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

MACHINE LEARNING IN THERMO-ELECTROCHEMICAL CELL MATERIAL IDENTIFICATION: FEATURE SELECTION AND ANALYSIS

Riaz Muhammad[1], Asad Ullah[2], Sana Ullah[*3], Ahmad Nisar[1], Fawad Ali[1], Samahat Ullah[1], Umair Ahmad[2]

[1]*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]*Department of Mechanical Engineering, University of Engineering and Technology, Peshawar 25120, Pakistan*

[*3]sanaullah.uet31@gmail.com

DOI: https://doi.org/10.5281/zenodo.19083658

Keywords
low grade energy harvesting, thermoelectric material, the rmoelectrochemcial cell, machine learning.

Article History
Received: 18 January 2026
Accepted: 03 March 2026
Published: 18 March 2026

Corresponding Author: *
Sana Ullah

Abstract

This research shows the use of Machine Learning models to predict the seebeck coefficient, feature selection, and analysis of ionic thermoelectric material using different feature selection strategies. The approach comprises data collection of ionic thermoelectric material including (matrix + ion donor) combinations with features and target (seebeck coefficient) from different publish papers. Applying different feature selection strategies with cross validation mean absolute error and mean square error to determine optimum feature subset which improve accuracy and generalization of model. Subsequently multiple models were trained on each respective feature subset. To evaluate their performance Decision Tree model was the best model exhibit high R^2 and low mean absolute error and root mean square error trained on univariate selected feature subset. It is revealed that seeebck coefficient is dominated over few strong predictors, adding more features reduce the accuracy of model and introduce noise or overfitting. This finding also expose that reduction of features significantly accelerates the discovery of matrix, ion donor combinations for thermoelectrochemical cell. Further to achieve additional superior robustness and generalization the top selected model was subjected to hyper parameter optimization process. SHapley Additive exPlanations (SHAP) and correlation analysis was performed to interpret model behavior, determined most influential features, and relationship between features and target seebeck coefficient. FractionCSP3 of the matrix and NumRotatableBonds of the ion donor were identified the most important features using SHAP analysis. It is also found that FractionCSP3 of the matrix show positive correlation, while NumRotatableBonds of the ion donor exhibit negative correlations with seebeck coefficient. The Decision Tree models trained on univariate selected features predicted many promising combinations, especially polyurethane-based, cellulose-based, and PVA-based along with gelatin ionogel, and PAM hydrogel systems with predicted Seebeck coefficients up to 42.8 mV/K.

Introduction

Global energy demand for renewable, clean and long lasting maintainable energy continues to escalate as societies required for reducing their dependence on fossil fuels and moderating those changes which impacts the climate [1]. Yet a substantial fraction of globally generated energy is dissipated as low-grade waste heat [2]. Some traditional technology like steam turbine and sold-state thermoelectric generator having a great contributions in this temperature regime, leaving this vast energy largely untapped [3]. Subsequently, efficient technologies capable of converting low temperature gradient into electricity are essential for improving sustainable energy development [4]. These devices are costly, restricted to elevated-temperature applications, and limited by low efficiency, rare material availability, and high production expenses. These limitations create a reasonable motivation to find alternative technologies for conversion low-grade thermal energy into electrical energy [5]. Thermoelectrochemical cell (TECs) recently positioned themselves as one of the significant technology to overcome the challenges and issues, as they convert low temperature difference into electrical power through entropy-driven redox processes.

TECs consist of two electrodes (Hot and cold) immersed in a redox-active electrolytes [6]. When temperature gradient is applied between the electrodes establish a temperature-induced entropy generating potential difference. The temperature dependent change in electrode potential can be express using the seebeck coefficient (Si = ΔV/ΔT) which is define as the voltage produced per unit temperature difference, serves as the key performance parameter of TEC systems [7]. Their simple structure, mechanical flexibility, environmental compatibility, low-cost material and higher seebeck coefficient at low temeperture gradient make TECs prominent low-grade heat harvesting technology compare to other. vApart from the conventional waste heat recovery, TECs also find special interest in wearable electronic system as they provide long lasting power to these devices [8]. Wearable electronics are becoming essential components of modern healthcare, enabling continuous physiological monitoring, real time diagnostics, and seamless data communication [9] However, despite these benefits, the optimization of TEC performance is still a challenge. This is because the ionic seebeck, is strongly influenced by molecular structure, ion mobility, charge redistribution and solvation entropy. Consequently identifying optimal matrix ion donor combination is experimentally costly, time consuming and challenging [10]. Despite progress in ionic thermoelectric (i-TE) materials, there remains no comprehensive machine learning approach for predicting accurately Si from fundamental molecular features. This gap highlights the need for machine-learning (ML) frameworks capable of accelerating material discovery and guiding the rational design of high performance.

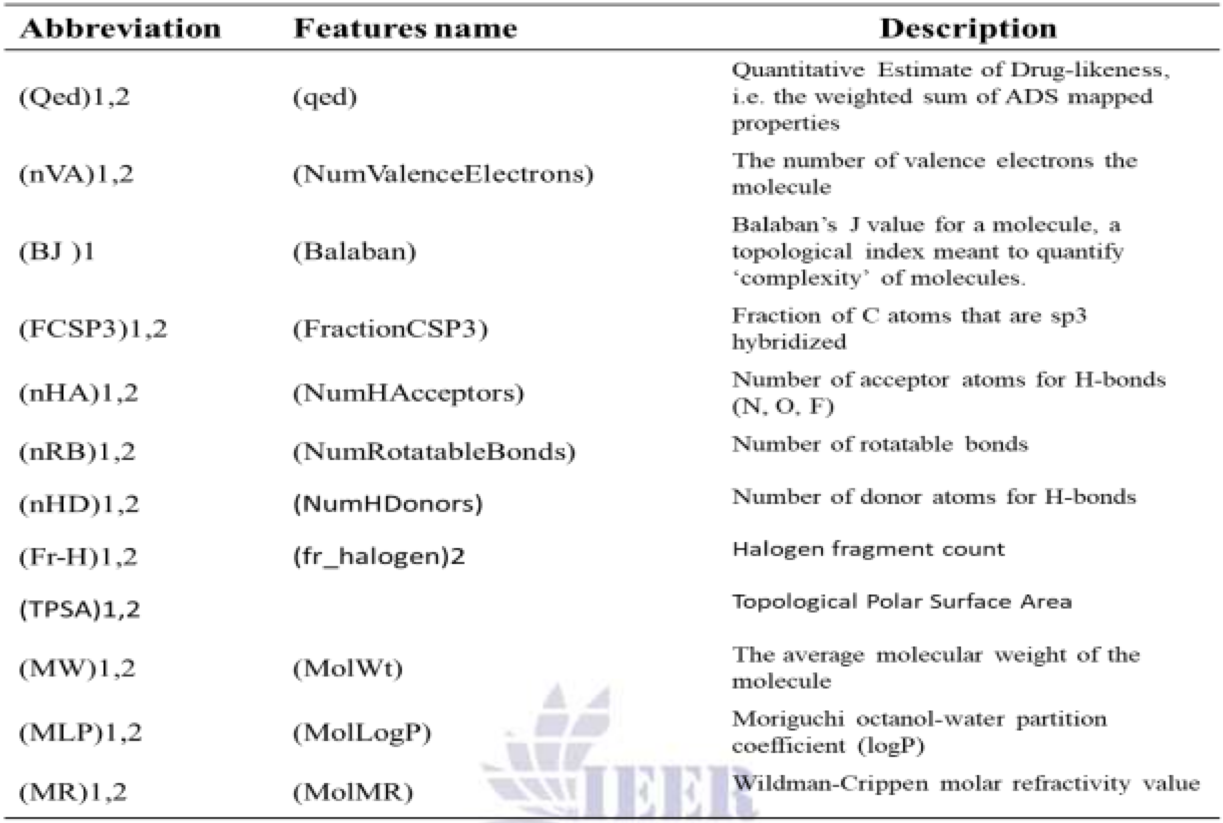
Recent innovations in i-TE materials have highlighted the vital role of molecular structure in defining thermos diffusion and entropy transport. For instance, Wu et al. (2025) have exploited innovative machine learning approaches to reveal key molecular parameters that govern thermoelectric properties and validated high-performance ionogel materials [11]. Likewise, Franco et al. (2022) have utilized molecular dynamics simulations to examine ion mobility and entropy transport in temperature gradients, underscoring the essential role of the Soret coefficient in thermoelectric properties [12]. These study reveal that molecular parameters have significant role in seebeck coefficient prediction. Their research provides a smooth way to understand the behavior and property of ionic TEC electrolytes. ML methods also demonstrated the significant potential broader material-science applications. The challenges in ML-based materials discovery were deliberated by Butler et al. (2018), who highlighted the transformative role of data-driven methods [13]. The role of ML in accelerating materials discovery through the combination of ML predictions and experimental validation was further demonstrated by Wu et al. (2019) [14]. In the situation of thermoelectrics, scalable ML frameworks integrating Random Forest, XGBoost, and deep neural networks for the prediction of thermoelectric power factors were established by Vaitesswar et al. (2024) [15]. In addition, interpretability methods such as SHapley Additive exPlanations (SHAP), introduced by Lundberg et al. (2020), enable the quantitative evaluation of feature contributions, thus enabling the transition from predictive modeling to physical understanding. Although these works provide a favorable route, there is still lack of ML method application used in TECs cell [16].

