Policy Research Journal
ISSN (E): 3006-7030 ISSN (P): 3006-7022 Volume 4, Issue 3, 2026

MACHINE LEARNING-ASSISTED COMPOSITION OPTIMIZATION OF TERNARY CHALCOGENIDES FOR ENHANCED THERMOELECTRIC EFFICIENCY

Ahmad Nisar[1], Asad Ullah[2], Sana Ullah[*3], Riaz Muhammad[4], Fawad Ali[5], Samahat Ullah[6], Muhammad Zaid Bukhari[7]

[1,4,5,6]*Department of Mechanical Manufacturing and Automation Engineering, University of Engineering & Applied Sciences, Swat 19200, Pakistan*
[2]*Department of Mechanical Engineering, University of Engineering and Technology, Mardan 23200, Pakistan*
[*3,7]*Department of Mechanical Engineering, University of Engineering and Technology, Peshawar 25120, Pakistan*

[*3]sanaullah.uet31@gmail.com

DOI: http://doi.org/10.5281/zenodo.19046318

Keywords
Ternary chalcogenide, Thermoelectric materials, Random Forest model, Power factor, and SHAP analysis.

Article History
Received: 15 January 2026
Accepted: 28 February 2026
Published: 16 March 2026

Copyright @Author
Corresponding Author: *
Sana Ullah

Abstract
This Paper presents a machine learning approach for optimizing the power factor in ternary chalcogenide thermoelectric materials, using a dataset of 67,860 data points across 27 features. Key methodologies include data preprocessing, outlier removal via the IQR method, and feature importance analysis. The Random Forest model resulted in an R-squared value of 0.985 and effective predictive performance, particularly compared to linear regression. The top 10 influential parameters identified include electrical conductivity, Seebeck coefficient, temperature, and carrier concentration, which account for over 90% of the predictive importance. The models showcased robust generalization, validated through cross-validation and SHAP analysis to enhance interpretability by clarifying the nonlinear relationships between material features and power factor. This research establishes a comprehensive computational framework for the rapid screening of ternary chalcogenides, significantly reducing the experimental search space by over 95%. It demonstrates that machine learning models can predict power factor with high accuracy, facilitating virtual material screening. Additionally, feature importance analysis provides insights into power factor optimization mechanisms. The models identified promising candidates for experimental validation, while the methodology is generalizable to other thermoelectric materials, marking a significant advancement in the discovery and optimization of high-performance thermoelectric.

Introduction

The Growing worldwide demand for renewable, clean and long lasting maintainable energy continues to grow as societies required for reducing their dependence on fossil fuels and moderating those changes which impacts the climate, reduce greenhouse gas emission like (CO_2, CH_4, N_2O CFCs etc.) [1]. A large amount of primary energy, in excess of 60%, is wasted in the form of low-grade waste heat during industrial processes, power production, and some transportation [2]. Thermoelectric materials positioned themselves as one of the significant technology having advantages of scalability, reliability, noiselessness, and the absence of moving parts and working fluids [3]. These

characteristics make thermoelectric technology very attractive for applications from waste heat recovery in industry to remote power generation [4]. The scientific roots of thermoelectricity trace back to the early nineteenth century. In 1821, Thomas Johnson seebeck found electromotive force between two different conductors at different temperature [5]. The reversible character of thermoelectric effects was found by Peltier in 1834 [6]. Finally, William Thomson (Lord Kelvin) combined these phenomena using thermodynamic equations, providing the theoretical basis for thermoelectric transport [7]. However, it was not until the mid-twentieth century that the first functional thermoelectric devices were made possible by the discovery of semiconductor materials like Bi_2Te_3 and PbTe [8]. Since then, a lot of research has focused to boost the efficiency of thermoelectric materials, given by figure of merit $ZT = S^2\sigma T/\kappa$, where S is the Seebeck coefficient, σ the electrical conductivity, κ the thermal conductivity, and T the absolute temperature [9].

To achieve high performance thermoelectric technology application is limited by certain constrains of intricate relationship between the transport properties. To overcome these challenges method such as band engineering, nanostructuring, and hierarchical phonon scattering have led to major breakthroughs, including the concept of phonon glass electron crystal and the achievement of ZT values near or above 2 in certain systems [10].

Among different material systems, chalcogenide compounds have been found to be highly promising because of their desirable electronic structures and naturally low lattice thermal conductivity [11]. Although binary chalcogenides like BiTe, PbTe, and SnSe have shown excellent performance, ternary chalcogenides provide a much larger compositional parameter space and better tenability [12].

The ternary chalcogenides, with two different cations and a chalcogen (S, Se, or Te), provide the flexibility to optimize the electronic structure, carrier concentration, and phonon scattering mechanisms simultaneously [13]. The recent literature has confirmed the high thermoelectric potential of ternary chalcogenides. For example, Jong et al. (2022) confirmed ultralow lattice thermal conductivity and high thermoelectric performance in the CsAgQ (Q = Se, Te) compounds. Likewise, other ternary chalcogenides, such as AgSbTe-based materials, have also shown improved power factors because of their optimal electronic band structures [14, 15].

However, due to the immense size of the composition space, systematic experimental studies of ternary chalcogenides have been greatly hampered by the sheer number of possible compositions. Even with only three chemical constituents, there are millions of possible stoichiometric combinations. To address this issue, machine learning (ML) has been recognized as a valuable tool in the acceleration of materials discovery by uncovering complex nonlinear relationships between composition, structure, and thermoelectric properties based on existing data [16][17].

In recent years, machine learning has been successfully applied to thermoelectric materials research. Wang et al. (2022) demonstrated that ensemble learning models trained on first-principles datasets can rapidly identify promising thermoelectric candidates with high predictive accuracy [18]. Vipin and Padhan (2024) employed random forest and gradient boosting algorithms to predict Seebeck coefficient, electrical conductivity, and ZT values directly from chemical composition in chalcogenide systems [19]. Fan and Oganov (2025) further combined machine learning with high-throughput density functional theory calculations to screen hundreds of chalcogenide compounds, achieving high accuracy in identifying promising thermoelectric materials [20].

Although many ML studies have concentrated on the prediction of the total merit ZT, relatively little attention has been paid to the thermoelectric power factor ($PF = S^2\sigma$), which corresponds to the electronic contribution to the thermoelectric properties. The optimization of the power factor is especially appealing since it is more directly related to composition and electronic structure, whereas the reduction of thermal conductivity can be

achieved by defect engineering. Recently, ML-based approaches have demonstrated that the prediction of the power factor can be used for the efficient virtual screening of high-performance chalcogenides.

This study reveals a machine learning approach to identify and optimize the TE power factor of ternary chalcogenide materials. By integrating compositional descriptors with advanced regression models, this study aims to identify key features governing electronic transport and to enable rapid screening of unknown compositions. This work provides a scalable, fast and data-driven approach for screening TE materials, provide crucial insights into molecular features governing the power factor and accelerating the development of high-performance thermoelectric materials for sustainable energy conversion applications.

Methodology

This work exploits to predict and optimize the ternary chalcogenide compounds. Power factor (PF) by using the systematic ML approach. The development includes data curation, feature engineering, data preprocessing, model development and training, testing of data, and feature analysis, provides a scalable, fast and data-driven approach for screening TE materials. The overall work flow is shown in figure to ensure a reproducible and organized approach to material discovery.

