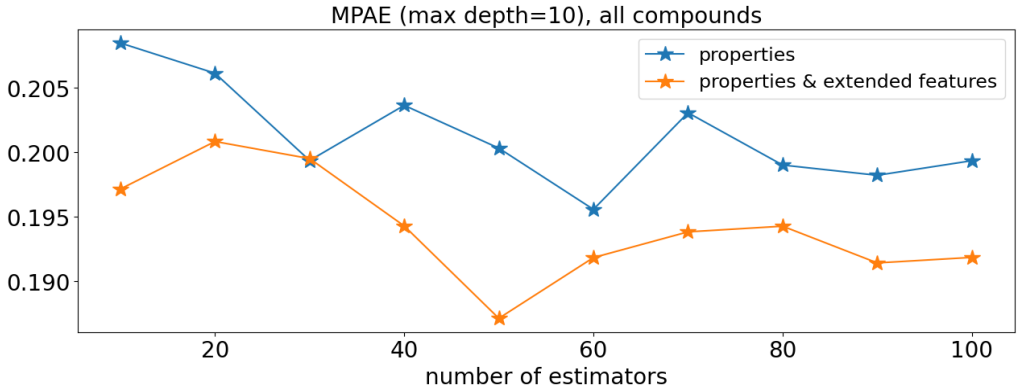

Figure 4. MPAE of model with maximum depth of 10 over different number of trees..

Figure 5 displays the MPAE of the learned RFR model over the testing set with an increasing maximum depth, where there are 50 decision trees. As the maximum depth increases, the prediction error decreases, until a maximum depth of around 10 is reached and the prediction error stops decreasing. It suggests that a RFR model with decision trees of maximum depth 10 is able to reach optimal performance.

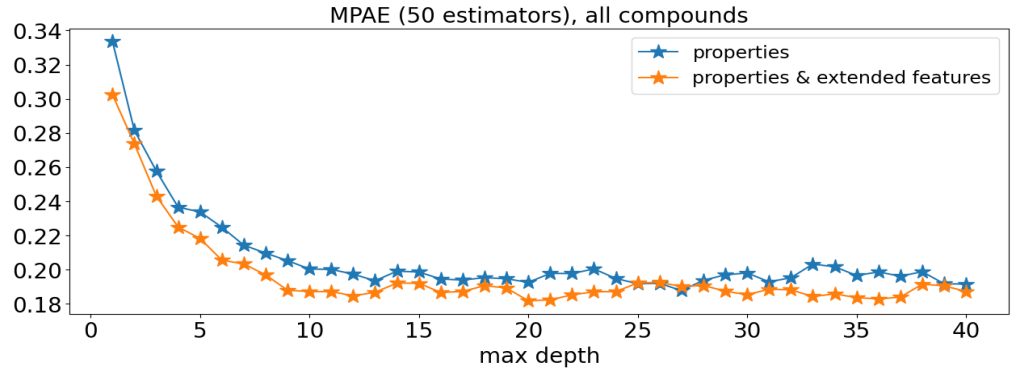

Figure 5. MPAE of model with 50 decision trees over different maximum depths.

Model Performance with and without Extended Features

Figure 4 and Figure 5 both show that with extended features, the RFR model achieved lower prediction error consistently over different number of decision trees and maximum depths. The top row of Table 3 shows the MPAE of RFR models trained with and without extended features.

Table 3. Comparison of MPAE of RFR models with and without extended features over testing set, where the number of trees is 50 and the maximum depth is 10.

Compound group	Material properties only	Material properties and extended features
All compounds	20.0%	18.7%
Metal	15.5%	14.8%

Transition Metal	18.8%	17.2%
Halogen	22.3%	18.7%
Chalcogen	20.6%	20.8%
Alkaline Earth Metal	19.8%	18.7%
Post Transition Metal	17.8%	16.8%

Feature Importance

In a RFR model, a feature is selected at each node to divide compound samples, where the variance of their melting temperatures is high, into two subsets where the variance of melting temperatures in each subset is reduced. We can measure how each feature decreases the variance through the sample split. For each feature, how on average it decreases the variance over different trees is defined as the feature importance.