A comprehensive framework is required to evaluate feature selection, recognizes informative molecular descriptors, and explain the underline physical meaning of these within the context of TEC electrolytes. This research focus to develop a comprehensive ML framework to predict the seeebck coefficient of TECs cell electrolytes. A curated dataset of matrix-ion donor with 23 features and target value (seebeck coefficient) was constructed from experimentally reported literature. Subsequently multiple ML models were also train on optimum feature subsets derived from different feature selection methods. These models are evaluate and compare to get the best model exhibit high coefficient of determination (R^2) value and lower mean absolute error (MAE) and root mean square error (RMSE) values. By integrating predictive modeling with interpretability analysis, this work aims to provide both accurate Seebeck coefficient prediction and mechanistic understanding of molecular factors governing ionic thermodiffusion, thereby accelerating the development of high-performance TEC materials for low-grade heat harvesting applications.

Methodology:

A small curated dataset less than 100 samples of TECs was compiled from different peer-reviewed literature [17,21], including matrix and ion donor combinations mostly comprising one matrix and one ion donor and molecular descriptors potentially tied with i-TE performance shown in table 1. Each sample resembled to specific electrolyte system with an experimentally reported seebeck coefficient, defined as the target variable.

Table 1: list the 23 selected features for matrix and ion donor, having high influence on seebeck coefficient. Where subscript '1' denotes matrix-derived features and '2' denotes ion donor-derived features.



Abbreviation	Features name	Description
(Qed)1,2	(qed)	Quantitative Estimate of Drug-likeness, i.e. the weighted sum of ADS mapped properties
(nVA)1,2	(NumValenceElectrons)	The number of valence electrons the molecule
(BJ)1	(Balaban)	Balaban's J value for a molecule, a topological index meant to quantify 'complexity' of molecules.
(FCSP3)1,2	(FractionCSP3)	Fraction of C atoms that are sp3 hybridized
(nHA)1,2	(NumHAcceptors)	Number of acceptor atoms for H-bonds (N, O, F)
(nRB)1,2	(NumRotatableBonds)	Number of rotatable bonds
(nHD)1,2	(NumHDonors)	Number of donor atoms for H-bonds
(Fr-H)1,2	(fr_halogen)2	Halogen fragment count
(TPSA)1,2		Topological Polar Surface Area
(MW)1,2	(MolWt)	The average molecular weight of the molecule
(MLP)1,2	(MolLogP)	Moriguchi octanol-water partition coefficient (logP)
(MR)1,2	(MolMR)	Wildman-Crippen molar refractivity value

Preceding to the model training, the dataset undergoes through systematic processing to ensure statistical reliability and consistency. Numerical features were necessarily scaled to prevent imbalance, to avoid bias incomplete entries were removed and units were standardized across all variables. The output variable, namely the seebeck coefficient, was subjected to a logarithmic transformation to reduce variance and improve data stability for ML model training.

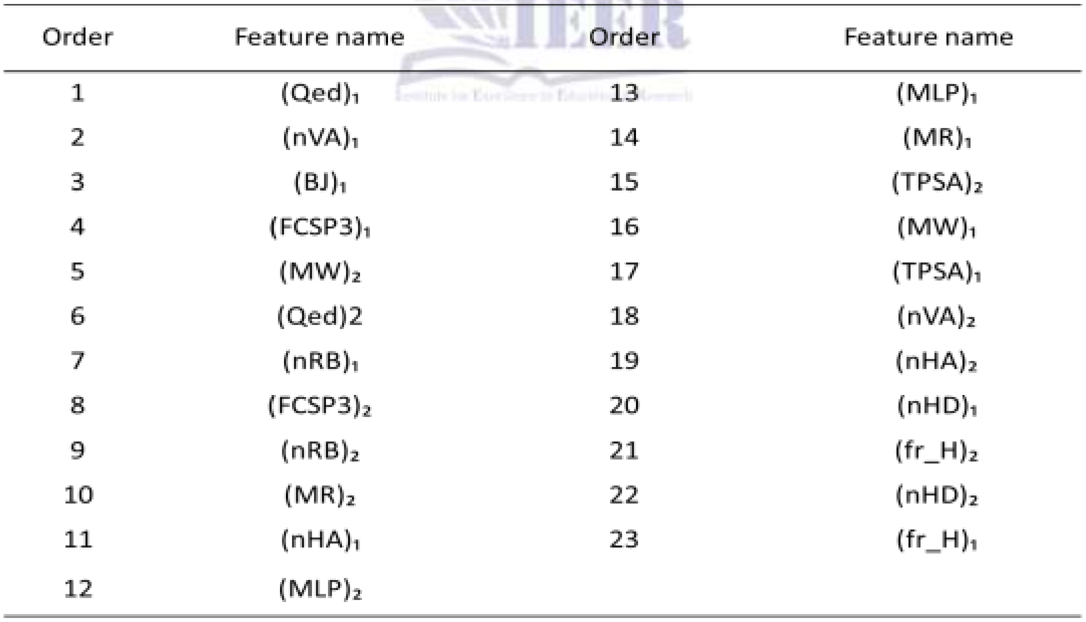
Subsequently to set the performance, the feature selection was performed using Random Forest and Extreme Gradient Boosting (XGBoost) as baseline models trained on complete feature set. Selecting the best baseline model by evaluating the R^2, MAE, and RMSE of both models. Followed by three complementary feature selection strategies were systematically implemented and compared: Recursive Feature Elimination (RFE), univariate feature selection, and LASSO regression [22,24]. These methods implemented using best baseline model. RFE with cross validation removed the least important feature based on the model-derived importance scores. Similarly univariate feature selection evaluate each feature according to its correlation with target variable, while LASSO feature selection with L1 regularization reduce the less important coefficient to zero, obtaining optimize subset of features on error minimization. The optimal number of subset determined by these feature selection method using cross-validated MAE and cross-validated MSE evaluation.