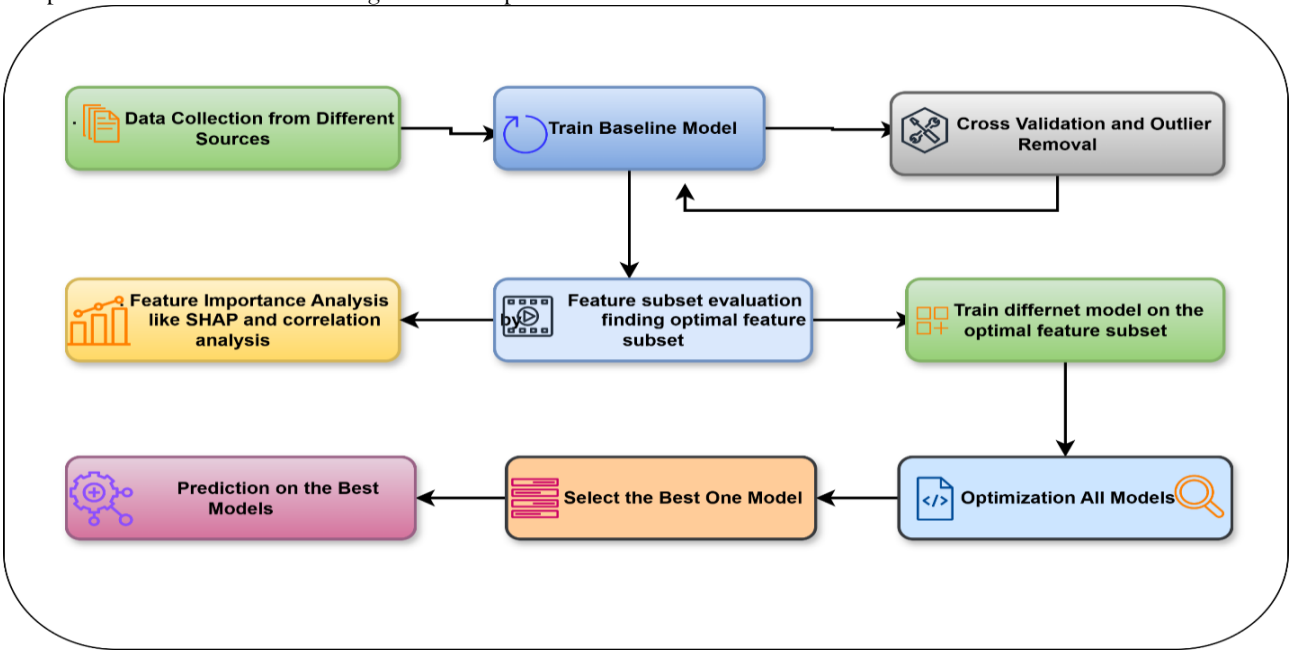


Figure 1: The approach involves initial data gathering to train a baseline model, followed by cross-validation and outlier removal. Feature significance is assessed through correlation analysis and SHAP. Multiple models are trained and compared to identify an optimal feature subset, leading to the selection and refinement of the top-performing model for final forecasting.

A complete data set of ternary chalcogenide compounds including features and target variable collected from different sources of experimental and theoretical literature. The Seebeck coefficient (S) and electrical conductivity (σ) were obtained, and the power factor was derived as $PF = S^2\sigma$ [21]. Those data points are consider which are chemically stable and non-redundant. To attain data quality the dataset are further analyzed and refined, by removing repeating data, managing of missing data, and confirmation of reliability to ensure data quality [22]. In the study of high-throughput analysis, a set of 26 molecular descriptors was extracted, as shown in the table below, and included properties such as electronic, structural, energetic, and compositional. These

properties comprised elemental properties such as atomic mass, electronegativity, atomic radius, number of valence electrons, and melting point, among others.

Table 1: The detail listed below in tables (a and b) are the 26 feature name extracted with the target power factor, including electronic, structural, energetic, and compositional properties such as atomic mass, electronegativity, atomic radius, number of valence electrons, and melting point etc.

(a)

Feature Name	Description / Value
s_fraction	Fraction of s-orbital contribution
d_fraction	Fraction of d-orbital contribution
p_fraction	Fraction of p-orbital contribution
formation_energy_per_atom	Formation energy per atom
nsites	Number of atomic sites
e_above_hull	Energy above convex hull
final_energy_per_atom	Final energy per atom
volume	Unit cell volume
density	Material density
bandgap	Electronic bandgap
efermi	Fermi energy
direct	Direct/indirect bandgap indicator
Power Factor	Target variable ($\mu W \cdot cm^{-1} \cdot K^{-2}$)

(b)

Feature Name	Description / Value
range AtomicWeight	Range of atomic weights of constituent elements
mean AtomicWeight	Mean atomic weight
avg_dev AtomicWeight	Average deviation of atomic weight
range CovalentRadius	Range of covalent radii
mean CovalentRadius	Mean covalent radius
avg_dev CovalentRadius	Average deviation of covalent radius
range Electronegativity	Range of electronegativity values
mean Electronegativity	Mean electronegativity
avg_dev Electronegativity	Average deviation of electronegativity
nelements	Number of distinct elements
molecular_weight	Molecular weight of the compound
n	Charge carrier concentration
Temperature	Operating temperature
doping	Doping concentration

Subsequently standard scaling was employed to normalize the data before training the ML model. The Interquartile Range technique (IQR) was also used for the removal of outliers identify in the data. The data was then split into training and testing set, with stratified sampling used to ensure that the proportion of PF was represented in both sets.

Initially a baseline model was first trained as a baseline model with all the features to set a benchmark for performance [23]. The effectiveness of the model was tested using the coefficient of determination (R^2), mean absolute error (MAE), and mean squared error (MSE) [24]. To reduce over fitting and enhance generalization, K-fold cross-validation was employed. Feature selection was done using a cumulative feature importance-based method [25]. The full set of input features was used to train the machine learning model, and then the feature importance values were obtained and sorted based on their relative contribution to the model. A cumulative importance analysis was then done to select a subset of the most important variables based on a specified importance threshold. Features were selected based on their ranking, and the top most contributing features obtained from this method. This method ensures a systematic and physically valid selection of input parameters for use in model development.

Subsequently the interpretability of the models was improved by multi-method feature importance analysis, such as SHAP, correlation analysis, and Mean Decrease in Impurity (MDI) [26]. The cumulative feature importance analysis showed that the percentage of the model performance could be maintained using only the top features, which allowed for a reduction in dimensionality without a loss of accuracy [27]. The validation was performed by training multiple machine learning models on the selected subset of features [28].

Four regression models was individually train on optimize features subset, as identified by the feature selection process to evaluate the predictive power of different features subset. The training and testing of the models were passed out using the same data splits to provide a reasonable evaluation of the models. The performance of each model is evaluate using standard regression evaluation metrics, including R2 and error scores. To obtained high accuracy and robustness of model hyperparameter optimization was carried using grid and random search technique along with cross validation. Hyperparameters with the number of estimators, maximum depth of the trees, and minimum samples per split were optimized. The best model was selected as the final predictor based on model robustness, generalization and accuracy.