Figure 6 shows the feature importance in the RFR model trained on only material properties, where the number of decision trees was set to 50 and the maximum depth of trees was set to 10. Considering the random effect of sampling in RFR models, the figure shows that the properties of bulk modulus, shear modulus, formation energy, mean atomic radius, and mean row likely play a more important role in predicting melting temperature than the rest of the properties. The results in this figure also show that both compound properties and element-based properties are useful in predicting melting temperature.

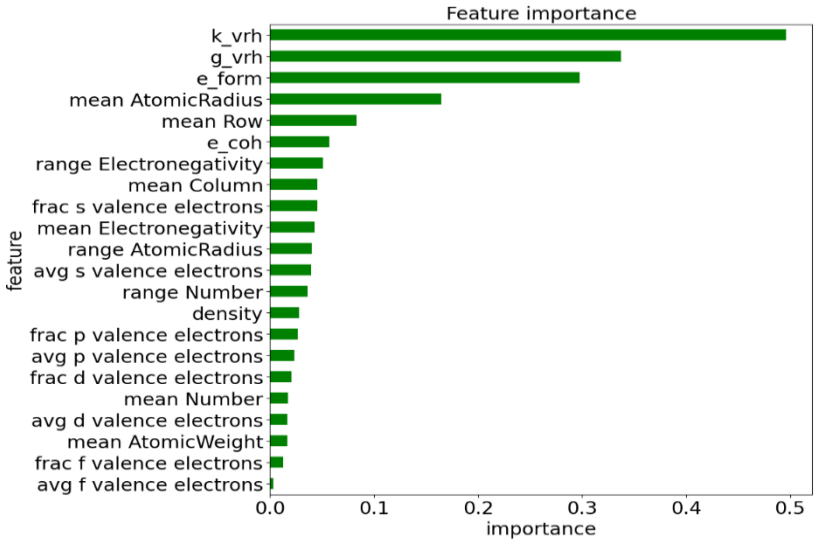

Figure 6. Ranked feature importance of compound properties and (element-based properties (one RFR model for all compounds with 50 decision trees of maximum depth 10).

Figure 7 shows the importance of the top 40 features in the RFR model trained with all 253 features. The majority of the top ranked features are extended features derived from the average number of valence electrons in the suborbitals, particularly the f-suborbital and the most important compound properties (formation energy, mean atomic radius, mean row, elastic moduli) observed in Figure 6. It is interesting to observe that some of these extended features with high feature importance are products of density and the average number of suborbital valence electrons, while density did not have high feature importance in the RFR model learned only with compound and element-based properties.

