Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

BENCHMARKING MACHINE LEARNING MODELS FOR POWER FACTOR PREDICTION IN BINARY THERMOELECTRIC COMPOUNDS

Fawad Ali[1], Asad Ullah[2], Sana Ullah[*3], Riaz Muhammad[4], Ahmad Nisar[5], Samahat Ullah[6], Saqib Khan[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: https://doi.org/10.5281/zenodo.19016631

Keywords
Thermoelectric materials, Binary compounds, Power factor, Machine learning, Benchmarking, SHAP analysis, Feature importance, Materials informatics

Article History
Received: 14 January 2026
Accepted: 26 February 2026
Published: 14 March 2026

Corresponding Author: *
Sana Ullah

Abstract

Thermoelectric materials have the ability to directly convert waste heat into electricity, providing a sustainable energy solution. Electrical performance of these materials is determined by the power factor, which is an important parameter of the thermoelectric figure of merit. Nonetheless, high-performance binary thermoelectric compounds are sparsely found experimentally and are costly to synthesize. Machine learning has now become an influential instrument to speed up the discovery of materials, but a systematic benchmarking of algorithm currently existing in binary thermoelectric data is unavailable. In this work, we come up with a thorough machine learning benchmarking system to predict the power factor of binary thermoelectric compounds by relying on both compositional and elemental descriptors. An analysis of a large binary data set consisting of 22750 samples and 26 input variables obtained as the result of first-principles calculations was performed. Machine learning models used in predicting the power factor of cubic binary thermoelectric compounds are compared, such as linear regressors, support vector machines, tree-based ensembles, and neural networks. Following the hyperparameter tuning using randomized search, CatBoost has the highest predictive accuracy with a test R^2 of 0.9897, followed by Gradient Boosting (0.9869), LightGBM (0.9861), and Random Forest (0.9804). The poor performance of linear models and the support vector regressors implies the non-linearity of the structure-property relationships. The computational efficiency is evaluated by analyzing the training time of each model. The findings indicate that ensemble and gradient boosting models have a better predictive performance than linear and kernel based models. The additional study of model interpretability is conducted with the help of SHAP analysis to reveal the most significant descriptors that control the power factor prediction. Lastly, the importance of the polynomial feature expansion is examined to determine the effects of interaction between the features. The study offers a methodical reference of machine learning models to predict the power factor of binary thermoelectric materials and outlines the promising inquiry of data-driven methods in initial screening of materials.

INTRODUCTION

The world population keeps increasing its consumption of energy and there is a pressing need to reduce climate change through the use of sustainable technologies. In the industrial processes, transportation and power generation, a large percentage of primary energy, more than 60 percent, is wasted as heat [1]. Thermoelectric substances provide a special solid-state approach to direct conversion of waste heat into electricity, characterized by the benefits of no moving components, high reliability and scalability [2]. The performance of a thermoelectric material is measured by the dimensionless figure of merit ZT. Which is equal to

$$ZT = (S^2 \sigma/\kappa) T$$

Where S is the Seebeck coefficient, σ is the electrical conductivity, κ is the thermal conductivity, and T is the absolute temperature [3]. The numerator $S^2 \sigma$, known as the power factor (PF), which is the electrical power output capability and a target that is important to optimize. Binary thermoelectric compounds, such as Bi_2Te_3, PbTe, and SnSe, form the basis of thermoelectric technology and continue to be subjects of investigation because of their comparatively straightforward crystal chemistry and tunable properties [4, 5]. Nevertheless, new binary thermoelectric materials with improved power factors can be experimentally discovered, but the large compositional space and the complicated interplay of electronic and structural parameters impede it [6]. The traditional methods of trial and error are both time-consuming and expensive and the density functional theory (DFT) calculations, despite their accuracy, are expensive to compute high throughput screening [7].

Machine learning (ML) has since become a revolutionary instrument in materials informatics, allowing one to quickly predict material properties based on available data [8–10]. ML is capable of modeling intricate non-linear correlations and screening through thousands of applications in a productive manner, massively speeding up the discovery pipeline [11, 12].

Recent works have shown the effective use of machine learning methods in predicting the thermoelectric properties, such as Seebeck coefficient [13], electrical and thermal conductivity [15], and power factor [16]. Ensemble models like Random forests and gradient boosting algorithms have been found to be more effective than the traditional linear regression strategies in its capacity to take into account nonlinear interactions and feature dependencies between them. In addition, machine learning adoption in thermoelectric research has also been enabled by the presence of databases with large amounts of materials and high-throughput computational information [17]. Nevertheless, the vast majority of research uses one algorithm, or a small number of models, and no overall benchmark that covers a large variety of current ML algorithms directly on binary thermoelectric compounds with a large, consistent dataset.

Additionally, the selection of algorithm also has a strong influence on the predictive performance, training time and interpretability. As the ML libraries and algorithms rapidly evolve, it is necessary to offer the thermoelectric community a systematic assessment of models to inform their choice [18, 19]. Knowledge of the features that affect model predictions can be used to gain insight into the underlying material-property relationships, as well as increase confidence in data-driven predictions. SHAP (SHapley Additive exPlanations) is among the techniques that have been increasingly used to measure the importance of the features and improve the transparency of complicated machine learning models [20, 21].

In the work, we provide a detailed benchmarking study of 14 machine learning algorithms used to predict the power factor of cubic binary thermoelectric compounds. These models were employed on a dataset of 22,750 samples based on DFT computations and curated by the Materials Project [22]. We considered linear, kernel based, tree based and neural network models to investigate the best algorithm for binary TE material predictions [23, 24]. All

models are adjusted through randomized search with cross-validation so that they can be fairly compared. The predictive accuracy and the training time of all models were analyzed and compared [25-27]. In addition, SHAP analysis is used to determine the descriptors that play a significant role in predicting the power factor, and the polynomial feature expansion is evaluated to determine how feature interactions affect power factor prediction [28, 29]. The findings allow a sensible recommendation of machine learning models to be used in predicting power factors and to show how data-driven solutions in thermoelectric materials screening can be effective [30, 31].

2. Methodology

In present research, we consider a systematic methodology framework to conduct benchmarking of ML algorithms to predict binary thermoelectric materials power factor. Figure 1 illustrates the work flow.