Ten ML models linear, tree-based, and ensemble techniques were trained for each optimal feature subset derived from three selection method. Scikit-learn library is used in python for the implementation of all these ML models and compared using R^2, MAE, and RMSE on both training and test sets for all models trained on different subset derived from three selection methods to ensure robust generalization and avoid overfitting [25]. The top model was selected based on superior testing performance across all feature selection strategies was further optimized to ensure accurate, reliable, interpretable, generalizable and reproducible prediction [26]. The optimized model was subsequently applied to unseen data for prediction seebeck coefficient. The seebeck coefficient predicted was then back-transformed to obtain the final Seebeck coefficient, permitting the identification of promising i-TE materials for TEC applications. Beyond predictive performance, model interpretability such as SHAP analysis and correlation analysis were performed on beast performing models which quantify the contribution of each univariate selected features with each other and with the predicted seebeck coefficient [27],[28]. These analysis elucidate to capture physical meaningful and complementary structural information regarding parameters of TECs behavior. Which shift the model from a prediction engine into a scientific discovery tool.

Result and discussion:

The developed ML framework facilitated a rigorous and data-driven assessment of the performance of i-TE materials via systematic benchmarking, feature identification, and predictive modeling. A base line models random forest and Xgboost were trained on all features. Selecting the Random forest as best baseline model achieving high R^2 value while maintaining the lowest MAE and RMSE values. Initial analysis exposed that not all descriptors contributed equally to predictive accuracy, stress the necessity of dimensionality reduction to remove redundant or weakly informative variables and to improve model generalization. Following the RFE with cross validation was first applied to rank all features to their predictive importance, listed in Table 2.

Table 2: list the RFE selected features rank for matrix and ion donor. Where subscript '1' denotes matrix-derived features and '2' denotes ion donor-derived features.



Order	Feature name	Order	Feature name
1	(Qed)$_1$	13	(MLP)$_1$
2	(nVA)$_1$	14	(MR)$_1$
3	(BJ)$_1$	15	(TPSA)$_2$
4	(FCSP3)$_1$	16	(MW)$_1$
5	(MW)$_2$	17	(TPSA)$_1$
6	(Qed)2	18	(nVA)$_2$
7	(nRB)$_1$	19	(nHA)$_2$
8	(FCSP3)$_2$	20	(nHD)$_1$
9	(nRB)$_2$	21	(fr_H)$_2$
10	(MR)$_2$	22	(nHD)$_2$
11	(nHA)$_1$	23	(fr_H)$_1$
12	(MLP)$_2$		

RFE isolated that 22 feature subset provide the best compromise between model accuracy and efficiency shown in figure 1. This finding was supported by lower cross validation error, adding more features would not improve the model's performance but could cause overfitting.

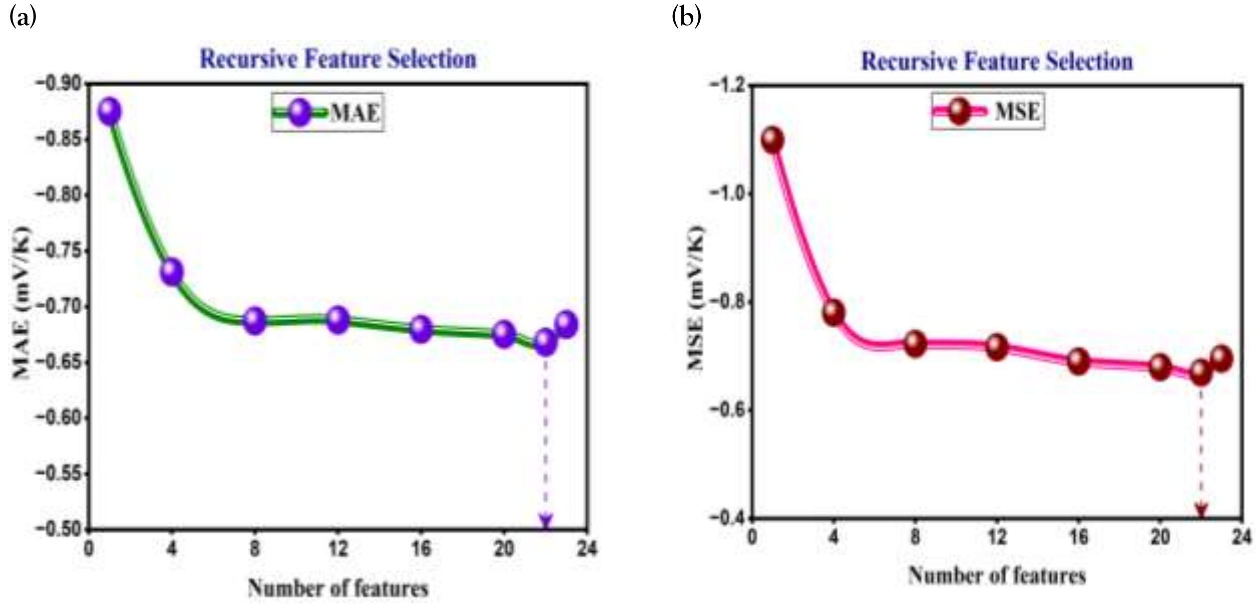
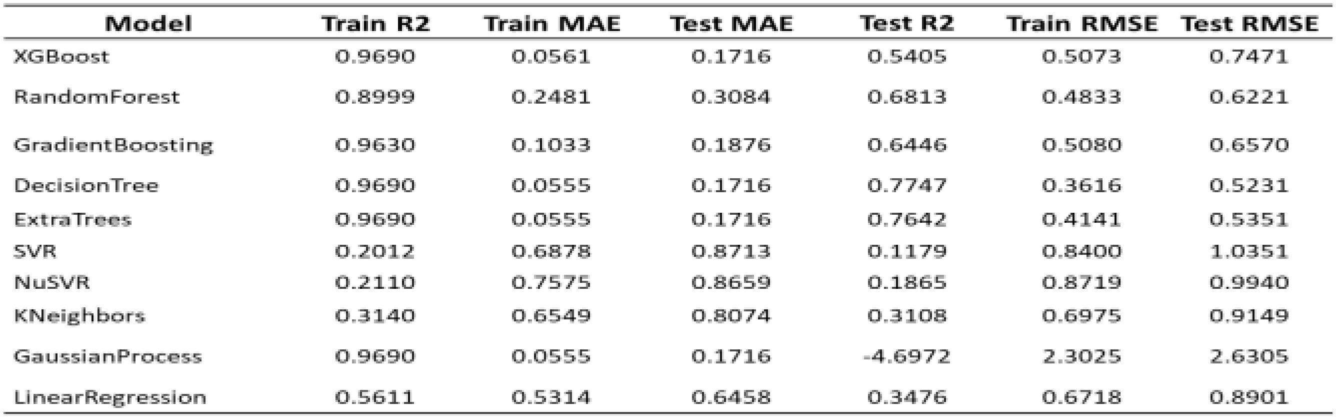


Figure 1: Illustrate the RFE with Cross-Validation identified an optimal number for both MAE and MSE metrics. **(a)** Shows the highest value of MAE for 22 RFE selected features compare to other sets. **(b)** Shows the highest value of MSE for 22 RFE selected features compared to other sets of features. Ensure that 22 RFE selected feature set as the optimal descriptor set for the predictive framework.

Ten ML models were trained using RFE selected 22-feature subset and evaluated their performance. Their statistical performance metrics, including R², MAE, and RMSE for both training and test sets, are reported in Table 3. The comparison highlights the difference in prediction accuracy and generalization across the models.