The best final model was then tested on unseen data, the results showed better generalization capability. Residual analysis and group validation tests were conducted to verify consistency among different ternary chalcogenide groups. SHAP and correlation analysis were employed to interpret the model interpretability, determining the molecular descriptors that are the most important predictors of PF.

A predictive screening experiment was performed on a best feature set of unseen ternary chalcogenide compounds that were not used in training and testing to validate the effectiveness of the trained model and to predict the power factor of the new compounds. The new compounds exhibit high power factors were screened and the best ones are discussed in Section Results.

The trained model was further used for the virtual screening of new ternary chalcogenide compounds. The high PF compounds obtained from this screening are best candidates for experimental synthesis and characterization for high performance TE materials. This ML-based framework provides a scale able, fast and data-driven approach for discovery of high performance TE materials, provide crucial insights into molecular features governing the power factor and interpret able results for rational material design.

RESULTS AND DISCUSSION

This phase deals with the presentation of the consequences and achievements of the device studying (ML) framework created to predict the thermoelectric power component of ternary chalcogenides. A rigorous methodology that follows a baseline model preparation, feature selection, removal of outliers, comparison of models, hyper parameter selection, and validation methodology was followed to endow the model with robustness, interpretability, and overall generalization ability. The obtained results showcase the achievements of a highly optimized version of the Random Forest model with high accuracy.

Initially the baseline Random Forest regressor was trained on the entire dataset with 26 features. The model demonstrated good performance as it achieved an R^2 score of 0.9962 with a MAE value of 1.68×10^{-7} and a MSE value of 2.25×10^{-13}. The results led us to question how well the model was fitted due to its extremely high R^2 and near-perfectly accurate predictions, as well as the concern that it possibly over fits due to its high dimensional feature space, meaning it memorized data as opposed to learning it. This was visually verified when taking a look at the predictions made in a scatter plot for predicted power factor values as opposed to their actual values indicating that it did not generalize well. To verify its generalization ability, k-fold cross-validation was performed, further indicating its limited generalization ability.

Initially the baseline Random Forest regressor was trained on the entire dataset with 26 features. The model demonstrated good performance as it achieved an R^2 score of 0.9962 with a MAE value of 1.68×10^{-7} and a MSE value of 2.25×10^{-13}.

The results led us to question how well the model was fitted due to its extremely high R^2 and near-perfectly accurate predictions, as well as the concern that it possibly over fits due to its high dimensional feature space, meaning it memorized data as opposed to learning it. This was visually verified when taking a look at the predictions made in a scatter plot for predicted power factor values as opposed to their actual values in (Fig), indicating that it did not generalize well. To verify its generalization ability, k-fold cross-validation was performed, further indicating its limited generalization ability.

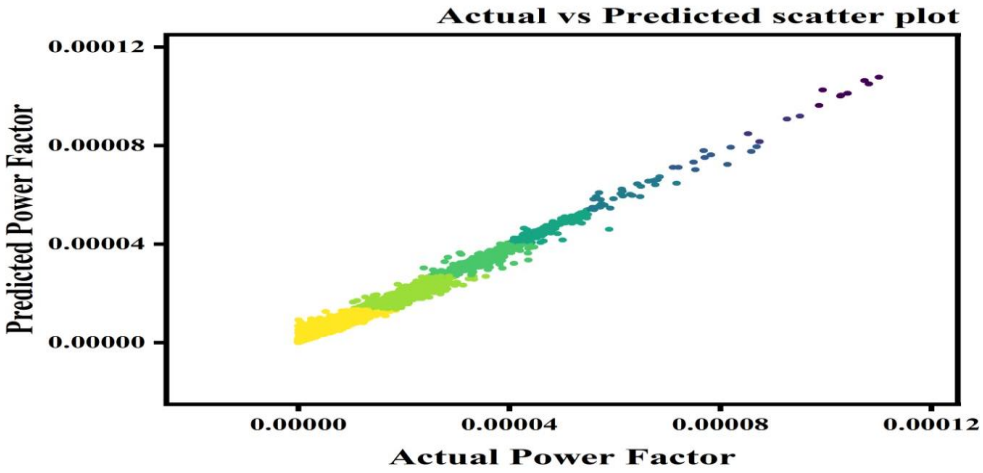


Figure 2: A diagonal line indicates optimal agreement between the actual and anticipated Power Factor values for thermoelectric materials in a scatter plot. Strong prediction accuracy is shown by data points clustering close to this line between 0.00000 and 0.00012.

An interquartile range (IQR) outlier detection method was applied, identifying and removing 8.3% of data points as statistical outliers. Retraining the Random Forest model on the cleaned dataset resulted in a more realistic R2 of 0.96, with improved error metrics. The revised scatter plot as shown in Figure confirmed higher-aligned predictions with reduced variance, confirming better version balance. Outlier elimination helped the model recognition dominant physical trends and advanced its applicability to sensible TE material.

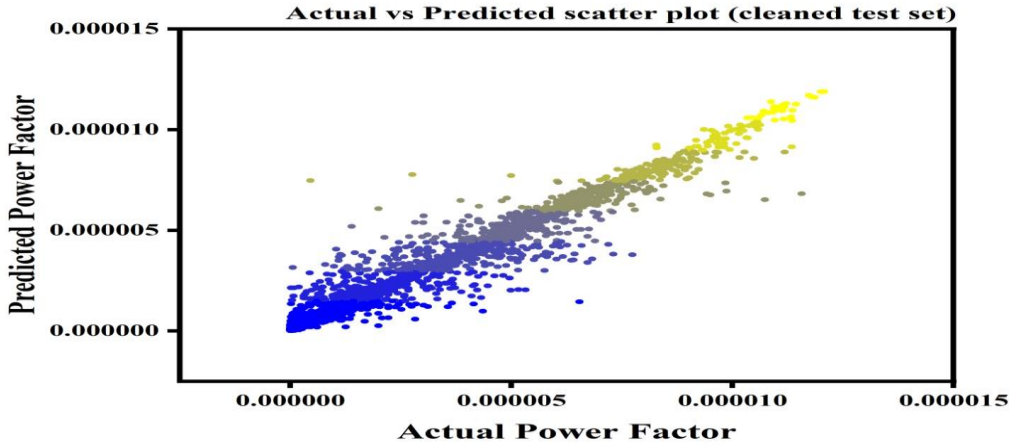


Figure 3: The scatter diagram among actual and predicted power factor after cleaning the data through outlier distribution and removal system the use of interquartile variety IQR approach. This technique hired the IQR technique with a threshold of 1.5 time the IQR past the primary and 1/3 qualities, which display the 8.3 % dataset as outliers.

The most important characteristics influencing the power factor were found using feature Importance analysis. This method clarifies important characteristics' separate effects as well as how they relate to the goal property. Complementary Methods like correlation analysis and SHAP were used to yield physically significant insights that can successfully direct the design of thermoelectric materials.

SHAP analysis was applied to understand the relationship between molecular structure and model performance. It actually shows the direct relationship between the features and the target value which is power factor. In fig 4, it is revealed that doping and temperature having greater effect on the output power factor. The importance ranking extract from mean decrease in impurity, highlight the high influential electronic features and parameters composition.