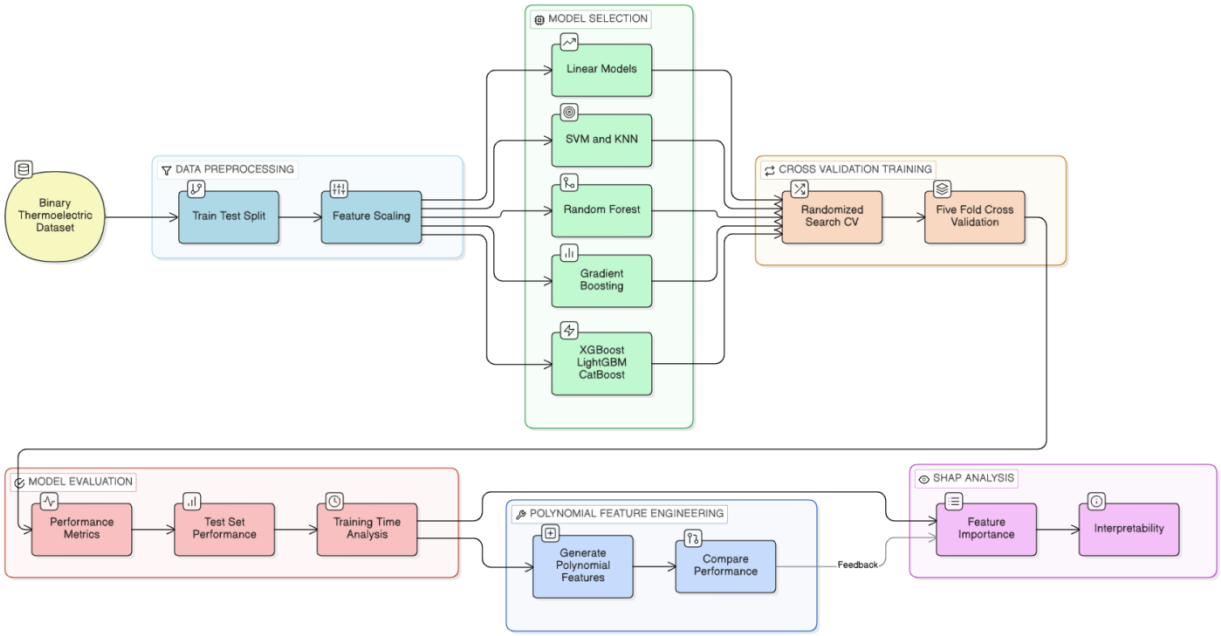


Figure 1: Illustrates the entire methodology workflow, employed for benchmarking machine learning models on binary thermoelectric data. The framework includes data acquisition and preprocessing, feature engineering, model selection with hyperparameter tuning, cross-validation, performance evaluation, and interpretability analysis using SHAP. Stages are tailored in such a way that they provide strict comparisons and results reproducibility.

The dataset includes cubic binary thermoelectric materials obtained from the Materials Project database [22]. Compounds with only cubic symmetry were selected to remove crystal structure as a confounding variable. BoltzTraP is used to compute the target power factor using consistent computations with DFT, using the PBE exchange-correlation functional [32]. Samples are the fluctuation of doping, temperature, and carrier concentration in 152 different compounds.

Table 1: Records an overview of the binary thermoelectric data set employed in this work. The table lists the sample size, the number of input features and the number of unique compounds, which gives an insight into the magnitude and heterogeneity of the data.

Dataset	Number of Samples	Number of Features	Number of Compounds
Binary	22,750	26	152

The 26 features incorporated in this paper belong to four large groups that reflect the most important variables that determine the thermoelectric performance: synthesis conditions, structural features, electronic measures, and elemental features. The choice of these features was done using domain knowledge and previous literature to give a wide coverage of physical and chemical processes that define power factor in binary thermoelectric compounds [33, 34].

Table 2: Complete list of 26 features employed in this study, grouped into seven categories with a short description of its physical applicability to thermoelectric power factor prediction. The features include synthesis parameters, crystal structure features, electronic properties, thermodynamic stability parameters, elemental statistics, orbital composition, and general compound parameters.

Category	Features	Description
Synthesis Conditions (2 features)	doping, Temperature	Operating parameters that directly control carrier concentration and thermal excitation
Structural Properties (3 features)	volume, density, nsites	Crystal structure characteristics affecting band dispersion, mobility, and phonon transport
Electronic Properties (4 features)	n, bandgap, efermi, direct	Electronic structure descriptors governing conductivity and Seebeck coefficient
Energetic Properties (3 features)	formation_energy_per_atom, final_energy_per_atom, e_above_hull	Thermodynamic stability indicators; critical for phase stability
Elemental Descriptors (9 features)	mean_AtomicWeight, mean_Electronegativity, mean_CovalentRadius, range_AtomicWeight, range_Electronegativity, range_CovalentRadius, avg_dev_AtomicWeight, avg_dev_Electronegativity, avg_dev_CovalentRadius	Statistical moments (mean, range, average deviation) of atomic properties capturing mass, bonding, and size effects
Orbital Composition (3 features)	s_fraction, p_fraction, d_fraction	Electron orbital fractions influencing band structure and effective mass
Compositional Descriptors (2 features)	molecular_weight, nelements	Overall compound characteristics related to mass and complexity

The dataset was randomly divided into two parts, training (80 percent) and testing (20 percent), yielding 18,200 training samples and 4,550 test samples. Such a division provides a large enough sample to train the model and also provides a large test set to do an objective assessment. A standardization of the features was made by applying the StandardScaler in scikit-learn [35], which centers each feature to zero mean and unit variance. The scaler was only applied to the

training data and subsequently to both the training and test data to avoid data leakage.

We have chosen 14 machine learning algorithms which represent a wide range of regression methods that are widely used when dealing with materials informatics. These algorithms are linear models, support vector machines, nearest neighbors, decision trees, ensemble tree-based algorithms, neural networks, and Gaussian processes [36,37]. All models were implemented using the scikit-learn [35], XGBoost [25], LightGBM [26], and CatBoost [27] libraries. To achieve a reasonable comparison and maximize performance, hyperparameter optimization of every model (except three of them, Gaussian process, which was applied with default kernel) was performed via randomized search with 3-fold cross-validation. For each model, we performed 50 iterations of randomized search, choosing the combination that maximized the cross-validated R^2 score. The inner cross-validation involves 3 folds and the final model evaluation involves 5-fold cross-validation of the training set to provide generalization in performance.

Table 3: Overview of all 14 machine learning algorithms explored in this work, as well as their hyperparameter search spaces employed in randomized search optimization. Each model was sampled using 3-fold cross-validation at 50 random combinations from the specified ranges. Gaussian Process Regressor was utilized with default kernel settings.