Table 3: Highlight the performance of multiple ML models train on 22 RFE selected feature subset using R², MAE, and RMSE values for both training and testing phases. The comparison summarize the difference in prediction accuracy and generalization across the models. The top-performing model Decision Tree — exhibiting the highest R² and lowest MAE and RMSE—is shown.



Model	Train R2	Train MAE	Test MAE	Test R2	Train RMSE	Test RMSE
XGBoost	0.9690	0.0561	0.1716	0.5405	0.5073	0.7471
RandomForest	0.8999	0.2481	0.3084	0.6813	0.4833	0.6221
GradientBoosting	0.9630	0.1033	0.1876	0.6446	0.5080	0.6570
DecisionTree	0.9690	0.0555	0.1716	0.7747	0.3616	0.5231
ExtraTrees	0.9690	0.0555	0.1716	0.7642	0.4141	0.5351
SVR	0.2012	0.6878	0.8713	0.1179	0.8400	1.0351
NuSVR	0.2110	0.7575	0.8659	0.1865	0.8719	0.9940
KNeighbors	0.3140	0.6549	0.8074	0.3108	0.6975	0.9149
GaussianProcess	0.9690	0.0555	0.1716	-4.6972	2.3025	2.6305
LinearRegression	0.5611	0.5314	0.6458	0.3476	0.6718	0.8901

Subsequently, univariate feature selection with cross-validation was than employed to independently evaluate the statistical relationship between each feature and the Seebeck coefficient. Feature rankings to their predictive importance are presented in Table 4.

Table 4: list the univariate selected features rank for matrix and ion donor. Where subscript '1' denotes matrix-derived features and '2' denotes ion donor-derived features.

Order	Feature Name	Order	Feature Name
1	$(FCSP3)_1$	13	(MW)2
2	(qed)2	14	(MW)1
3	$(nRB)_2$	15	(nVE)1
4	(BJ)1	16	(fr_H)2
5	(FCSP3)2	17	(nHA)2
6	(MR)2	18	(NHD)2
7	(nVE)2	19	(TPSA)1
8	(fr_H)1	20	(nHA)1
9	(MLP)2	21	(MLP)1
10	(nRB)1	22	(MR)1
11	(TPSA)2	23	(NHD)1
12	(qed)1		

Univariate selection reveal that 5 feature subset provide the best compromise between model accuracy and efficiency. This analysis was supported by lower cross validation error shown in figure 2, representing that a highly compact feature space was adequate to capture the dominant structure–property relationships.

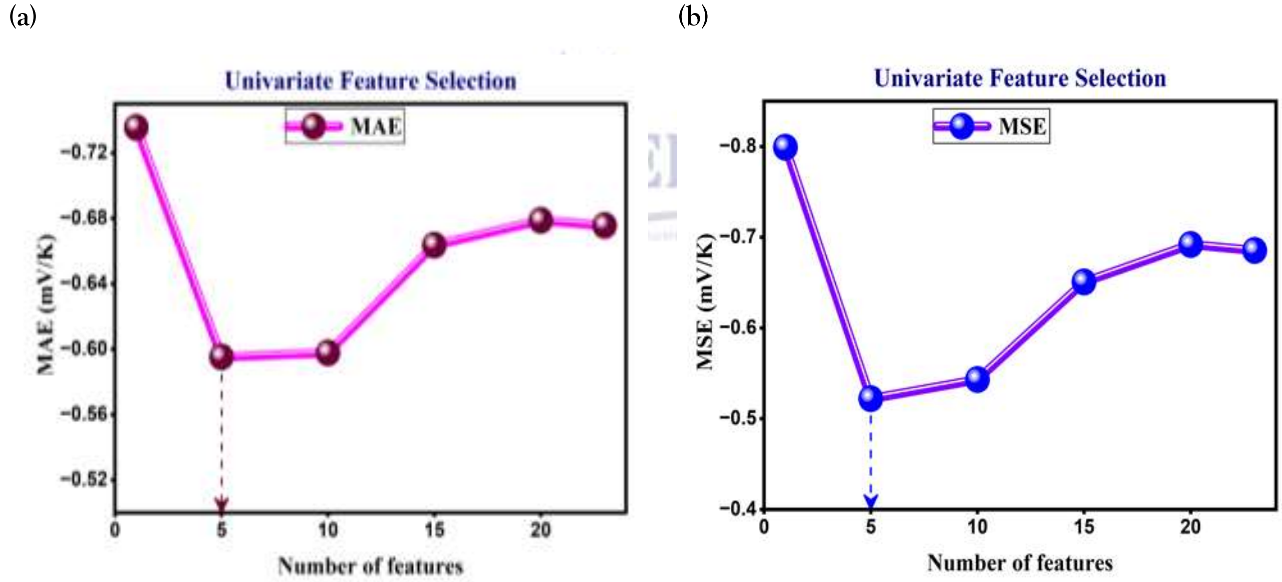
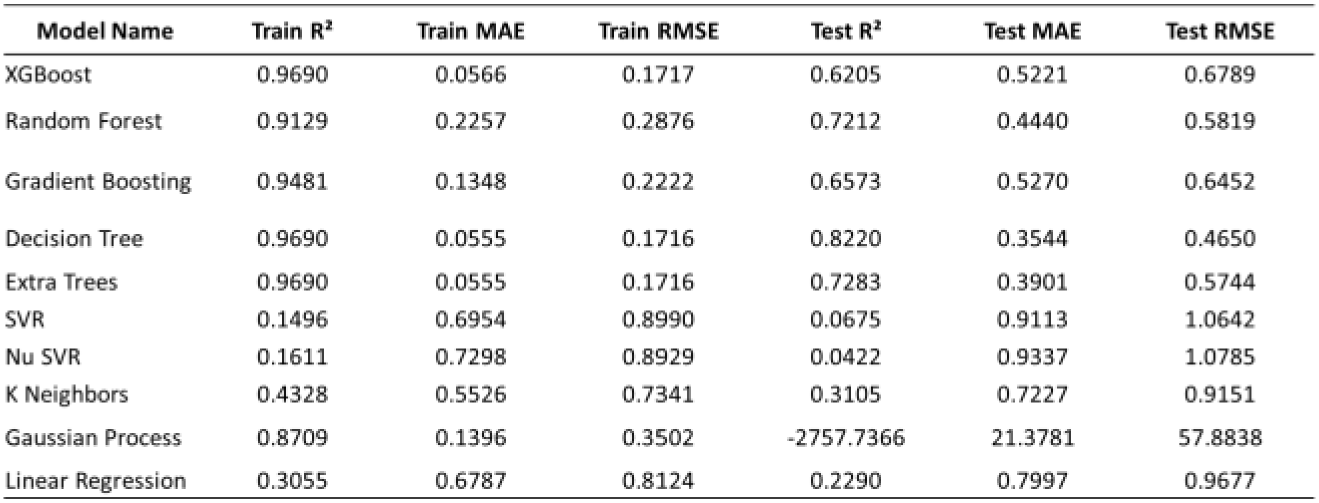


Figure 2: Illustrate the univariate selection method with Cross-Validation identified an optimal number for both MAE and MSE metrics. **(a)** Shows the highest value of MAE for 5 RFE selected features compare to other sets. **(b)** Shows the highest value of MSE for 5 univariate selected features compared to other sets of features. Ensure that 5 univariate selected feature set as the optimal descriptor set for the predictive framework.