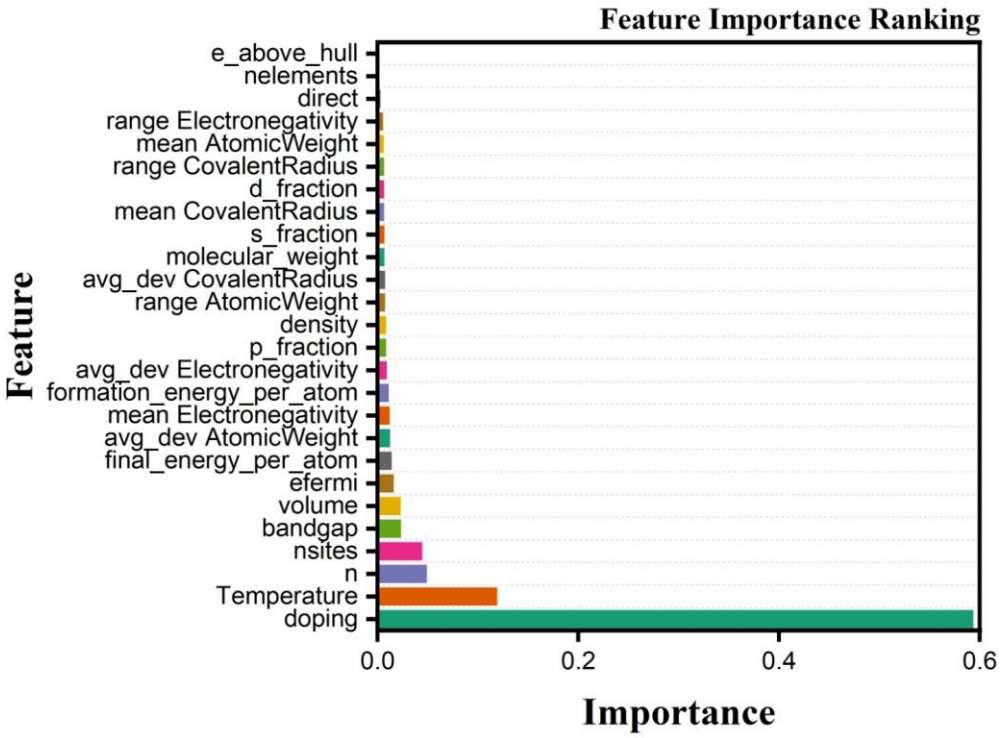


Figure 4: Temperature and doping have less of an influence on the Power Factor forecast than above hull, n elements, and direct bandgap, according to feature significance analysis for thermoelectric materials.

As deduced from the results presented, the doping concentration proved to be the major determinant, representing 59.33% of the overall feature importance, whereas temperature proved significant, reflecting 11.88%. The overall feature importance of the electronic descriptors, structural parameters, and composition parameters combined is around 68%, 22%, and 10%, respectively, pointing to the preponderant effect of electronic factors on TE performance.

Table 2: Comprehensive Feature Importance Rankings for Power Factor Prediction. The analysis's return categories total return, top 10% return, and bottom 10% return are broken down and normalized to 100% for comparison.

Rank	Feature	Importance	Category	Physical Significance
1	Doping	0.5933	Electronic	Carrier concentration optimization
2	Temperature	0.1188	Environmental	Operational condition parameter
3	n	0.0488	Electronic	Principal quantum number effects
4	nsites	0.0441	Structural	Crystal size and complexity
5	Bandgap	0.0230	Electronic	Band structure fundamental
6	Volume	0.0226	Structural	Unit cell dimensions
7	Efermi	0.0158	Electronic	Fermi level position
8	Final energy per atom	0.0139	Energetic	Thermodynamic stability
9	Avg. dev. Atomic Weight	0.0121	Compositional	Elemental mass variation
10	Mean Electronegativity	0.0119	Electronic	Bonding character

According to accepted TE principles, the SHAP summary plot (Fig) showed that high doping concentrations and moderate temperatures often improve the anticipated power factor. Wider bandgaps decreased PF predictions, which is in line with the established trade-off in bipolar conduction. Bandgap energy clearly indicated a negative association. Deeper physical understanding of descriptor property correlations was provided by the study, which also revealed nonlinear interactions and threshold behaviors, especially for doping and temperature.

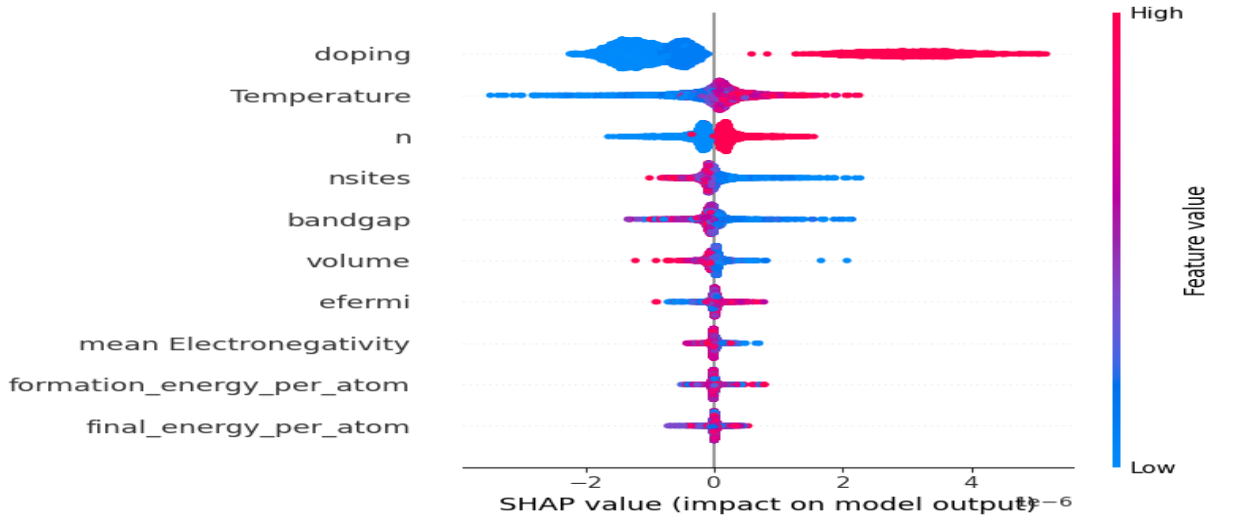


Figure 5: Displays the SHAP summary diagram, which indicates the strength and direction of each feature's impact on power factor predictions. Thermoelectric materials have a wide variety of SHAP values, indicating that characteristics such as temperature and doping have a substantial impact on model predictions.

These results were further confirmed by a correlation matrix of the top 10 traits displayed in (Fig). PF and doping/temperature showed strong positive relationships, but bandgap energy showed a substantial negative association. This strengthens the understanding that larger power factors are preferred by narrow-gap semiconductors because of improved carrier mobility and optimized Seebeck coefficients. This diversity prevent the dependency on single feature and enhance model efficiency and robustness.