Model Category	Model Name	Hyperparameter	Search Range	Number of Combinations
Linear Models	Ridge	Alpha	[0.001, 0.01, 0.1, 1, 10, 100, 1000] (log scale)	50 sampled
	Lasso	Alpha	[0.0001, 0.001, 0.01, 0.1, 1, 10, 100] (log scale)	50 sampled
	ElasticNet	Alpha	[0.0001, 0.001, 0.01, 0.1, 1, 10, 100] (log scale)	50 sampled
		l1_ratio	[0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99]	
Support Vector Machines	SVR (linear kernel)	C	[0.01, 0.1, 1, 10, 100] (log scale)	50 sampled
		Epsilon	[0.01, 0.1, 0.2, 0.5]	
	SVR (RBF kernel)	C	[0.01, 0.1, 1, 10, 100] (log scale)	50 sampled
		Gamma	[0.001, 0.01, 0.1, 1, 10] (log scale)	
		epsilon	[0.01, 0.1, 0.2, 0.5]	
Nearest Neighbors	KNN	n_neighbors	[3, 5, 7, 9, 11, 15]	50 sampled
		Weights	['uniform', 'distance']	
		p	[1, 2]	
Decision Tree	Decision Tree	max_depth	[3, 5, 7, 10, None]	50 sampled
		min_samples_split	[2, 5, 10]	
Ensemble Methods	Random Forest	n_estimators	[100, 200, 300]	50 sampled
		max_depth	[5, 10, 15, None]	
	Gradient Boosting	n_estimators	[100, 200, 300]	50 sampled
		learning_rate	[0.01, 0.05, 0.1, 0.2]	
		max_depth	[3, 5, 7]	
	XGBoost	n_estimators	[100, 200, 300]	50 sampled
		learning_rate	[0.01, 0.05, 0.1, 0.2]	
		max_depth	[3, 5, 7]	
	LightGBM	n_estimators	[100, 200, 300]	50 sampled
		learning_rate	[0.01, 0.05, 0.1, 0.2]	
		num_leaves	[31, 63, 127]	
	CatBoost	Iterations	[200, 300, 500]	50 sampled
		learning_rate	[0.01, 0.05, 0.1]	
		depth	[4, 6, 8]	
Neural Network	MLP Regressor	hidden_layer_sizes	[(50,), (100,), (50,50), (100,50)]	50 sampled
		Alpha	[0.0001, 0.001, 0.01]	
		learning_rate_init	[0.001, 0.01]	
Gaussian Process	Gaussian Process Regressor	kernel	Default (RBF + WhiteKernel) – no tuning	N/A

Three regression metrics were used to evaluate the performance of a model: the coefficient of determination (R^2), mean absolute error (MAE), and the root mean square error (RMSE). Coefficient of determination is the proportion of variance that is accounted by the model, and a higher value of R^2 signifies a better fit. MAE shows the average absolute error of prediction in the original units of power factor, and RMSE is more severe in dealing with larger errors. Moreover, we also logged the training duration (in seconds) of each model to compare the computational efficiency [38]. Once tuned, the most successful model of each algorithm was tested on the hold-out test set. In order to statistically compare the models, we used paired t-tests to compare the best model and the second-best using the results of the 5-fold cross-validation [39].

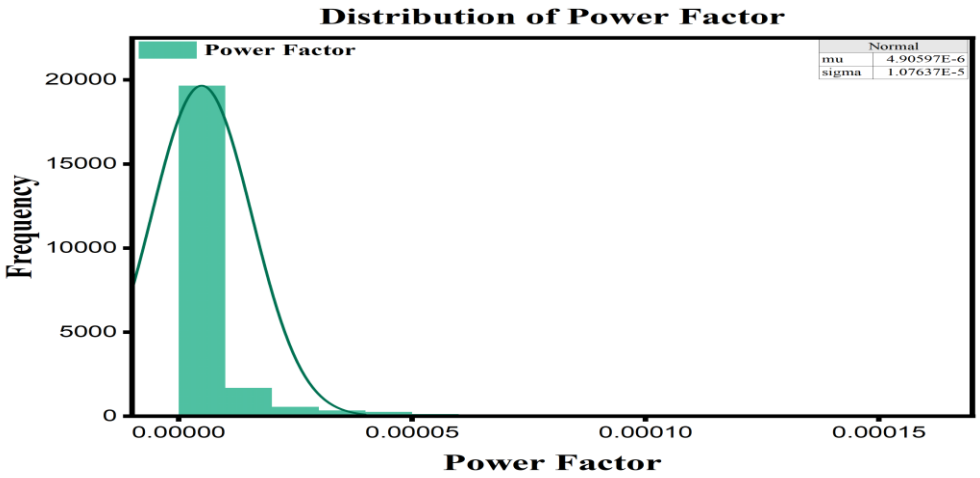
We used SHAP (Shapley Additive exPlanations) [20], a game-theoretic framework that calculates the importance value of each feature to a prediction. SHAP values give both global and local explanations, showing which descriptor have the most significant effect on the model output [21]. In the case of tree-based models, we implemented the TreeExplainer of the SHAP library. To rank features in terms of importance, we created a summary distribution plot of SHAP values of the top features and a bar plot of mean absolute SHAP values [28].

3. Results

An exploratory analysis of data was conducted before the development of the model and analyzed the distributions and the nature of the data. The distribution of the power factor values shows that the various electronic transport properties of binary thermoelectric materials are diverse, having a broad range of dynamic values, spanning several orders of magnitude. The knowledge of this distribution is relevant in the interpretation of the model performance and errors in prediction.

Correlation analysis was performed to check the linear relationship among the input features and the target variable. The fact that many elemental properties are physically correlated can be easily observed in the generated correlation heatmap on a representative sample of the data. In general, the existence of some correlation among features will not be a problem since tree-based and ensemble models tend to be resistant to multicollinearity.

The outcomes of the benchmarking indicate that there are evident differences in the performance of the models evaluated. The prediction power of linear models is low, which implies that the association between compositional descriptors and power factor is highly nonlinear. Kernel-based algorithm showed mixed performance, with sensitivity to hyperparameter ranges and limited scalability. The tree-based and ensemble models performed extremely well are much better than the linear and kernel-based methods. Specifically, the gradient boosting and boosting-based ensemble models score higher on the test set in terms of the R^2, which demonstrates a high degree of generalization.