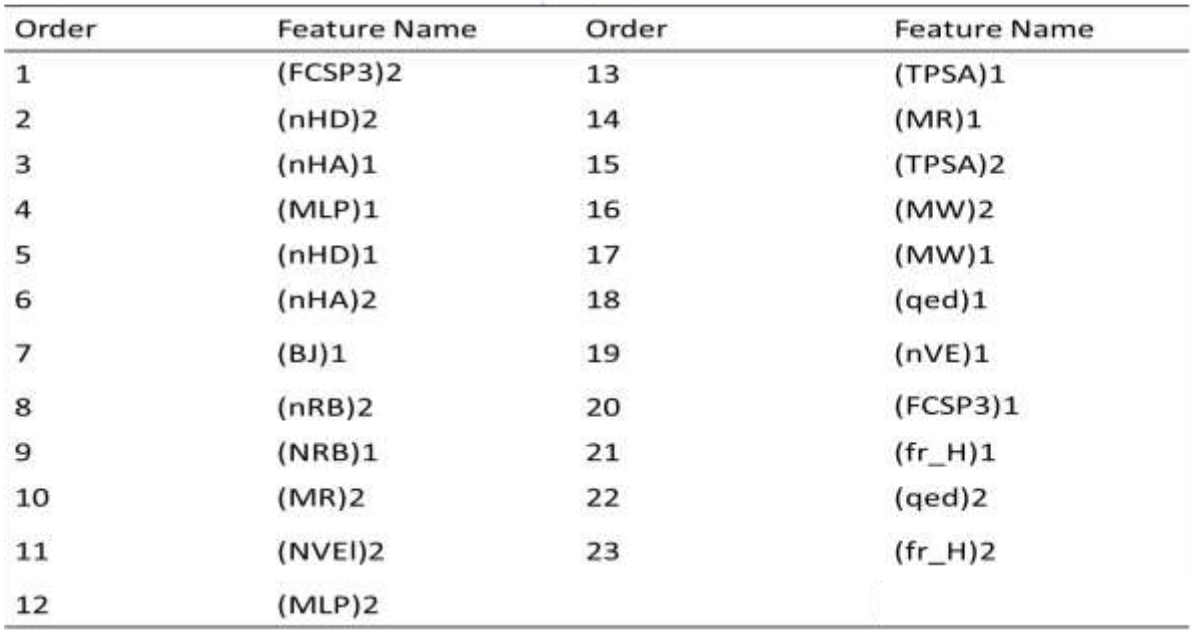
Based on univariate feature selection optimize subset, ten ML models were trained and evaluated, with their corresponding R², MAE, and RMSE values summarized in Table 5.

Table 5: Highlight the performance of multiple ML models train on 5 univariate selected feature subset using R², MAE, and RMSE values for both training and testing phases. The top-performing model Decision Tree exhibiting the highest R² and lowest MAE and RMSE—is shown.



Model Name	Train R²	Train MAE	Train RMSE	Test R²	Test MAE	Test RMSE
XGBoost	0.9690	0.0566	0.1717	0.6205	0.5221	0.6789
Random Forest	0.9129	0.2257	0.2876	0.7212	0.4440	0.5819
Gradient Boosting	0.9481	0.1348	0.2222	0.6573	0.5270	0.6452
Decision Tree	0.9690	0.0555	0.1716	0.8220	0.3544	0.4650
Extra Trees	0.9690	0.0555	0.1716	0.7283	0.3901	0.5744
SVR	0.1496	0.6954	0.8990	0.0675	0.9113	1.0642
Nu SVR	0.1611	0.7298	0.8929	0.0422	0.9337	1.0785
K Neighbors	0.4328	0.5526	0.7341	0.3105	0.7227	0.9151
Gaussian Process	0.8709	0.1396	0.3502	-2757.7366	21.3781	57.8838
Linear Regression	0.3055	0.6787	0.8124	0.2290	0.7997	0.9677

Table 6: list the LASSO selected features rank for matrix and ion donor. Where subscript '1' denotes matrix-derived features and '2' denotes ion donor-derived features.



Order	Feature Name	Order	Feature Name
1	(FCSP3)2	13	(TPSA)1
2	(nHD)2	14	(MR)1
3	(nHA)1	15	(TPSA)2
4	(MLP)1	16	(MW)2
5	(nHD)1	17	(MW)1
6	(nHA)2	18	(qed)1
7	(BJ)1	19	(nVE)1
8	(nRB)2	20	(FCSP3)1
9	(NRB)1	21	(fr_H)1
10	(MR)2	22	(qed)2
11	(NVEI)2	23	(fr_H)2
12	(MLP)2		

LASSO selection result reveal that 9 and 13 feature subset provide the best compromise between model accuracy and efficiency. Nine features minimized MAE, whereas thirteen features minimized MSE enables simultaneous feature selection and regularization, yielding sparse, interpretable models by lessening irrelevant coefficients to zero shown in figure 3.