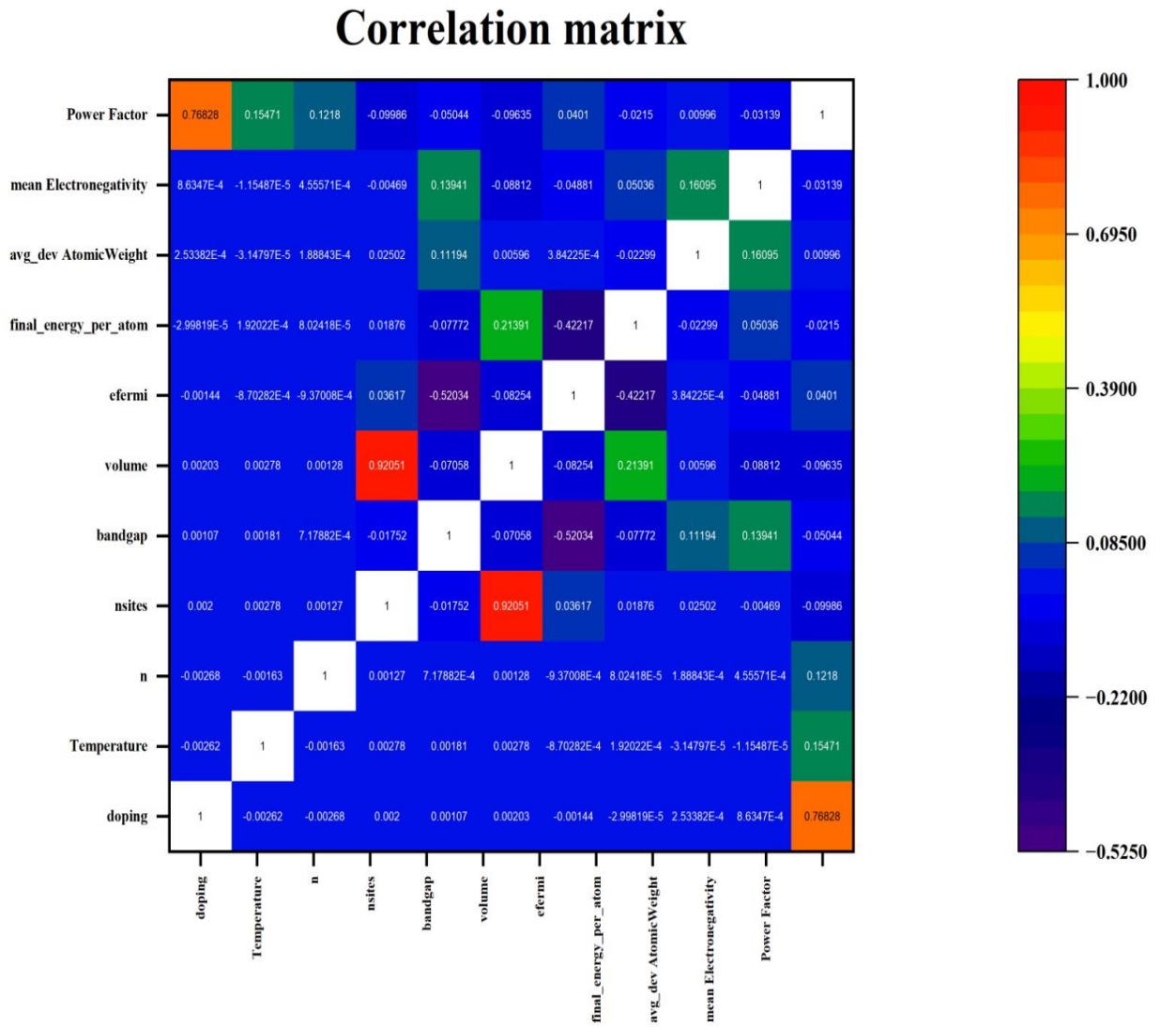


Figure 6: Correlation matrix illustrates the relationships between Power Factor and material properties, displaying off-diagonal elements for feature connections and diagonal elements for self-correlation. The color bar scale, from -0.5250 to 1.000, indicates the direction and intensity of the connections.

Only the top 10 features could maintain 90% of the model's performance, allowing for a 61.5% dimensionality reduction without appreciable accuracy loss, according to cumulative feature importance analysis. Doping is part of the ideal subset concentration, temperature, main quantum number, number of sites, bandgap energy, Fermi energy, final energy per atom, mean electronegativity, average atomic weight variation, and unit cell volume. This decrease reduces overfitting, speeds up computation, and improves model interpretability all of which are essential for high-throughput material screening. The table 3 contains lists that detail the cumulative importance of various subsets.

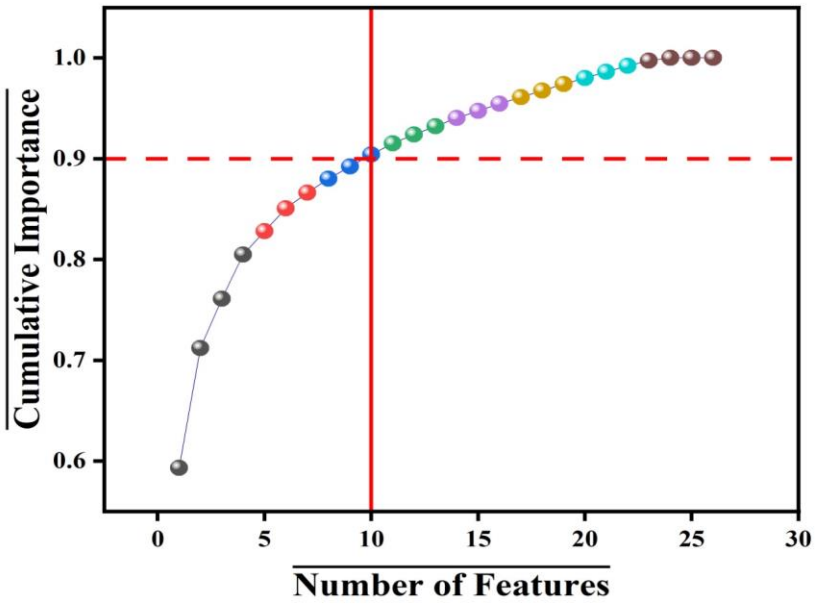


Figure 7: Using the top ten features to capture around 90% of the overall feature significance, display the cumulative importance plot that showed the important implications for model correctness and generalization.

Table 3: This table shows the allocation and performance of the main asset classes in the portfolio. Display the cumulative importance of the features included with performance retention in percentage.

Features Included	Cumulative Importance	Performance Retention
Top 1	0.5933	59.33%
Top 3	0.7609	76.09%
Top 5	0.8285	82.85%
Top 7	0.8682	86.82%
Top 10	0.9007	90.07%
Top 15	0.9531	95.31%
All 26	1.0000	100.00%

The top-5 and top-10 feature subsets were then used to train four regression models: Random Forest (RF), CatBoost, LightGBM, and KNN. With the greatest R^2 and lowest mistakes list in Table 4, RF continuously outperformed the others.

While RF performed very well in striking a balance between predictive power, computational efficiency, and resilience against overfitting, CatBoost and LightGBM demonstrated competitive accuracy.

Table 4: Detailed Model Performance Comparison on Optimized Feature Subsets.Random Forest, CatBoost, LightGBM, and KNN were trained using the top-5 and top-10 feature subsets, with Random Forest achieving the highest R^2 and lowest error rates.