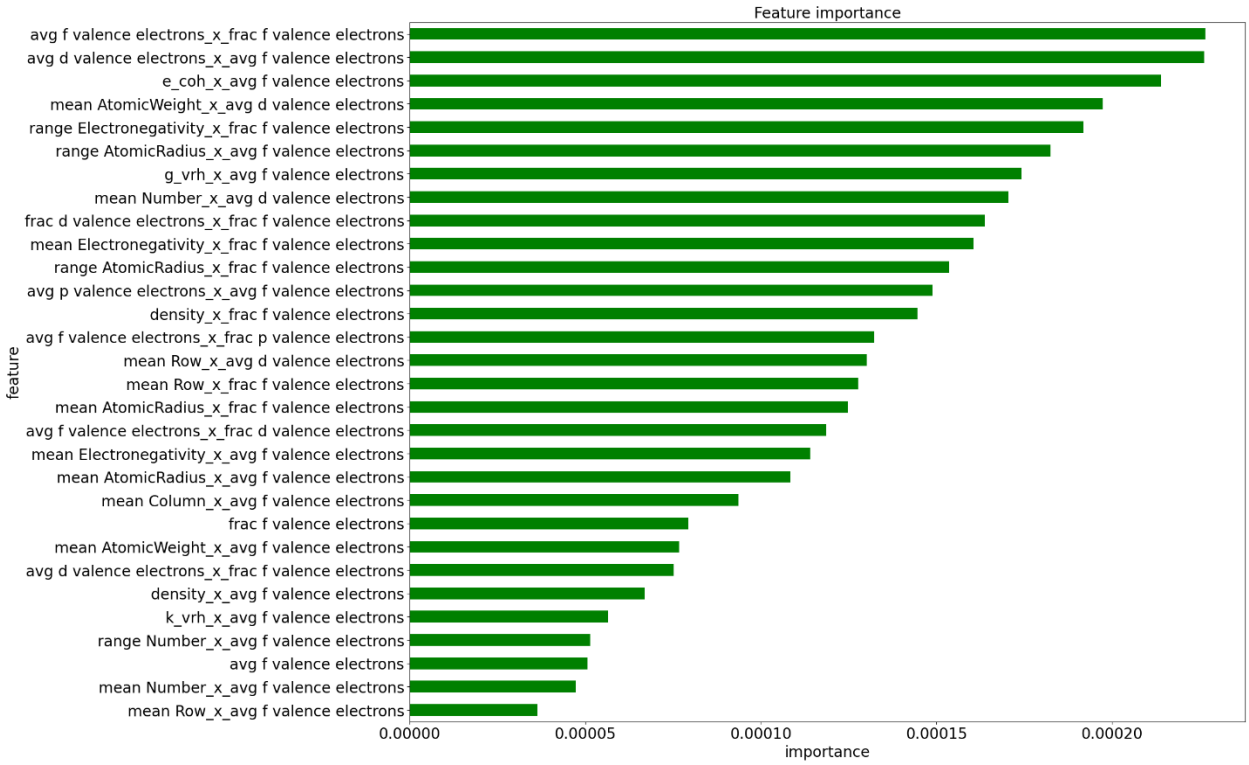

Figure 7. Ranked feature importance of compound properties, element-based properties, and extended features (one RFR model for all compounds with 50 decision trees of maximum depth 10).

We make two observations from these results. First, extended features that use the average number of valence electrons in suborbitals have higher feature importance than the compound and element-based properties in the RFR model. Second, density is shown to have higher feature importance only when used jointly with the average number of valence electrons in suborbitals.

Separate Models for Compound Groups

For each compound group, we trained two RFR models specific to the compound group to predict melting temperature, one model without extended features, and the other model with extended features.

Hyperparameters (Number of Decision Trees, Maximum Depth)

Figure 8 displays the MPAE of the learned RFR model over the testing set with an increasing number of trees, where the maximum depth of each tree was set to 10. For all compound groups except for the halogen group, optimal performance is reached at around 15 decision trees. For the halogen group, the optimal performance is reached at around 40 decision trees. Because the number of samples in the halogen group is much smaller compared to the other compound groups, the results are expected to have a higher random effect.

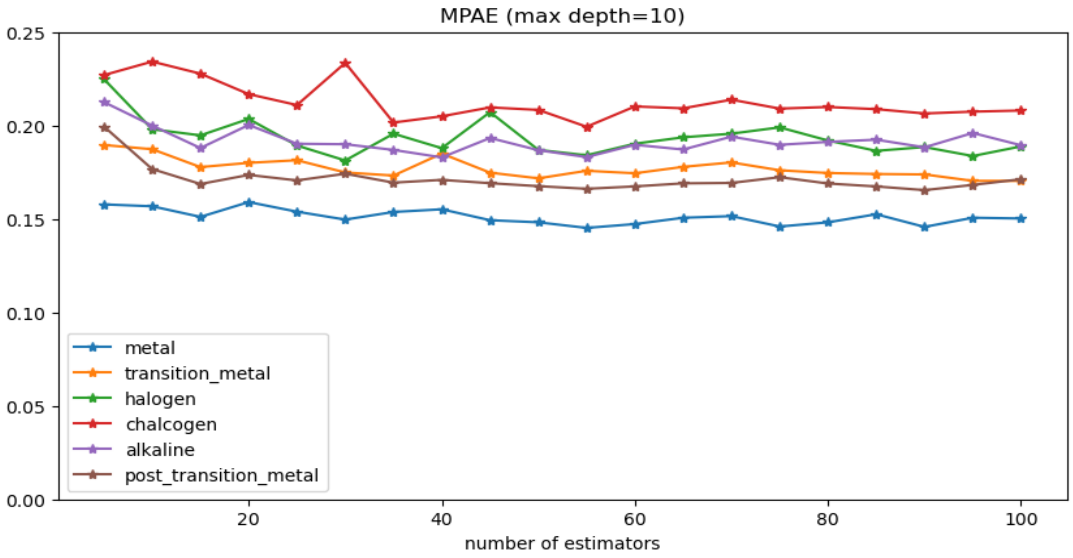

Figure 8. MPAE of RFR models specific to compound groups with different number of trees and a maximum depth of 10.