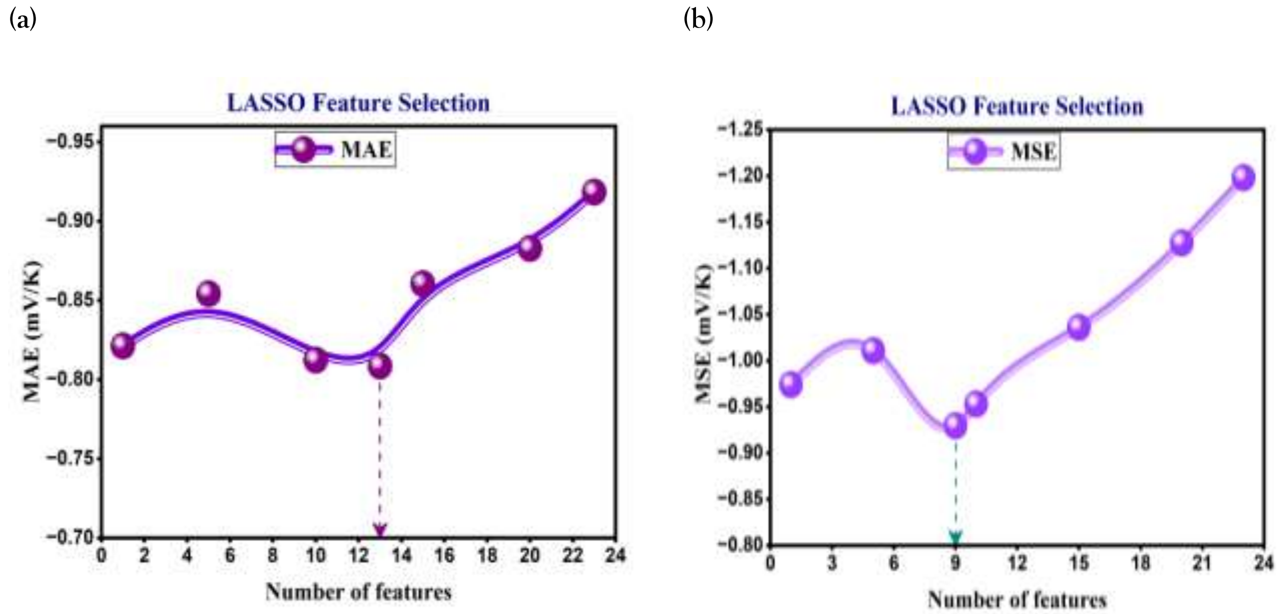


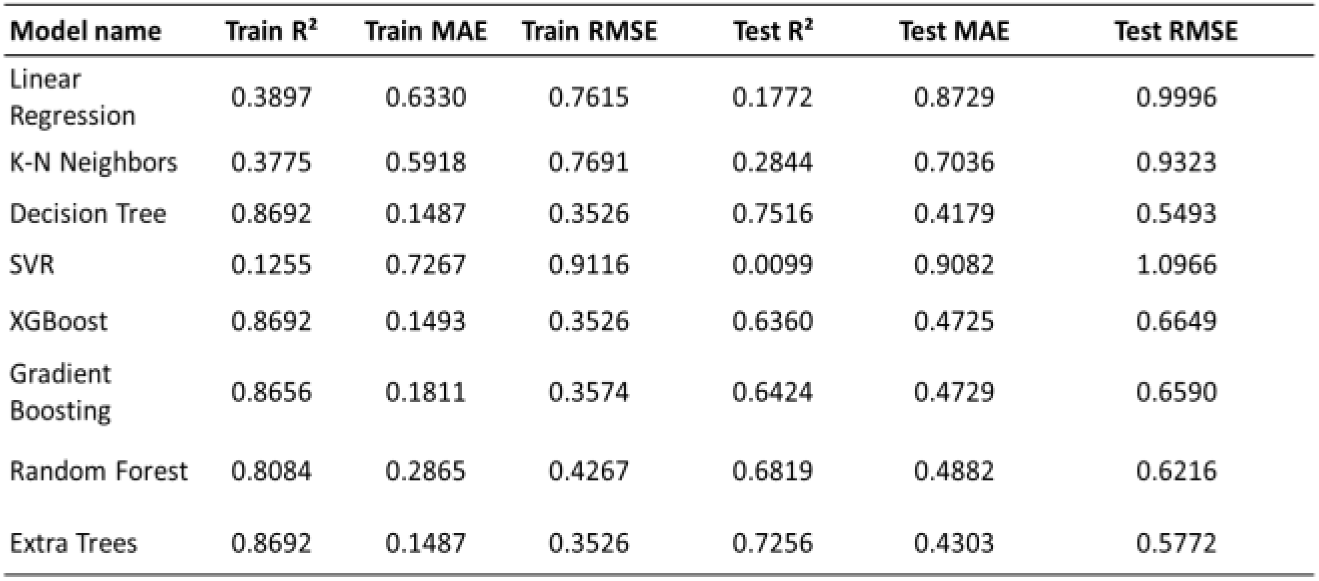
Figure 3: Illustrate the LASSO selection method with Cross-Validation identified an optimal number for both MAE and MSE metrics.**(a)** Shows the highest value of MAE for 13 LASSO selected features compare to other sets. **(b)** Shows the highest value of MSE for 9 LASSO selected features compared to other sets of features.

Subsequently both LASSO selected subsets were used to train and benchmark ten ML models, their statistical performance on training and test datasets is reported in Table 7 and Table 8.

Table 7: Summarize the performance of multiple ML models train on 9 LASSO selected feature subset using R², MAE, and RMSE values for both training and testing phases. The top-performing model Decision Tree exhibiting the highest R² and lowest MAE and RMSE—is shown.

Model Name	Train R²	Train MAE	Train RMSE	Test R²	Test MAE	Test RMSE
Linear Regression	0.2145	0.7024	0.8640	0.0686	0.9057	1.0636
K-Nearest Neighbors	0.4576	0.5635	0.7179	0.2507	0.7092	0.9540
Decision Tree	0.8692	0.1491	0.3526	0.7523	0.4133	0.5485
SVR	0.2707	0.6074	0.8325	0.1035	0.8448	1.0434
XGBoost	0.8692	0.1499	0.3526	0.6665	0.4808	0.6364
Gradient Boosting	0.8649	0.1812	0.3583	0.6641	0.4860	0.6387
Random Forest	0.8145	0.2763	0.4199	0.6618	0.5107	0.6409
Extra Trees	0.8692	0.1491	0.3526	0.7019	0.4554	0.6017
Gaussian Regressor	0.7947	0.2912	0.4416	0.6422	0.5034	0.6592
Linear SVR	0.2640	0.7124	0.8363	0.2848	0.8171	0.9320

Table 8: Summarize the performance of multiple ML models train on 13 LASSO selected feature subset using R², MAE, and RMSE values for both training and testing phases.. The top-performing model Decision Tree—exhibiting the highest R² and lowest MAE and RMSE—is shown.



Model name	Train R²	Train MAE	Train RMSE	Test R²	Test MAE	Test RMSE
Linear Regression	0.3897	0.6330	0.7615	0.1772	0.8729	0.9996
K-N Neighbors	0.3775	0.5918	0.7691	0.2844	0.7036	0.9323
Decision Tree	0.8692	0.1487	0.3526	0.7516	0.4179	0.5493
SVR	0.1255	0.7267	0.9116	0.0099	0.9082	1.0966
XGBoost	0.8692	0.1493	0.3526	0.6360	0.4725	0.6649
Gradient Boosting	0.8656	0.1811	0.3574	0.6424	0.4729	0.6590
Random Forest	0.8084	0.2865	0.4267	0.6819	0.4882	0.6216
Extra Trees	0.8692	0.1487	0.3526	0.7256	0.4303	0.5772

Comparing MAE and MSE across different selection method on optimum set reveals that univariate selected 5 feature subset achieved the lowest errors indicating superior predictive accuracy shown in figure 4.

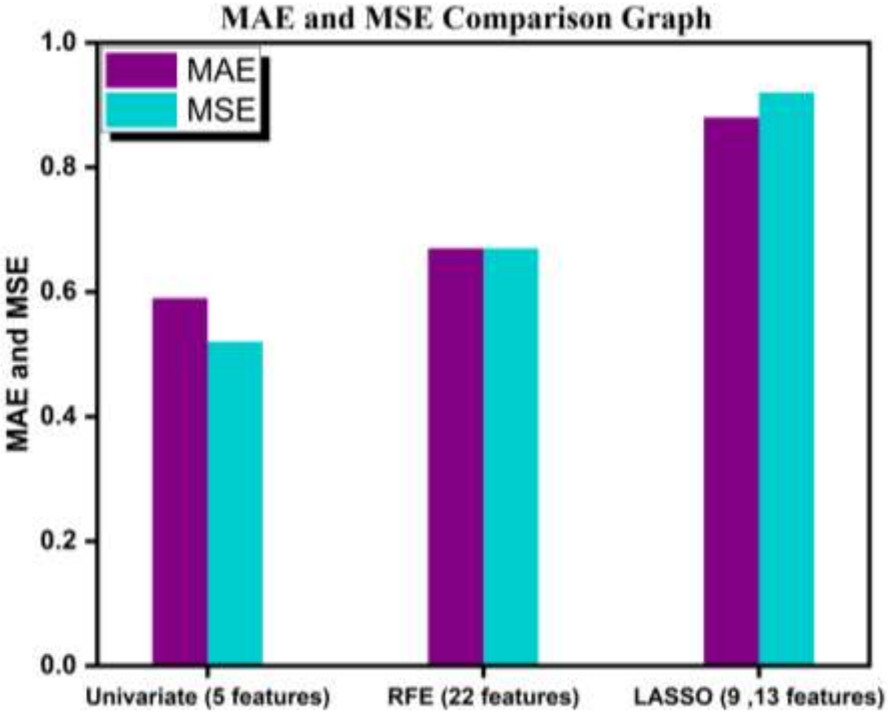


Figure 4: Compares the MAE and MSE values for optimum features subset obtained through RFE, univariate and LASSO selection methods. It is revealed that univariate selected feature subset show highest performance exhibiting lower MAE and MSE.

The overall model ranking was established by integrating the comparison of models developed using the three feature selection strategies. The comparison clearly show that Decision Tree model trained on optimum feature subset selected by three feature selection strategies was the top performer compare to other models exhibiting the highest R^2, accompanied by low MAE and RMSE value on both training and testing dataset.