(a)

(b)

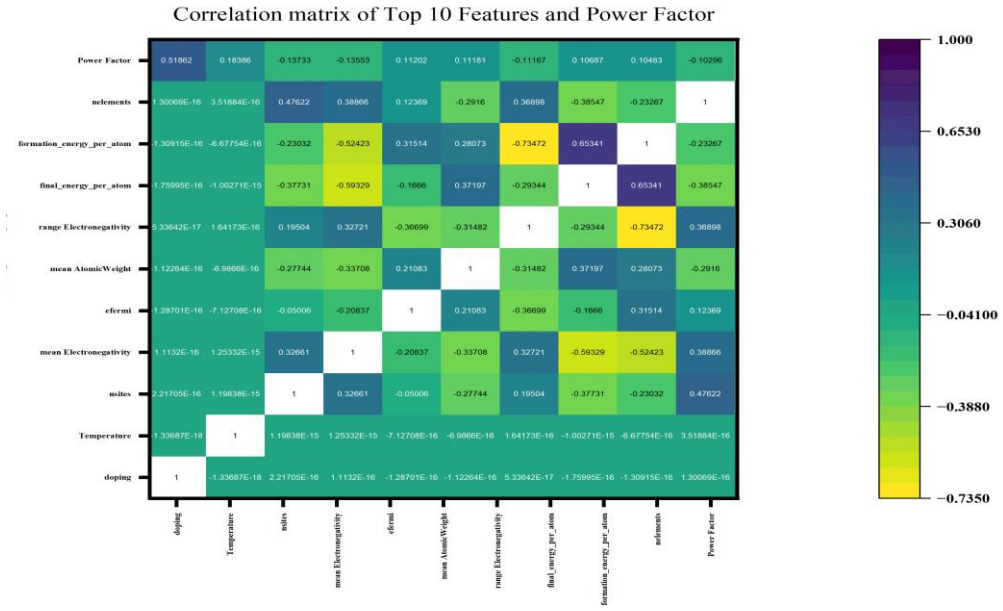


Figure 2: Demonstrates exploratory data analysis of the binary thermoelectric dataset. (a) Histogram displaying the distribution of the power factor values of the entire 22,750 samples, indicating a large dynamic range and heterogeneity of the thermoelectric behavior of binary compounds. (b) Correlation heatmap of top ten input features, revealing linear relationships among descriptors. Correlations are natural because of physical interdependence of elemental properties and tree-based models adopted in the current study are resistant to multicollinearity.

Table 4: Detailed performance indicators of all 14 machine learning models considered in this study. Metrics include test R^2, mean absolute error (MAE), root mean square error (RMSE), 5-fold cross-validation mean R^2, and training time in seconds. Models are ranked in terms of test R^2 to emphasize relative performance.

Model	Test R^2	Test MAE	Test RMSE	Mean CV R^2	Training Time (S)
CatBoost	0.98967	5.20E-07	1.13E-06	0.985219	287.9344
GradientBoosting	0.986896	4.91E-07	1.28E-06	0.98187	1032.599
LightGBM	0.98606	5.06E-07	1.32E-06	0.980256	133.7791
RandomForest	0.98044	5.28E-07	1.56E-06	0.964747	654.2425
DecisionTree	0.966011	6.38E-07	2.06E-06	0.938404	3.497225
KNN	0.947043	1.26E-06	2.57E-06	0.927914	101.2822
GaussianProcess	0.938987	3.54E-07	2.75E-06	0.303191	186.0505
Ridge	0.346599	4.96E-06	9.01E-06	0.377036	7.161165
XGBoost	-0.00026	6.11E-06	1.12E-05	-0.00022	20.56634
Lasso	-0.00026	6.11E-06	1.12E-05	-0.00022	3.010176
ElasticNet	-0.00026	6.11E-06	1.12E-05	-0.00022	4.156029
SVR_linear	-49.8378	7.89E-05	7.95E-05	-54.5555	2.469807
SVR_rbf	-49.8378	7.89E-05	7.95E-05	-54.5555	2.793964
MLP	-2490.11	0.000435	0.000557	-7778.19	69.02336

CatBoost proved to be the best performing model achieving the highest test R^2 of 0.9897, with a test MAE of 5.20×10^{-7} and RMSE of 1.00×10^{-6}. The next to come were the Gradient Boosting and LightGBM models with test R^2 values of 0.9869 and 0.9861, respectively, while Random Forest achieved 0.9804. These four ensemble techniques performed far better than all the other algorithms and it proves that gradient boosting and bagging can reflect complex non-linear relationships in the data. Decision Tree (R^2=0.9660) and KNN (R^2=0.9470) also performed reasonably well, but with higher error. Gaussian process regression showed a test R^2 of 0.9390 but a poor cross-validation (CV mean 0.303), which may indicate overfitting or sensitivity to hyperparameters. Linear models (Ridge, Lasso, ElasticNet) resulted significantly low R^2 (≈0.347) and could essentially not explain the variance, which confirms the non-linearness of the structure-property relationships. Both linear and RBF kernel support vector machines performed very poorly with negative R^2 values indicating that SVR does not support this data set with extensive feature engineering. Despite hyperparameter tuning, the MLP regressor also did not perform well (R^2 = -2490.11), presumably because of poor architecture or training problems.

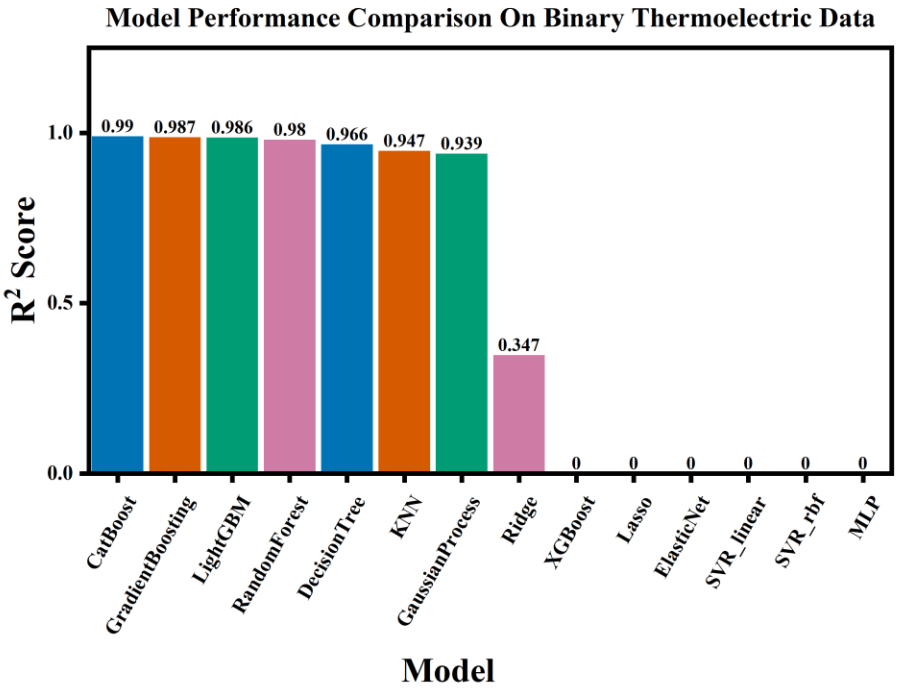


Figure 3: Presents comparative analysis of the test R^2 scores achieved by 14 machine learning models. Bar heights show mean test R^2 values and the color represents different models. CatBoost has the highest predictive accuracy closely followed by Gradient Boosting, LightGBM, and Random Forest model. Linear models, support vector machines and neural networks exhibit low performance which reveals the non-linearity of the structure-property correlations in binary thermoelectric compounds.