Figure 9 displays the effect of increasing the maximum depth, where the number of trees was set to 50, to be consistent with the RFR models trained on all compounds. The model reaches optimal performance when the maximum depth increases to 10.

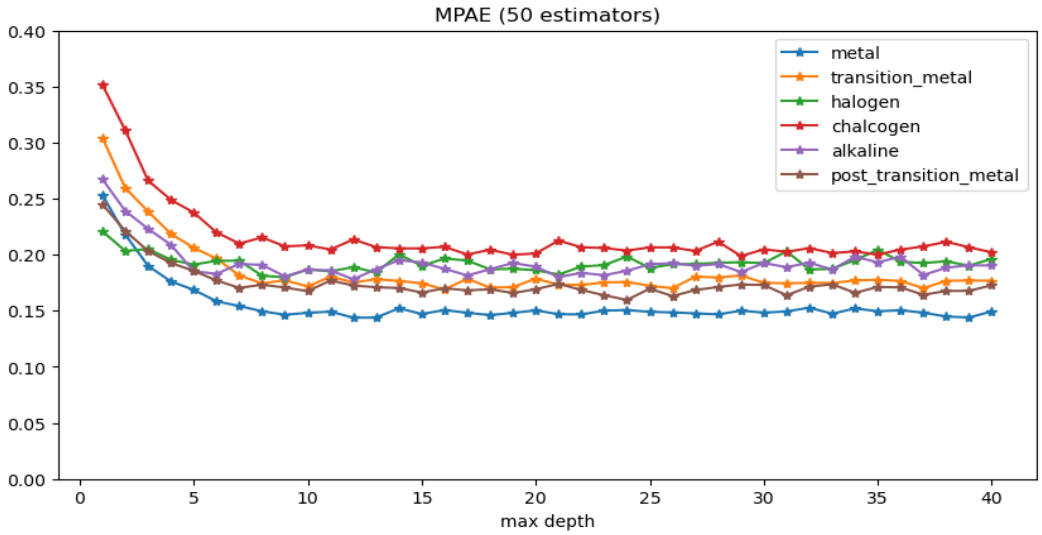

Figure 9. MPAE of RFR models specific to compound groups with different maximum depth of trees and 50 decision trees.

Table 3 shows the MPAE of RFR models specific to each compound group. Compared to a single RFR model for all compounds, RFR models specific to compound groups on average have lower prediction error. Note that the results from the halogen compound group are not as reliable as other groups due to its smaller sample size. In both experiments shown in Figure 8 and Figure 9, the metal, post transition metal, and transition

metal compound groups have lower prediction error than the alkaline earth metal, halogen, and chalcogen compound groups.

Model Performance with and without Extended Features

We compare specific models trained for each compound group with and without extended features to evaluate the effect of using extended features. Figure 10 compares the MPAEs of RFR models specific to compound groups, with and without using extended features, over different numbers of decision trees and different maximum depths of trees. For all compound groups except for the chalcogen compound group, the prediction error was reduced when extended features were used to train the models.