Among the evaluated models, Decision Tree model train on 5-univariate selected feature subset demonstrated superior performance compared to the same model trained on RFE and LASSO selected optimum feature subset shown in figure 5. This ML model enabling clear visualization of the ranked decision rules through which molecular descriptors influence ionic Seebeck coefficients.

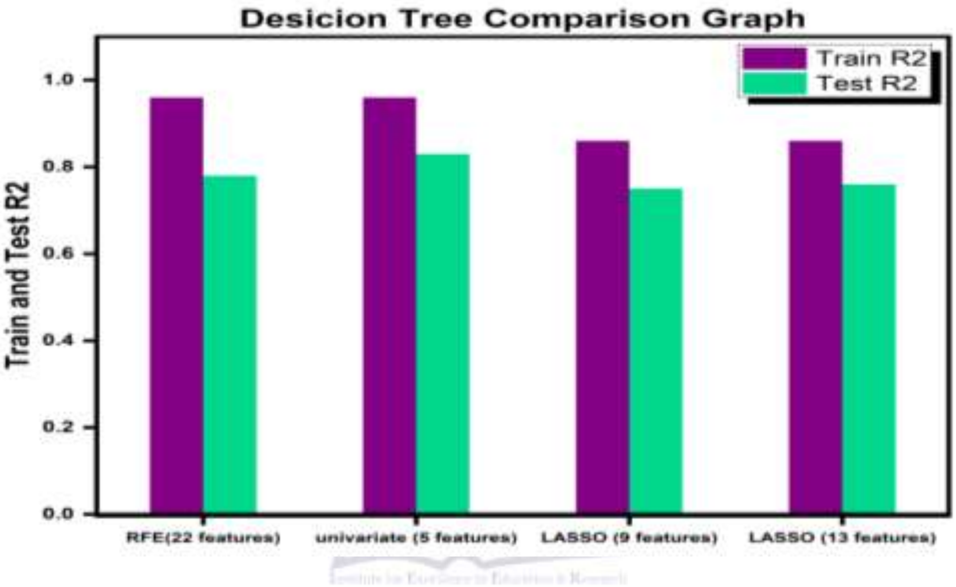


Figure 5: Compares the training and testing R^2 scores for model Decision Tree trained on different features subset obtained through RFE, univariate and LASSO selection methods. It is revealed that Decision Tree train on univariate selected feature subset show highest performance exhibiting high R^2 score.

To further minimize overfitting and improve the performance accuracy of top selected model was selecting for additional subsequent optimization for refinement. Scatter plots of the optimized decision Tree show a close-fitting clustering by comparing training and testing set values around the ideal regression line shown in figure 6. Based on the results, the optimized Decision Tree model was selected as the final predictive model for high-performance electrolyte screening.

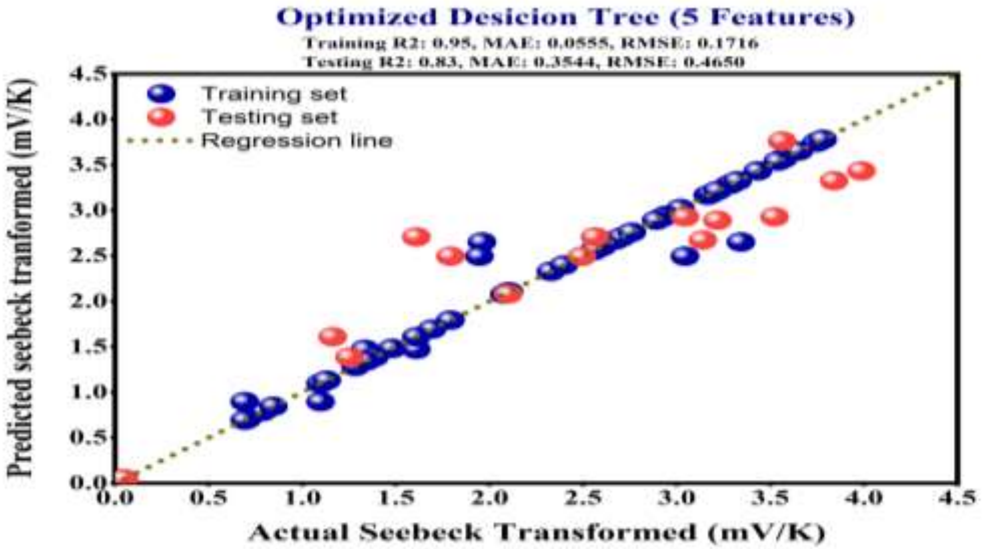


Figure 6: Present scatter plots for top model after optimization, respectively show the comparison between predicted and actual seebeck transformed (mV/K). Each point represents an individual sample. Both training and testing datasets are distinguished by different colors the dashes straight line is ideal regression line. Blue color spheres represent the training set and red color spheres represent testing set.

Further decision was dedicated to explain the model interpretability which equally critical for finding the meaningful insight that can guide the process of thermos diffusion in TEC systems. ML model were used to analyze the relationship between 5 univariate selected features and the target value seebeck coefficient. In count to accomplish accurate prediction and identification of the most influential molecular descriptors, interpretability of the model also play a vital role in elucidating the key physicochemical factors that govern ionic thermos diffusion, thermoelectric reaction in electrolyte systems and entropy transport by charge carrier under temperature gradient.

Using SHAP analysis to determine the relationship between molecular descriptors by calculating each descriptor contribution as shown in figure 7. This analysis explain direction (positive / negative effect) and molecular descriptors behavior across combinations. The analysis revealed that, all features had a constant SHAP value of -0.8 with a high feature value of 0.9, which means that high values of these descriptors are linked to a low value of the predicted target variable. (FCSP3)1, (nRB)2 show high impact on model, (qed)2 show mixed contribution while (BJ)1, and (FCSP3)2 overall effect are smaller than the top features but still measurable influence.

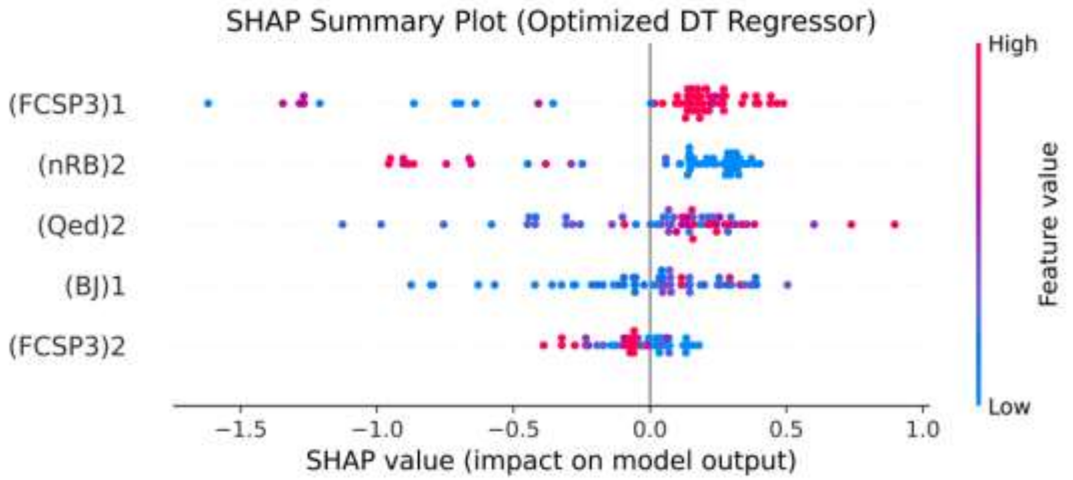


Figure 7: Demonstrate the SHAP summary plot for 5 univariate- selected features, showing the impact of feature on the final prediction. Where subscript '1' and '2' denotes matrix-derived and ion donor-derived features. From top to bottom, the features are listed in descending order of importance. The horizontal axis represents SHAP values, where right and left shifted points indicate positive and negative contribution to predicted seebeck coefficient, and the color gradient show the magnitude of feature value. In plot it is shown that (FCSP3)1 and (nRB)2 both show positive contribution to the seebeck coefficient.