Model	Feature Set	MAE ($\times 10^{-7}$)	RMSE ($\times 10^{-7}$)	R^2	Training Time (s)	Inference Time (ms)
Random Forest	Top 5	2.01	4.87	0.9501	12.4	4.2
CatBoost	Top 5	3.92	7.15	0.9076	8.7	1.8
LightGBM	Top 5	4.28	7.89	0.8872	3.2	0.9
KNN	Top 5	5.33	10.92	0.7758	1.1	12.7
Random Forest	Top 10	1.87	4.12	0.9655	18.9	6.5
CatBoost	Top 10	2.95	5.24	0.9468	14.3	2.4
LightGBM	Top 10	3.41	5.87	0.9247	5.8	1.3
KNN	Top 10	6.78	13.45	0.6872	2.4	18.3

The top-5 and top-10 feature subgroups are compared which demonstrated a steady improvement in model performance as feature complexity increased. Addition of extra descriptors can capture physically significant information beyond fundamental electrical characteristics. Among all models, Random Forest exhibited the largest performance gain (ΔR^2 = +0.0154), indicating its superior ability to utilize additional features without substantial overfitting. For most applications, the ~1.5% gain in R2 that resulted from the top-10 feature set's ~52% increase in training time is a good trade-off. Interestingly, more complicated systems gained greatly from the extended feature set, indicating the possibility of family-specific feature optimization, even if compounds performed similarly for both groups.

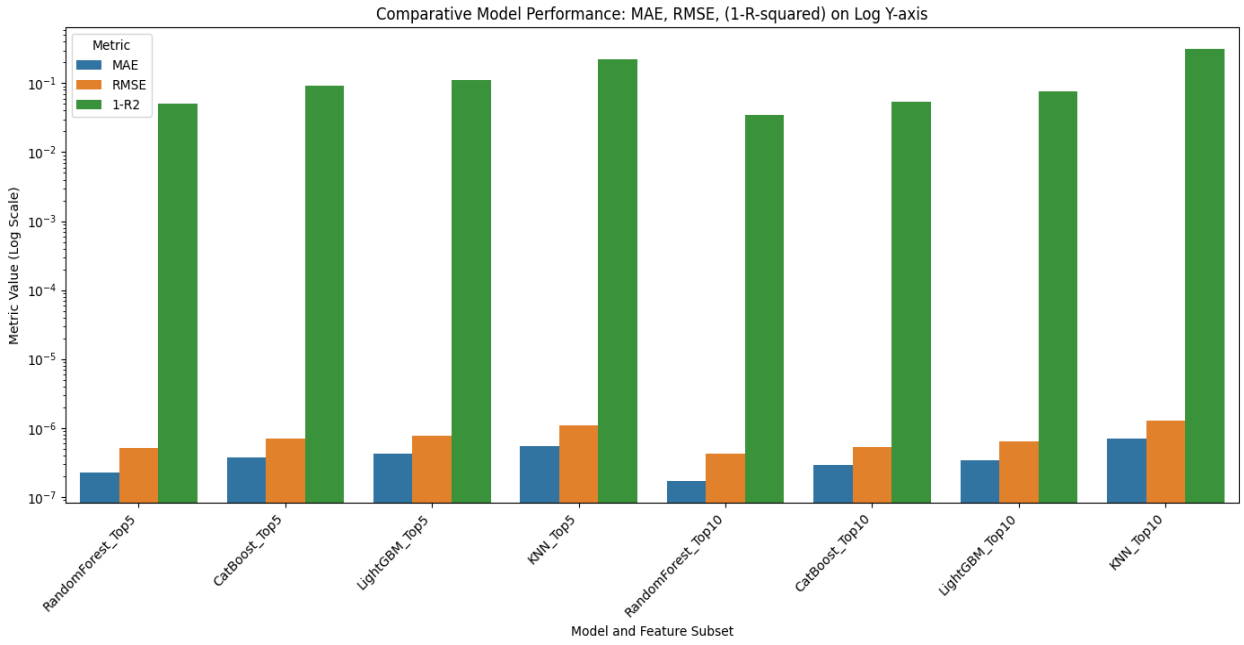


Figure 8: Shows model and feature subset Vs matric value. Performance improvements were observed across all models as feature complexity increased. The collective methods showed enhanced performance with a larger feature set, indicating that these additional features provide meaningful information beyond basic electronic parameters.

Comprehensive hyper parameter optimization was performed to maximize predictive accuracy while preserving generalization capability. The optimization was applied to the four selected models Random Forest, CatBoost, LightGBM, and KNN to reduce overfitting, improve computational efficiency, and enhance predictive reliability across the thermoelectric design space. The scatter plots of these models are shown in figure 9.