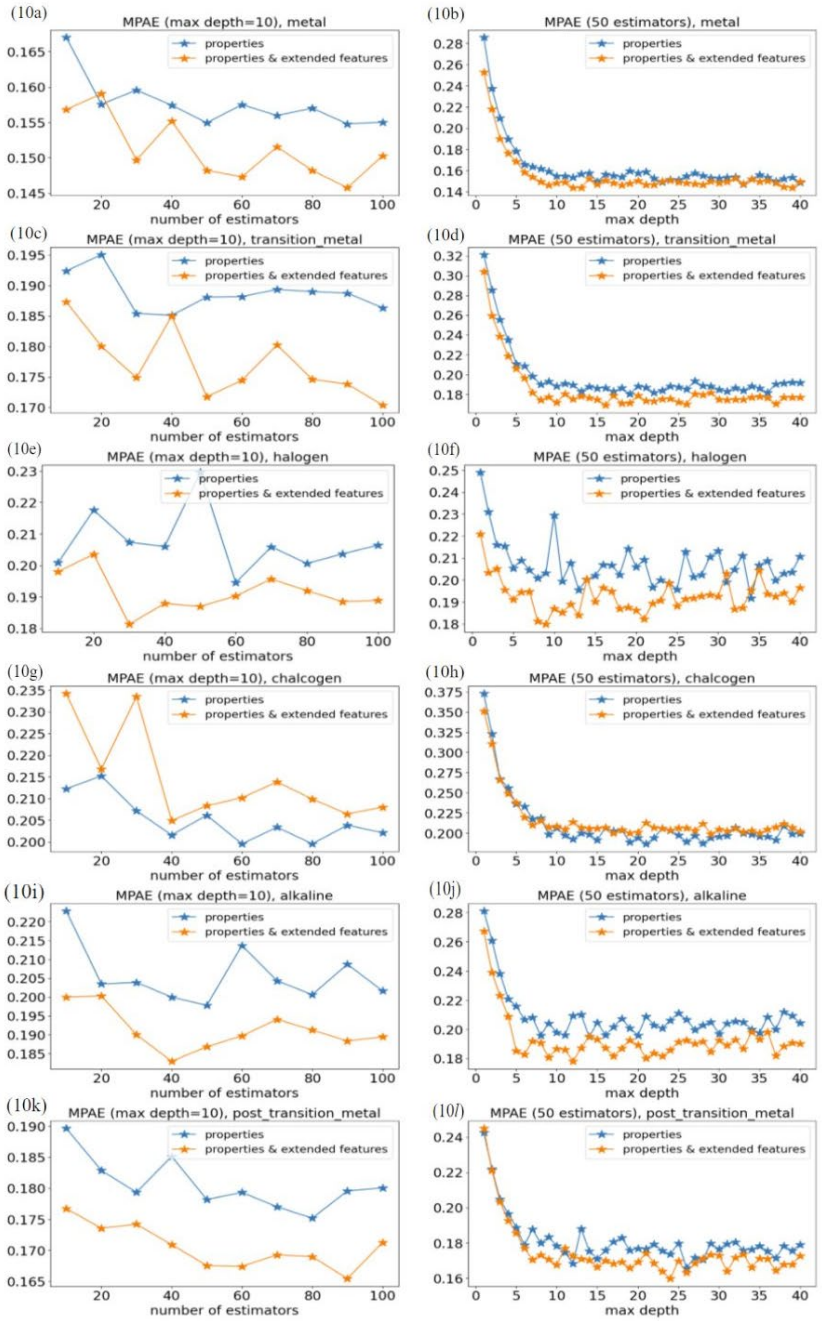

Figure 10. Comparison of MPAE's of RFR models specific to compound groups, with and without extended features. Left column: models with different number of decision trees (maximum depth is 10). Right column: models with different maximum depths of trees (number of trees is 50).

Feature Importance

Figures 11 and 12 show feature importance in RFR models specific to compound groups, where the number of decision trees was set to 50 and the maximum depth of trees was set to 10. The following observations are made. For the metal compound group, the most important features for improving prediction accuracy are the formation energy, mean atomic radius, mean row, mean column, and the range of element electronegativity. For the transition metal compound group, the most important features to accurately predict melting temperature are similar to the metal compound group, but also include the fraction of f-suborbital valence electrons. In the halogen compound group, the most important features to predict melting temperature are the mean column, mean electronegativity, mean row, range of electronegativity of atoms, and average valence electrons in the p-suborbital. In the chalcogen compound group, the most important features for predicting melting temperature are the bulk modulus, cohesive energy, shear modulus, mean row, and range of atomic number. In the alkaline earth metal compound group, the features most important to predicting melting temperature are the bulk modulus, mean column, mean electronegativity of atoms, mean row, and average valence electrons in the s-suborbital. Finally, in the post transition metal compound group, the most important features to predict melting temperature are the bulk modulus, mean atomic radius, mean row, mean electronegativity of atoms, and the range of the atomic number of elements.