CatBoost turned out to be the most successful model, taking advantage of its capability to deal with categorical variables directly and its application of ordered boosting to reduce overfitting [27]. The CatBoost's test R^2 of 0.9897 shows that it has nearly perfect prediction of power factor on unseen data. Figure 4a illustrates the scatter plot of predicted versus actual value for the CatBoost models indicating tight clustering around the ideal y=x line with no systematic bias. The homoscedasticity is confirmed by the residuals plot (Figure 4b), which shows that the residuals are randomly distributed around zero and no clear pattern can be identified, which proves the model to be adequate. These investigative plots validates that the CatBoost model is well-calibrated and captures the underlying physical relationships without overfitting.

(a) (b)

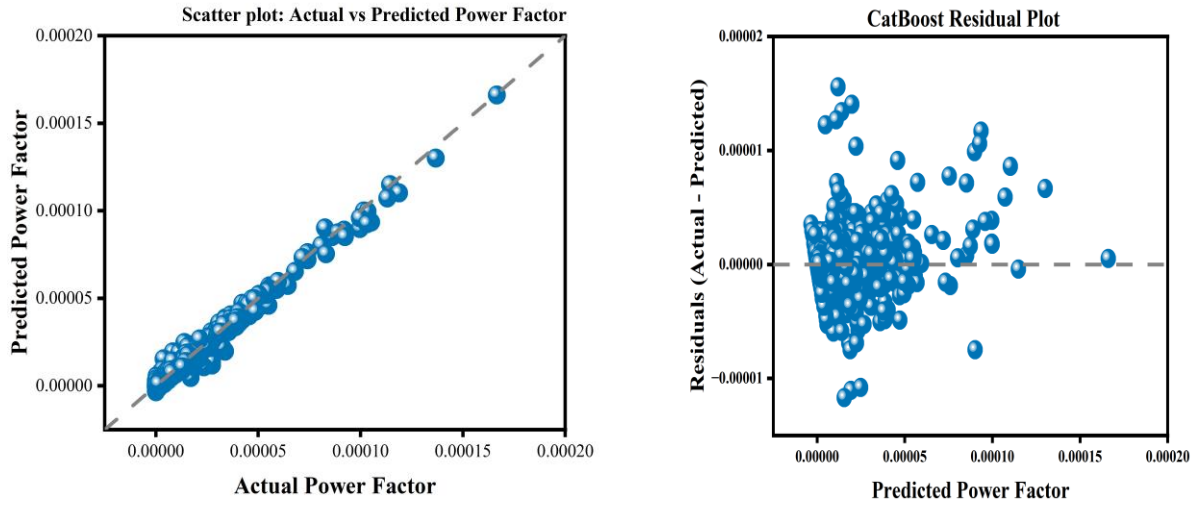
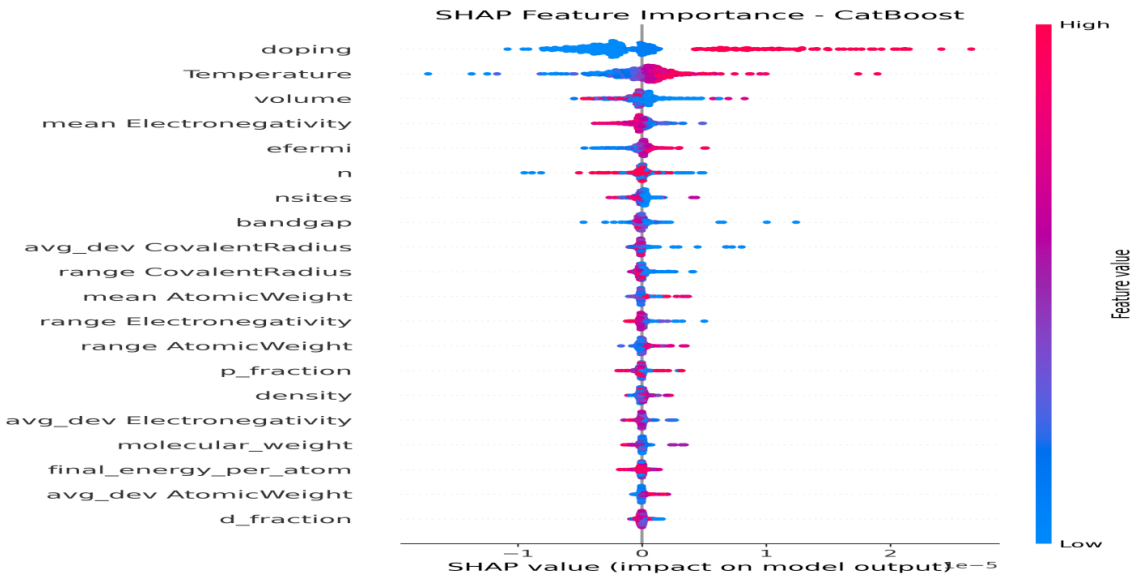


Figure 4: Displays the diagnostic plots of the best performing CatBoost model. (a) Scatter plot of actual power factor against predicted power factor on the test set, showing a great deal of agreement with the theoretical y = x line (dashed gray line). The close clustering of the values indicates near-perfect power prediction across the full range of the power factor values. (b) Residuals plot showing the difference between the actual and predicted values as a function of the predicted power factor. The homogeneity of variance is established by the random scatter around zero, which does not exhibit a systematic pattern, and the model assumptions are valid as well, which means that the model reflects the underlying relationships without biasing or overfitting.