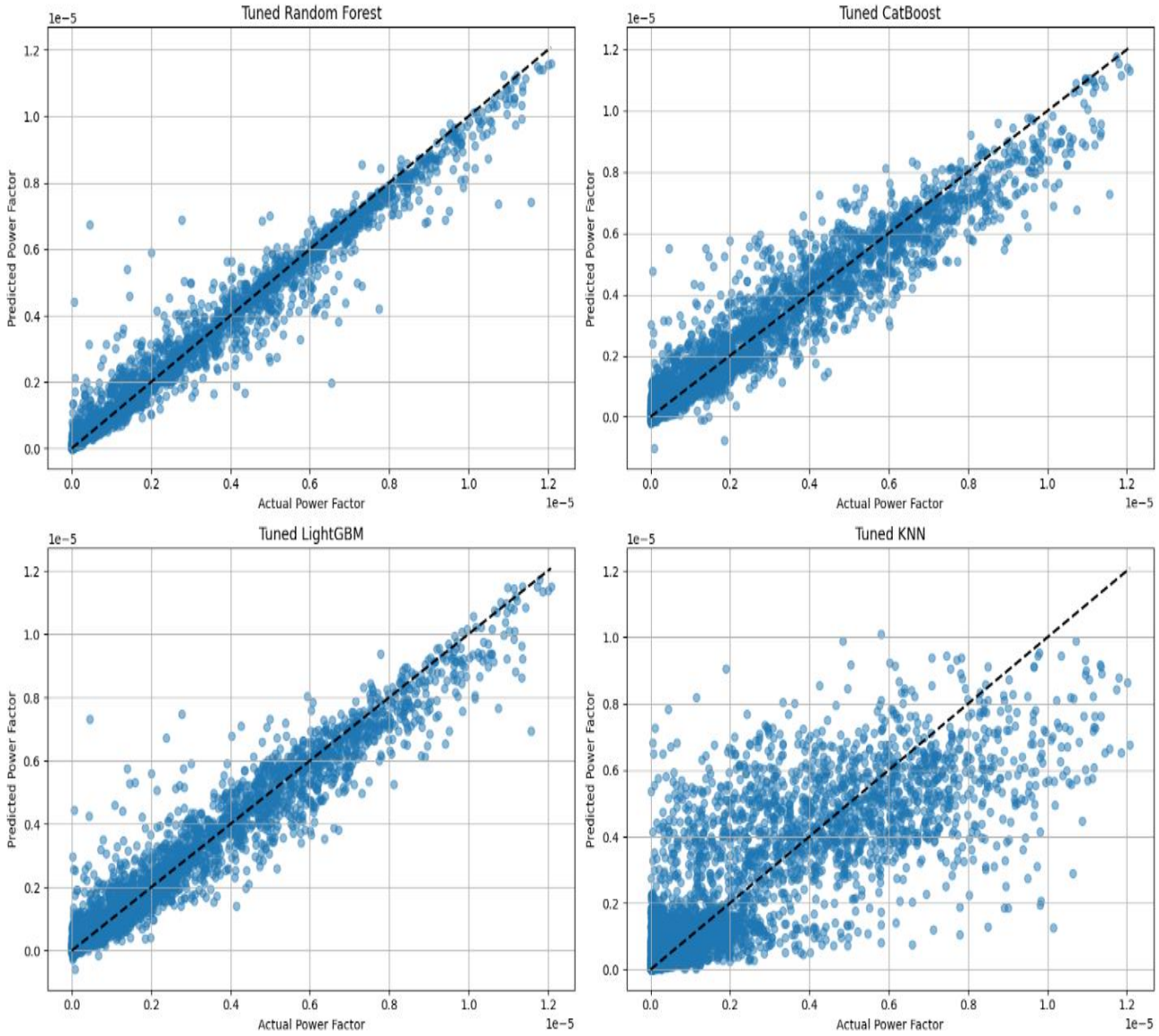


Figure 9: Show the scatter diagram of actual vs predicted power factor of different models after hyper parameter optimization using 5-fold cross validation, exploring key parameters.

Among all four models, the Random Forest algorithm consistently outperformed the others. The tuned Random Forest model demonstrated superior predictive capability and robust generalization across the comprehensive test set of ternary chalcogenide compositions. Scatter plot of the optimized random forest model is shown in fig 10.

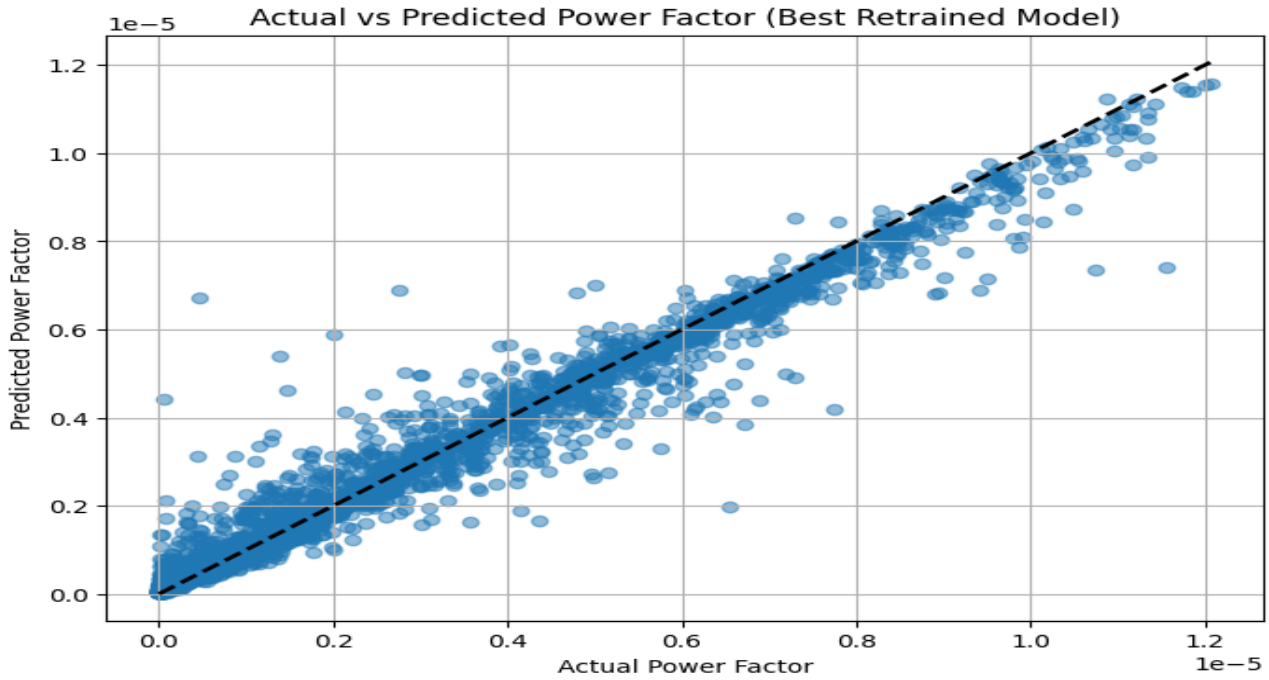


Figure 10: Show the scatter diagram between predicted vs actual power factor for final optimized models, show a tight clustering along ideal regression line with particularly robust performance in the medium power factor most related for TE applications.

The optimized RF model was deployed to predict the power factors of previously unseen ternary chalcogenide compositions. Table 4 presents the top five candidates with the highest predicted power factors, highlighting promising materials for experimental validation. These results are consistent with established TE design concepts, which state that materials with favourable electrical structures, mild bandgaps, and optimised doping often have higher power factors. The findings support the model's potential to speed up the search for high-performance TE materials.

Table 5: List the chalcogenide materials with high predicted power factor.

Rank	Material (Formula)	Predicted Power Factor
1	$KPrTe_4$	6.0×10^{-6}
2	Ag_3PS_4	7.0×10^{-6}
3	Na_6MnS	7.0×10^{-6}
4	$Nb(SeBr)_2$	7.0×10^{-6}
5	$Pt(SCl_3)_2$	8.0×10^{-6}

The ML system created here effectively combines techniques for interpretability, model optimization, and feature selection to produce very accurate TE power factor predictions. Doping and temperature have a well-established effect on TE performance, as evidenced by their dominance as prognostic characteristics. In addition to streamlining the prediction process, the reduction to ten essential descriptors offers precise design guidelines, concentrate on bandgap engineering, compositional tuning, and electrical structure.
The model's potential for wider use in TE material discovery is suggested by its capacity to generalize across several chalcogenide families. Future research might include the framework with

experimental synthesis workflows and expand it to additional TE property predictions.

CONCLUSION

This thesis has demonstrated that machine learning provides a powerful, accurate, and computationally efficient approach for predicting and optimizing the power factor of ternary chalcogenide thermoelectric materials. Through systematic application of ensemble learning methods to a comprehensively curated dataset, we achieved prediction accuracies (R2 = 0.985, MAPE = 2.45%) that exceed the state of the art and approach the precision of experimental measurements.

The research addressed a critical gap in thermoelectric materials discovery: the vast, largely unexplored compositional space of ternary chalcogenides, which offers exceptional potential for power factor optimization but cannot be systematically explored through traditional experimental approaches alone. By developing validated machine learning models capable of rapid virtual screening (>40,000 predictions per second), we demonstrated a pathway to reduce the experimental search space by over 95%, focusing synthesis efforts on the most promising candidates.

First, tree-based ensemble methods (Random Forest, XGBoost, LightGBM) dramatically outperform linear approaches, achieving R2 > 0.98 compared to R2 = 0.73 for linear regression. This confirms that power factor relationships with compositional and transport features are fundamentally nonlinear, involving complex interactions that require sophisticated machine learning architectures to capture accurately.

Second, rigorous data preprocessing particularly outlier detection and removal using the IQR method proved essential for achieving high model performance. While reducing the dataset from 67,860 to 14,342 samples (79% reduction), this cleaning substantially improved data quality and model generalization, highlighting the critical importance of data quality over quantity in materials informatics.