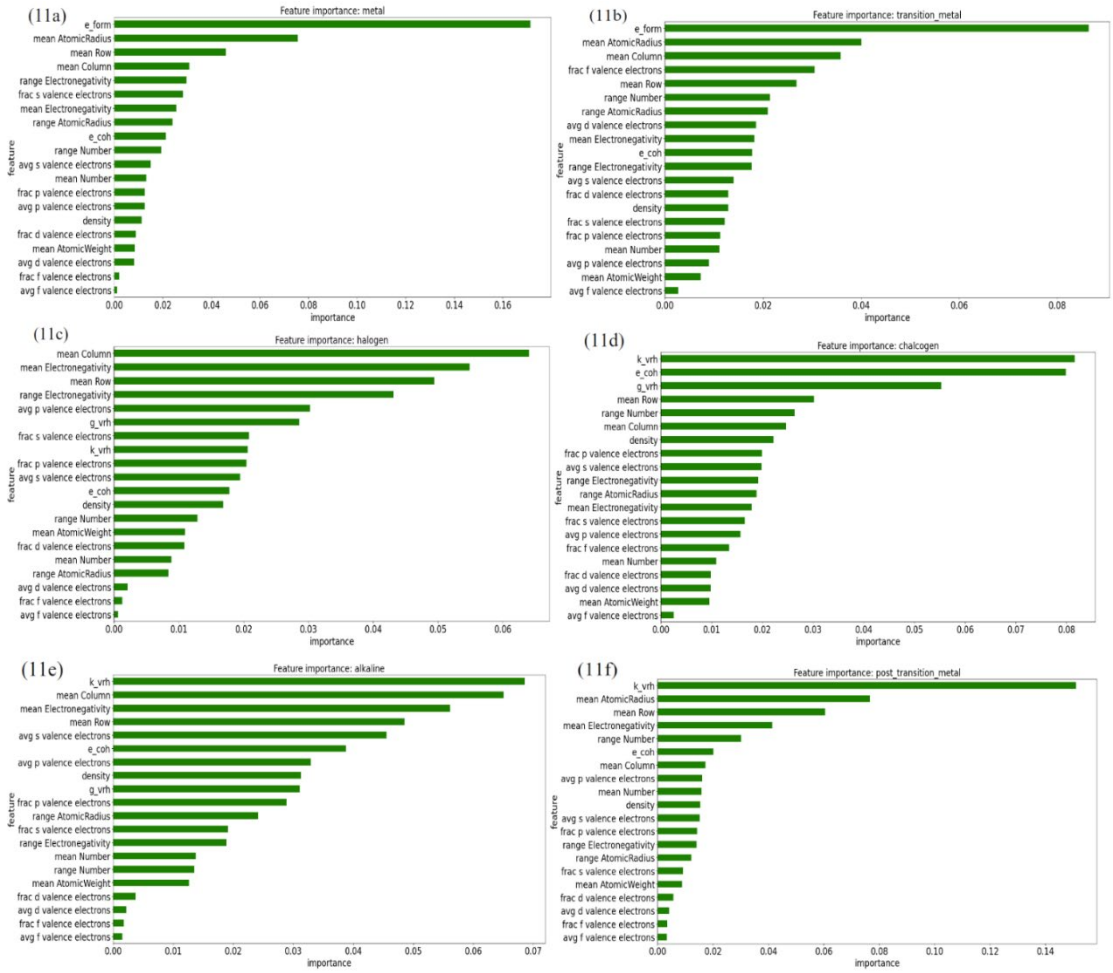

Figure 11. Feature importance in RFR models specific to each compound group, without using extended features. (11a) through (11f) describe the metal, transition metal, halogen, chalcogen, alkaline, and post transition metal compound groups, respectively.