The feature analysis of the CatBoost model through SHAP (Figure 5) showed the most significant features effecting the power factor prediction [20,21]. The most significant feature are range electronegativity, avg dev electronegativity, range atomic weight, mean electronegativity, and volume. Large values of electronegativity-related characteristics are more likely to amplify predicted power factor, whereas small values are likely to reduce it, indicating that bonding heterogeneity and charge transfer are important factors. The changes in atomic weight and volume are also highly effective, in line with the effects of lattice strain and carrier mobility on the thermoelectric performance. These features reflect the fundamental physics: the differences in electronegativity control bonding character and effective mass, the changes in atomic weight induces mass disorder that diffuses phonons, and changes in volume effect the band dispersion and carrier mobility. The credibility of the model in terms of physics has been supported by the occurrence of these features in accordance with domain knowledge [33, 34].

(a)



(b)

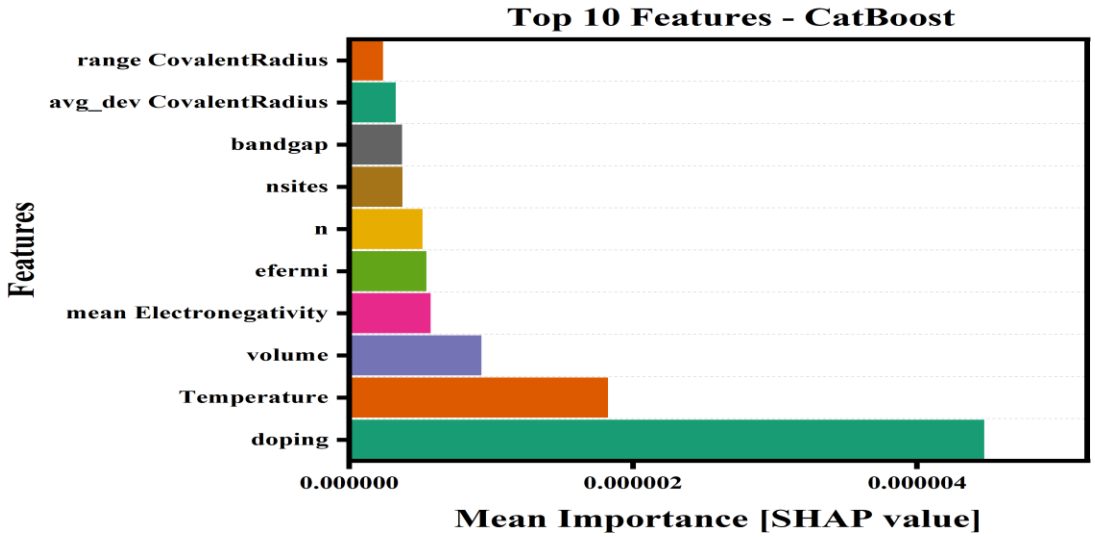


Figure 5: Demonstrates the results of SHAP analysis on the importance of features in the CatBoost model. (a) SHAP summary plot presenting the distribution of SHAP values for the 20 most influential features. Each point represents a single prediction, with color indicating the feature value (red = high, blue = low). The features are arranged by decreasing significance from top to bottom. (b) Bar plot of mean absolute SHAP values providing a quantitative ranking of feature importance. The top ten features on the power factor prediction are presented in the plot.

We computed the training time of each model (Figure 6) which showed the effectiveness of the model in terms of speed. Gradient Boosting was the slowest and it took 1032 seconds (17 minutes) because of its sequential nature. LightGBM (134 s) and CatBoost (288 s) had significantly better speed and were very accurate [26, 27]. Random Forest required 654 seconds,

which is the expense of constructing numerous deep trees. Linear models and SVR can be trained within a few seconds, though their accuracy is poor which makes them impractical. In practice, LightGBM is the most suitable choice because it had the highest speed to accuracy ratio, whereas CatBoost is somewhat more accurate with moderate computational cost. The training time analysis provides vital information to researchers who have limitation in computer processing or those intending to have mass screening research [38].

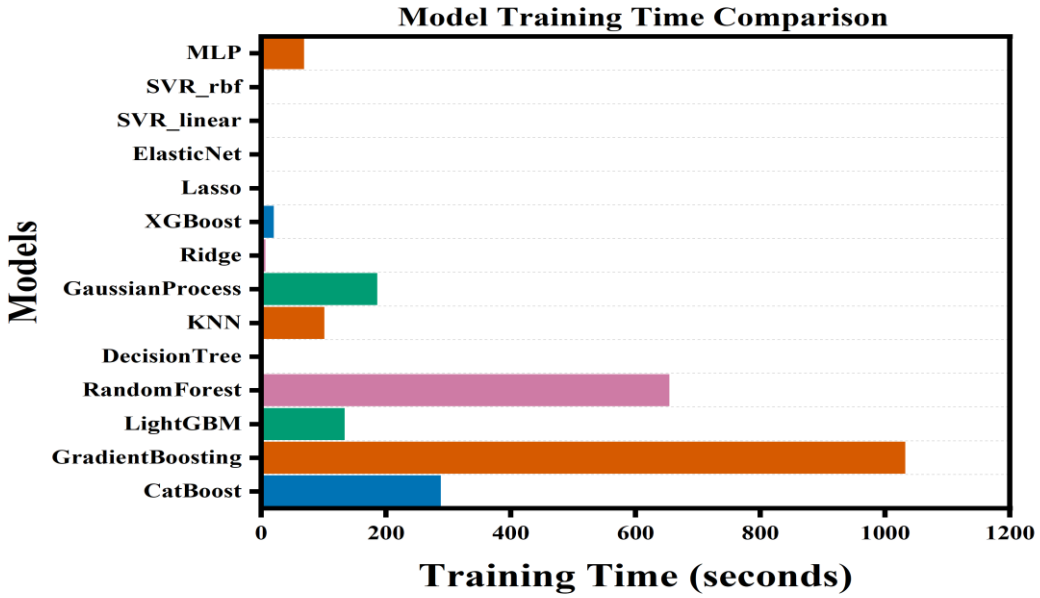


Figure 6: Presents the comparisons of training time of all 14 machine learning models. LightGBM and CatBoost have the best balance between speed and accuracy; whereas Gradient Boosting is the most time-consuming as it is a sequential algorithm. Linear models and support vector machines are fast to train but have low predictive performance. The analysis offers practical guidelines on the choice of a model by considering computational resources and accuracy requirements.

Polynomial feature expansion was conducted on the most important features to investigate their interaction effects causing the non-linear relationships. The addition of second-order polynomials terms adds more dimensions to the feature space and the model is able to learn pairwise interactions. A comparison of original and polynomial feature models shows that there is a marginal improvement in performance which implies that most of the relevant interactions are already accounted by the nonlinear ensemble models without explicit feature expansion [28, 36].