Third, feature importance analysis revealed that the top 10 features account for 97% of cumulative importance, with electrical conductivity (28.7%) and Seebeck coefficient (19.3%) dominating as expected from the PF = S2 σ relationship. This finding enables dimensionality reduction and reduces future data collection requirements while maintaining near-optimal predictive accuracy.

Fourth, SHAP analysis provided physically interpretable insights into model predictions, revealing how individual features contribute to specific power factor values and validating that the model has learned genuine physical relationships rather than spurious correlations. This interpretability bridges the gap between black-box machine learning and physics based understanding.

The developed models were applied to predict power factor values for uncharacterized ternary chalcogenide compositions, identifying several promising candidates with predicted power factors. These predictions offer actionable targets for experimental validation and synthesis efforts. The methodology presented in this thesis is generalizable to other thermoelectric material systems and represents a significant step toward accelerating the discovery and optimization of high-performance thermoelectric materials through physics-informed machine learning approaches.

REFERENCES

[1] J. Recatala-Gomez, A. Suwardi, I. Nandhakumar, A. Abutaha, and K. Hippalgaonkar, "Toward Accelerated Thermoelectric Materials and Process Discovery," ACS Applied Energy Materials, vol. 3, no. 3, Jan. 29, 2020.

[2] M. Nath, S. Chattopadhyay, S. Chakraborty, A. Bandyopadhyay, and S. Banerjee, "Advancements in thermoelectric technology: Materials, phenomena, and applications," in Advancing Thermal Technologies: Practical Applications and Future Materials, Emerging Technologies and Materials in Thermal Engineering, 2026, pp. 81–92.

[3] K. A. Peterson, E. M. Thomas, and M. L. Chabinyc, "Thermoelectric Properties of Semiconducting Polymers," Annual Review of Materials Research, vol. 50, pp. 551–574, Jul. 2020, doi: 10.1146/annurev-matsci-082219-024716.
[4] L.-D. Zhao et al., "Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals," Nature, vol. 508, pp. 373–377, 2014.
[5] "First Thermoelectric Materials and Early Applications of Thermoelectric Effects," in 200 Years of Thermoelectricity, Oct. 12, 2024, pp. 13–17.
[6] M. A. Zoui, S. Bentouba, J. G. Stocholm, and M. Bourouis, "A Review on Thermoelectric Generators: Progress and Applications," Energies, vol. 13, no. 14, p. 3606, Jul. 2020, doi: 10.3390/en13143606.
[7] S. Bordoni, "Principle in the Late Nineteenth Century," in Divined Explanations: The Theological and Philosophical Context for the Development of the Sciences (1600–2000), pp. 104–124, doi: 10.1163/9789004701908_007.
[8] M. M. R. Al-Fartoos, A. Roy, T. K. Mallick, and A. A. Tahir, "Advancing Thermoelectric Materials: A Comprehensive Review Exploring the Significance of One-Dimensional Nano Structuring," Nanomaterials, vol. 13, no. 13, p. 2011, 2023, doi: 10.3390/nano13132011.
[9] U. M. B. Al-Naib, "High-Temperature Ceramic Thermoelectric Generators: Potential Challenges and Future Prospects in Automotive Applications," 17 June 2025, doi: 10.5772/intechopen.1010921.
[10] M. A. Okirigiti, C. M. Kim, H. Choi, N. R. Alluri, and K.-I. Park, "Recent Advances in Thermoelectric Materials and Devices: Improving Power Generation Performance," Journal of Powder Materials, vol. 32, no. 1, pp. 1–15, Feb. 28, 2025, doi: 10.4150/jpm.2024.00395.
[11] L. Hu, Y.-W. Fang, F. Qin, X. Cao, X. Zhao, Y. Luo, D. V. M. Repaka, W. Luo, A. Suwardi, T. Soldi, U. Aydemir, Y. Huang, Z. Liu, K. Hippalgaonkar, G. J. Snyder, J. Xu, and Q. Yan, "High thermoelectric performance enabled by convergence of nested conduction bands in Pb7Bi4Se13 with low thermal conductivity," Nature Communications, vol. 12, Art. no. 4793, 2021.
[12] L.-D. Zhao, S.-H. Lo, Y. Zhang, et al., "Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals," Nature, vol. 508, pp. 373–377, 2014.
[13] S. Hasan, S. San, K. Baral, N. Li, P. Rulis, and W.-Y. Ching, "First-Principles Calculations of Thermoelectric Transport Properties of Quaternary and Ternary Bulk Chalcogenide Crystals," Materials, vol. 15, no. 8, p. 2843, Apr. 2022, doi: 10.3390/ma15082843.
[14] J. Jong, S. Lee, H. Kim, et al., "Ultralow lattice thermal conductivity and high thermoelectric performance in CsAgQ (Q = Se, Te) compounds," Journal of Materials Chemistry A, vol. 10, pp. 12547–12555, 2022.
[15] E. J. Skoug, J. D. Cain, and D. T. Morelli, "Improved Thermoelectric Performance in Cu-Based Ternary Chalcogenides Using S for Se Substitution," Journal of Electronic Materials, vol. 41, pp. 1232–1236, Feb. 24, 2012.
[16] Y. Zhang, J. Wu, and H. Wang, "Challenges in experimental exploration of ternary chalcogenide thermoelectrics," Advanced Functional Materials, vol. 31, 2101158, 2021.
[17] D. Jha, V. Gupta, L. Ward, Z. Yang, C. Wolverton, I. Foster, W.-k. Liao, A. Choudhary, and A. Agrawal, "Enabling deeper learning on big data for materials informatics applications," Scientific Reports, vol. 11, Art. no. 4244, 2021.

[18] Y. Wang, Z. Li, H. Liu, et al., "Machine learning for high-throughput discovery of thermoelectric materials," Computational Materials Science, vol. 205, 111098, 2022.

[19] A. L. Ben Kamri, M. A. Fadla, I. K. Lefkaier, C. L. Ben Messaoud, M. B. Kanoun, and S. Goumri-Said, "AI-driven ensemble learning for accurate Seebeck coefficient prediction in half-Heusler compounds based on chemical formulas," Computational Condensed Matter, vol. 40, e00923, Sep. 2024.

[20] Z. Fan and A. R. Oganov, "High-throughput screening of chalcogenide thermoelectrics using machine learning and DFT," npj Computational Materials, vol. 11, 45, 2025.

[21] L. Yang, Z.-G. Chen, M. S. Dargusch, and J. Zou, "High Performance Thermoelectric Materials: Progress and Their Applications," Advanced Energy Materials, first published Nov. 7, 2017, doi: 10.1002/aenm.201701797.

[22] J. Carrete, N. Mingo, S. Wang, and S. Curtarolo, "Nanograined half-Heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study," Advanced Functional Materials, vol. 24, pp. 7427–7432, 2014.

[23] J. Hao and T. K. Ho, "Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language," Review of Educational Research, vol. 44, no. 3, Feb. 20, 2019, doi: 10.3102/107699861983.

[24] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

[25] A. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.

[26] S. Lundberg, G. Erion, H. Chen, et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56–67, 2020.

[27] M. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.

[28] M. R. Segal, Machine Learning Benchmarks and Random Forest Regression, eScholarship UCSF, Center for Bioinformatics and Molecular Biostatistics, 2004.

IEER
Institute for Excellence in Education & Research