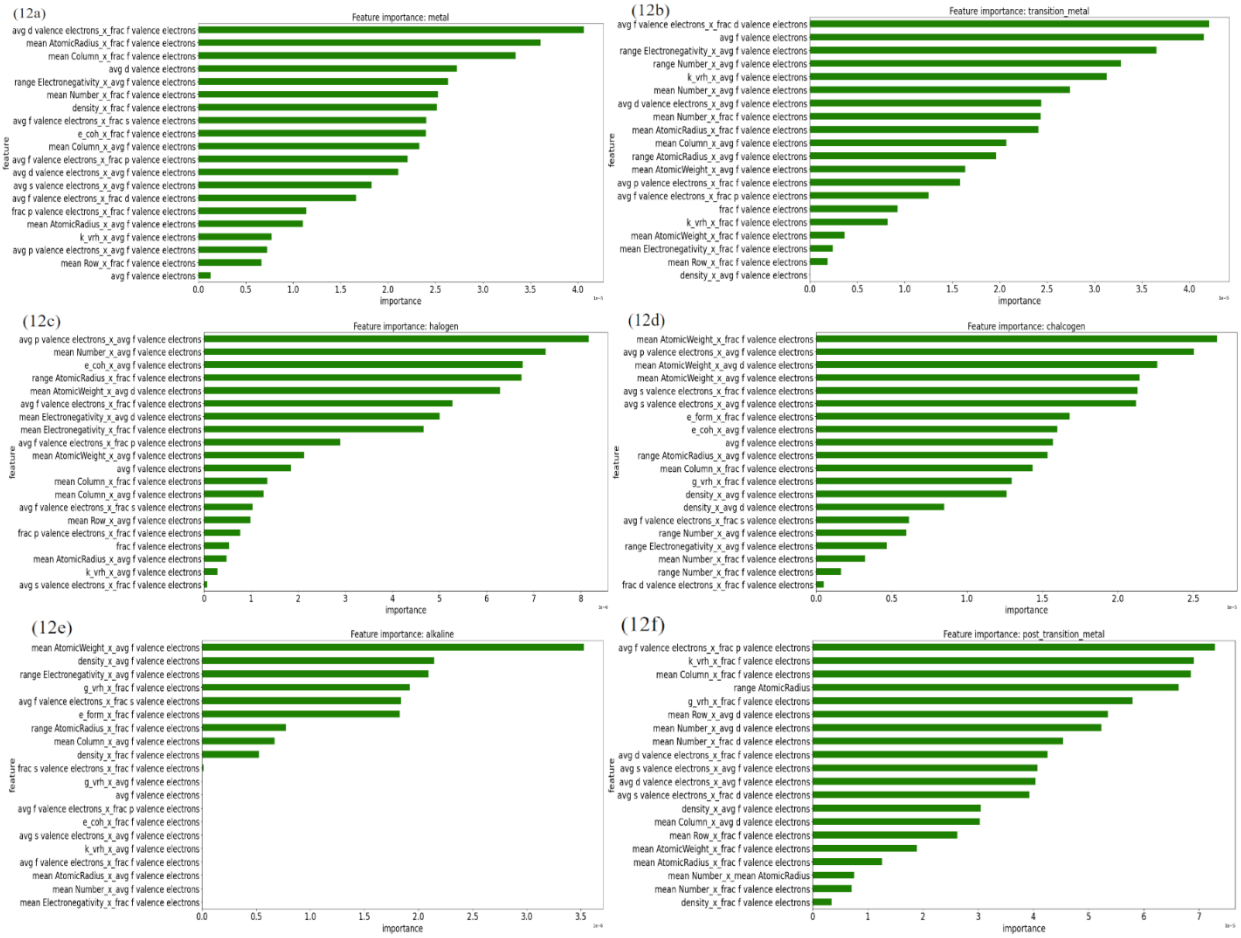

Figure 12. Feature importance in RFR models specific to each compound group, with extended features. (12a) through (12f) describe the metal, transition metal, halogen, chalcogen, alkaline, and post transition metal compound groups, respectively.

For RFR models for specific compound groups trained with both material properties and extended features, the majority of the high-importance features are extended features that jointly use the average number of valence electrons in the f-suborbital and element-based features. It is interesting to observe that in the RFR model for the alkaline earth metal compound group, only 10 features are selected by the RFR model.

Discussion

Through experimentation, we have demonstrated that machine learning models, in particular RFR models, achieved promising results. Compared to a single model for all compounds, models specific to compound groups had better performance on average in predicting melting temperature. We have observed a considerable difference in prediction accuracy observed across different compound groups. The model specific to the metal compound group performed most accurately, while the model specific to the halogen performed least accurately, even compared to the single model for all compounds, likely due to the small number of samples available in the halogen compound group. The use of extended features achieved slight performance improvement for all compound groups except for the chalcogen group.

In models trained with only material properties, both the compound properties and element-based properties were important in predicting melting temperature. The mean atomic row of the atoms was observed as an important feature for all compound groups. Other important properties vary considerably across different compound groups. Energy-related properties and electronegativity-related properties were shown to be important for predicting the melting temperatures of metal-containing compounds. The elastic moduli were observed to be important features for non-metal compound groups. When extended features are introduced, the majority of important properties for each compound group were extended features that jointly used the average number of valence electrons in the f-suborbital and an element-based property.