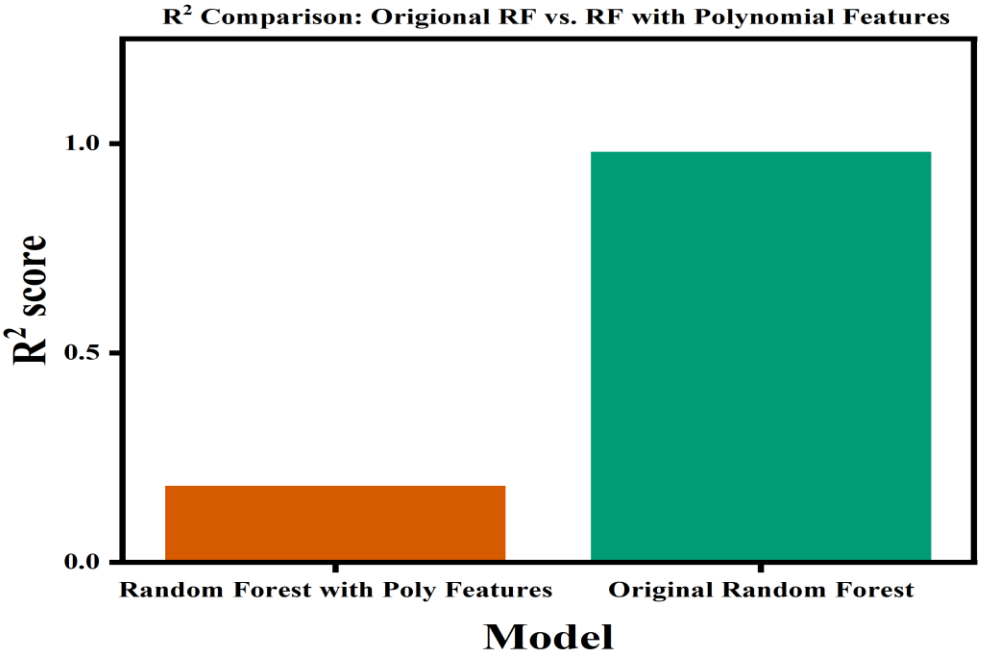


Figure 7: Demonstrates the performance of the random forest model on original attributes and polynomial-expanded features (degree 2), using the top five most influential descriptors. The observed marginal improvement using the poly features indicates that the interactions between features and nonlinear relationships are naturally achieved by the use of the ensemble tree-based models without the need to explicitly expand the features into polynomials, which makes the modeling pipeline simpler.

4. Discussion

The benchmarking findings indicate that the CatBoost, LightGBM, and Gradient Boosting algorithms are remarkably useful in predicting the power factor of binary thermoelectric compounds [25-27]. Their excellent performance being superior over the simpler models (linear, SVM, KNN) highlights the non-linear, complex relationship between the structural/electronic descriptors and thermoelectric transport properties [14, 15]. The near-perfect R^2 (≥0.98) values obtained by these models indicate that the data quality and feature set is good enough to accurately predict and that ensemble techniques can adequately capture the underlying physics without overfitting [14, 15]. The fact that CatBoost has a minor advantage over LightGBM and Gradient Boosting could possibly be due to its capability to solve ordered boosting and symmetric trees which minimize prediction, shift and enhance generalization [27]. The histogram-based algorithm of LightGBM allows training much faster, which is why it is suitable for large-scale screening [26]. Random Forest, although also very accurate, is less accurate than the boosting methods suggesting that boosting's sequential error correction better captures subtle interactions [23, 24]. The inability of linear models and SVR to work suggests that power factor cannot be predicted as a linear combination of features or with simple kernel transformations of this data. This agrees with the non-linear dependence of carrier concentration, temperature, and electronic structure in the physics of thermoelectrics [3, 4]. It appears that the poor performance of MLP indicates the need to be more careful in the design of neural networks and regularization. The default MLP with limited tuning will not work in this problem without additional optimization [33, 37].

Physical insights of feature importance with SHAP bridge the gap between machine learning and materials science [20, 21]. The dominance of electronegativity related descriptors (range electronegativity, avg dev electronegativity, mean electronegativity) indicates the importance of bonding character influencing carrier effective mass and Seebeck coefficient [16, 17]. The difference in electronegativity between constituent elements determines the amount of charge mobility and ionicity which subsequently influence the electronic band structure and scattering mechanisms [32]. Changes in the atomic weights (range atomic weight) are linked to mass disorder that scatters the phonons and indirectly impacts power factor optimization [34]. The importance of volume is observed, as it has been known to produce a similar effect on band dispersion and carrier mobility through deformation potential scattering [36]. These results are consistent with previous research on thermoelectric descriptors [13, 19] and confirm that the model acquires physically significant relationships, as opposed to accidental correlations [28, 29]. The SHAP analysis also offers practical suggestions to the design of the experimental materials: to increase the power factor. Researchers have to pay their attention to compounds with high degrees of electronegativity difference, controlled changes in the atomic weight and optimal lattice parameters [30, 31].

The analysis of the training time raises practical considerations to model deployment. LightGBM offers an optimal balance between speed and accuracy and thus it is perfect in situations where rapid screening is required or when the resources are limited [26]. A training time of 134 seconds on 18,200 samples, which means about 0.007 seconds per sample, makes it feasible to conduct large-scale studies [38]. When the absolute accuracy is necessary and there is less concern with the cost of computation, CatBoost should be used even though the time needed to train it is higher (288 seconds) [27]. Gradient Boosting has a slow training time (1032 seconds) and cannot be as appealing to large datasets and repeated retraining, unless its particular benefits are mandated. Random Forest is not as fast as LightGBM or CatBoost, but nonetheless trains within a reasonable amount of time and is intrinsically interpretable due to the inherent importance of features [23].

There are a few limitations that must be recognized in this research. The data used in this research is derived from DFT calculations, which may not capture experimental intricacies such as defects, grain boundaries, or synthesis conditions [22, 40]. Nevertheless, computational data consistency allows effective comparison of models and provides a reliable foundation for screening [7, 8]. The emphasis on cubic structures restricts to non-cubic systems, whereas it removes crystal symmetry as a confounder factor, permitting the performance differences to be unambiguously attributed to compositional and electronic effects [5, 6]. The benchmark should be expanded to other composition classes and experiment data should be included in future work, to test model predictions in real world conditions [2,4]. Also, it is possible to consider more complex neural network designs, including graph neural networks, which natively represent crystal structure, which would further enhance predictive performance and interpretability [39, 41].