Conclusion

In this work, we proposed a computational approach to predict melting temperature of compounds, which showed promising results. The machine learning based approach provides a low-cost, time efficient, and safe alternative to determine compound melting temperatures. In particular, the RFR models used in our work are shown to achieve a prediction accuracy of 15%-20% over different compound groups. Future research to improve prediction accuracy includes introducing more material properties, such as the average bond ionic character and molecular symmetry, and quantum properties into the feature set, as well as increasing the number of compounds in the dataset.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Hong Q.J., Ushakov S.V., van de Walle A., Navrotsky A. Melting temperature prediction using a graph neural network model: From ancient minerals to new materials. *Proc Natl Acad Sci U S A*. 2022 Sep 6;119(36):e2209630119. Epub 2022 Aug 31. PMID: 36044552; PMCID: PMC9457469. https://doi.org/10.1073/pnas.2209630119

[2] Lee, A., Sarker, S., Saal, J.E. et al. Machine learned synthesizability predictions aided by density functional theory. *Commun Mater* 3, 73 (2022). https://doi.org/10.1038/s43246-022-00295-7

[3] Guan P.-W., Viswanathan V. MeltNet: Predicting alloy melting temperature by machine learning. (2020). https://doi.org/10.48550/arXiv.2010.14048

[4] Legrain, F., Carrete, J., van Roekeghem, A., Curtarolo, S. & Mingo, N. How chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* 29, 6220–6227 (2017). https://doi.org/10.48550/arXiv.1703.02309

[5] Kirklin, S., Saal, J., Meredig, B. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput Mater* 1, 15010 (2015). https://doi.org/10.1038/npjcompumats.2015.10

[6] de Jong, M., Chen, W., Angsten, T. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci Data* 2, 150009 (2015). https://doi.org/10.1038/sdata.2015.9

[7] *The Materials Project*. Materials Project. https://legacy.materialsproject.org/

[8] *Pymatgen*. pymatgen. https://pymatgen.org/

[9] *Matminer*. matminer (Materials Data Mining) - matminer 0.9.0 documentation. https://hackingmaterials.lbl.gov/matminer/

[10] Meredig B, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*. 2014;89:094104 (2014). https://doi.org/10.1103/PhysRevB.89.094104.

[11] Jha, D., Ward, L., Paul, A. *et al*. *ElemNet*: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci Rep* 8, 17593 (2018). https://doi.org/10.1038/s41598-018-35934-y

[12] A. M. Krajewski, J. W. Siegel, J. Xu, Z.-K. Liu, Extensible structure-informed prediction of formation energy with improved accuracy and usability employing neural networks. *Comput. Mater. Sci.* 208, 111254 (2022). https://doi.org/10.1016/j.commatsci.2022.111254

[13] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model. *IEEE Trans. Neural Netw.* 20, 61–80 (2009). https://doi.org/10.1109/TNN.2008.2005605

[14] S. V. Krivovichev, Structural complexity and configurational entropy of crystals. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* 72, 274–276 (2016). https://doi.org/10.1107/s205252061501906x

[15] Jain, A. *et al*. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *Apl Materials* 1, 011002 (2013). https://doi.org/10.1063/1.4812323

[16] Jain, A. et al. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* 50, 2295–2310 (2011). https://doi.org/10.1016/j.commatsci.2011.02.023

[17] Hautier, G., Ong, S. P., Jain, A., Moore, C. J. & Ceder, G. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Phys. Rev. B* 85, 155208 (2012). https://doi.org/10.1103/PhysRevB.85.155208

[18] Hafner, J., Wolverton, C. & Ceder, G. Toward computational materials design: the impact of density functional theory on materials research. *MRS bulletin* 31, 659–668 (2006). https://doi.org/10.1557/mrs2006.174

[19] Schmidt, J., Marques, M.R.G., Botti, S. *et al.* Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater* 5, 83 (2019). https://doi.org/10.1038/s41524-019-0221-0

[20] Artem R. Oganov, Colin W. Glass; Crystal structure prediction using *ab initio* evolutionary techniques: Principles and applications. *J. Chem. Phys.* 28 June 2006; 124 (24): 244704. https://doi.org/10.1063/1.2210932