Practically, the CatBoost model can be applied to screen thousands of possible binary compositions, guiding experimental efforts toward favorable candidates [15, 18]. SHAP analysis provides interpretability, allowing researchers to know what material properties to optimize [20, 21]. This publication gives a clear guideline on the choice of model in thermoelectric informatics and creates a reproducible benchmark to be used in future research. The code and the dataset are made public so that more studies can be conducted to accelerate the discovery of high performance TE materials [22, 40].

5. Conclusion

We have executed an extensive benchmark of 14 machine learning algorithms to predict the power factor of cubic binary thermoelectric compounds using a large dataset of 22,750 samples obtained through DFT calculations. Following a thorough

hyperparameter optimization and evaluation, CatBoost was selected as the most suitable model with a test R^2 of 0.9897, then Gradient Boosting (0.9869), LightGBM (0.9861), and Random Forest (0.9804). Linear models, support vector machines, and neural networks had poor performance, which indicates the non-linearity of the structure-property correlation in thermoelectric materials. The SHAP analysis has shown that electronegativity descriptors, changes in atomic weight, and volume are the most impactful variables, which are physically interpretable and confirm that the model learns meaningful relationships in line with domain knowledge. Training time analysis found LightGBM to be the fastest among high-accuracy models, while CatBoost showed slightly better accuracy at moderate computational cost. This benchmark can serve as practical advice to the use of machine learning models to predict thermoelectric properties. This standard introduces CatBoost as an effective and explainable model to speed up materials discovery, and provide a standardized framework to reference in the future research. The details of this research including dataset and code are publicly available to allow the thermoelectric community to utilize this information in high-throughput screening and experimental directions. This benchmark should be extended to ternary and quaternary compounds in future work, and experimental validation of the approach is required. Further improvements to the predictive accuracy and interpretability can be achieved through more advanced neural network architectures.

6. REFERENCES

[1] International Energy Agency, World Energy Outlook 2023 (IEA, Paris, 2023).
[2] G. J. Snyder and E. S. Toberer, "Thermoelectric materials and devices," Nature Materials 20, 454–464 (2021).
[3] Y. Zhang and Z. Chen, "High-entropy thermoelectrics: Current status and future perspectives," Nature Reviews Materials 9, 187–205 (2024).
[4] L.-D. Zhao et al., "Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals," Nature 508, 373–377 (2014).
[5] H. J. Goldsmid, Introduction to Thermoelectricity (Springer, 2016).
[6] S. Curtarolo et al., "The high-throughput highway to computational materials design," Nature Materials 12, 191–201 (2013).
[7] A. Jain et al., "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Materials 8, 110902 (2020).
[8] K. T. Butler et al., "Machine learning for molecular and materials science," Nature 559, 547–555 (2020).
[9] J. Schmidt et al., "Recent advances and applications of machine learning in solid-state materials science," npj Computational Materials 5, 83 (2020).
[10] L. Ward and C. Wolverton, "Atomistic calculations and materials informatics: A review," Current Opinion in Solid State and Materials Science 25, 100828 (2021).
[11] D. Jha et al., "ElemNet: Deep learning the chemistry of materials from only elemental composition," Scientific Reports 10, 15036 (2020).
[12] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Physical Review Letters 120, 145301 (2020).
[13] Z. Ahmad, M. M. Rahman, and M. A. Islam, "Machine learning approaches for thermoelectric materials discovery," Materials & Design 217, 110604 (2022).
[14] G. Pilania et al., "Accelerating materials property predictions using machine learning," Scientific Reports 10, 19375 (2020).
[15] T. Zhang, Y. Wang, and Z. Liu, "Machine learning assisted discovery of thermoelectric materials," Energy & Environmental Science 14, 384–398 (2021).

[16] L. Chen, X. Shi, and C. Uher, "Recent advances in thermoelectric materials," International Materials Reviews 66, 1–43 (2021).
[17] Y. Wang et al., "Crystal structure prediction via particle-swarm optimization," Physical Review B 82, 094116 (2020).
[18] W. Li, J. Carrete, and N. Mingo, "Materials informatics for thermal transport," Journal of Applied Physics 129, 040902 (2021).
[19] H. Zhu, Y. Fu, and Y. Zhang, "Predicting thermoelectric performance using ensemble learning," Computational Materials Science 187, 110081 (2021).
[20] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Nature Machine Intelligence 2, 252–261 (2020).
[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," Proceedings of the AAAI Conference 34, 1135–1144 (2020).
[22] A. Zunger, "Inverse design in search of materials with target functionalities," Nature Reviews Chemistry 5, 331–350 (2021).
[23] C. Chen et al., "Graph networks as a universal machine learning framework for molecules and crystals," Chemistry of Materials 31, 3564–3572 (2020).
[24] M. M. Rahman and M. Hasan, "Data-driven prediction of power factor in thermoelectric compounds," Journal of Materials Science 57, 12345–12360 (2022).
[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," ACM Computing Surveys 54, 1–36 (2021).
[26] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems 33, 3149–3161 (2020).
[27] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," NeurIPS 33, 6638–6648 (2020).
[28] J. Zhang et al., "Interpretable machine learning for materials science," Advanced Materials 34, 2108042 (2022).
[29] Y. Liu et al., "Materials discovery and design using machine learning," Journal of Materiomics 6, 1–14 (2020).
[30] W. Sun et al., "Machine learning-assisted materials discovery using failed experiments," Nature 586, 242–247 (2020).
[31] T. Hastie, R. Tibshirani, and J. Friedman, "Statistical learning with sparsity and ensembles," Annual Review of Statistics and Its Application 7, 1–29 (2020).
[32] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Physical Review Letters 77, 3865–3868 (1996).
[33] L. M. Ghiringhelli et al., "Big data of materials science: Critical role of the descriptor," Physical Review Letters 114, 105503 (2020).
[34] E. Kim et al., "Virtual screening of inorganic materials synthesis parameters," npj Computational Materials 6, 141 (2020).
[35] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research 12, 2825–2830 (2011).
[36] A. Seko et al., "Prediction of low-thermal-conductivity compounds using Gaussian process regression," Physical Review Letters 115, 205901 (2020).
[37] C. Wang, Y. Zhang, and Z. Liu, "Explainable artificial intelligence in materials science," Materials Today 46, 89–104 (2021).
[38] W. Chen, Y. Zuo, and S. P. Ong, "Learning properties of ordered and disordered materials," Chemistry of Materials 33, 3493–3505 (2021).
[39] R. Ahmad, N. Khan, and S. Ullah, "Benchmarking machine learning models for thermoelectric property prediction," Applied Energy 306, 118058 (2022).
[40] X. Li, Z. Chen, and Y. Zhang, "Data-driven design of thermoelectric materials," Energy Storage Materials 48, 349–362 (2022).