Similarly mean absolute SHAP bar chart, explain the overall features importance clearly. It is a global interpretability analysis that measure the average contribution of each feature to the prediction as shown in figure 8. The analysis revealed that (FCSP3)1 had the highest mean absolute SHAP value, making it the most important feature, followed by (nRB)2 and (qed)2 while (BJ)1 and (FCSP3)2 had relatively lower values, respectively, yet significantly contributed. These results confirm the importance of molecular saturation, rotatable bonds, and drug-likeness in making predictions, and also check the significance of the chosen features by comparing them with the results of the previous univariate analysis.

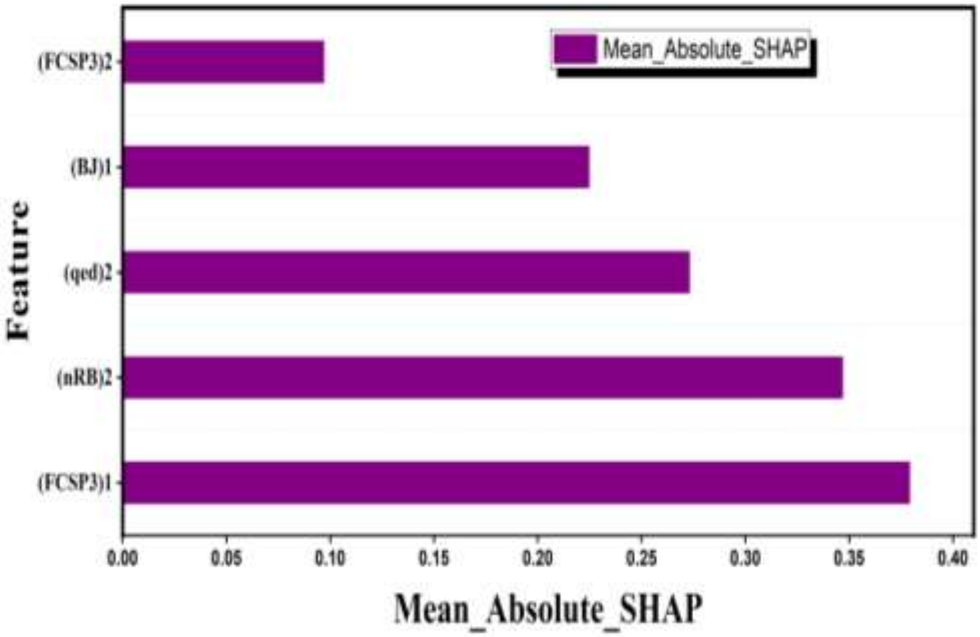


Figure 8: Demonstrates individual contribution of 5 univariate selected feature on target value. Where subscript ‘1’ denotes matrix-derived features and ‘2’ denotes ion donor-derived features. Higher the mean SHAP value, the more important the feature is. It is revealed that (FCSP3)1, (nRB) 2 is the most important feature.

Subsequently a correlation matrix was constructed for 5 univariate selected features as shown in figure 9, The analysis revealed generally low to moderate inter-feature correlations, with values ranging from −0.38 to 0.60. It is revealed that (FCSP3)1, (BJ)1 and (qed)2 show positive correlation with the seebeck coefficient while (nRB)2 and (FCSP3)2 are negative correlated with seebck coefficient. These results show that the selected molecular descriptors capture complementary structural and physicochemical information with marginal redundancy, supporting their combined use in multivariate modeling without significant risk of multicollinearity.

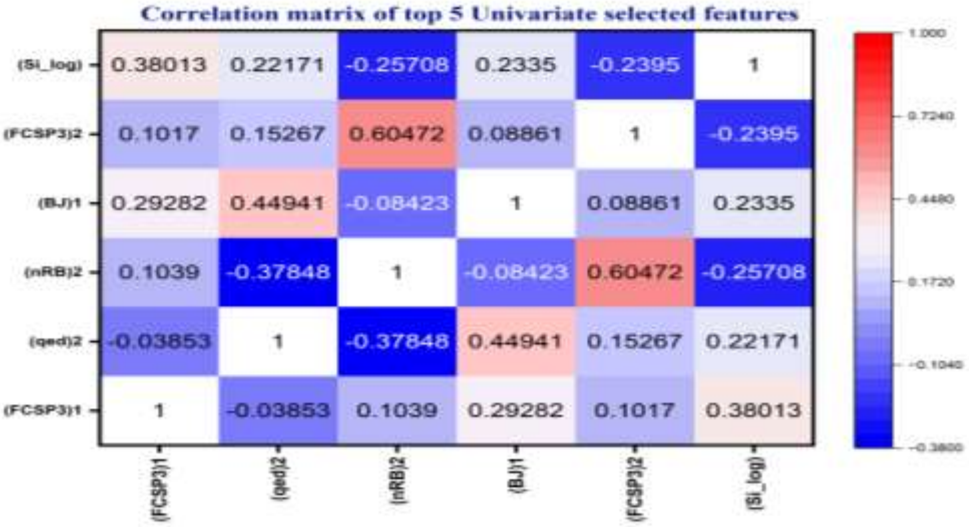


Figure 9: Demonstrate the correlation between 5 univariate selected features with each other and also with seebeck coefficient. The color gradient shows the strength and direction of the correlations, revealing both inter-feature dependencies and structural property relationships relevant to TECs. The values ranging from -0.38 to 1 indicating the strength and direction of correlation. It is found that (FCSP3)1, (BJ) 1 and (qed) 2 show the positive correlation, while (FCSP3)2 and, (nRB) 2 exhibit negative correlations with seebeck coefficient.

Finally the unseen dataset is subjected to optimized decision Tree model for prediction the seebeck coefficient. The model predicted several high Seebeck coefficients, particularly for systems containing polyurethane-based, cellulose-based, and PVA-based combinations. These materials revealed structural flexibility and favorable ion transport pathways, which are beneficial for TECs performance. The top five high Seebeck coefficients are presented in Table 9**.**

Table 9: Summarize the top matrix + ion donor combinations exhibit high Seebeck coefficient due to synergy between matrix-guided ion transport and temperature-sensitive ion donors.

Rank	Matrix	Ion Donor	Seebeck (mV/K)
1	Oxidized cellulose membrane	KNO_3	42.8
2	PVA	Quaternary ammonium salts	41.39
3	Gelatin ionogel	KSCN	41.39
4	PAM hydrogel	CsI	27.0
5	1-Dodecanol (neat)	Quaternary ammonium salts	25.6

Conclusion:

The present work explores a Machine Learning based approach to predict the seebeck coefficient in TECs. This study reveals that by using curated dataset and determine optimum feature subset using different features selection techniques such as recursive feature elimination method, univariate selection method and LASSO selection method improve the accuracy and generalization of final prediction. Similarly Training multiple models on optimum feature subset selected by different selection method improve predictive performance by allowing comparison and selection of best- performing algorithm .And also by combining models to achieve more robustness, enhance generalization and reliability by reducing bias, diminishing overfitting and ensuring models performance on unseen data. Moreover, the model interpretability analysis such as SHAP and correlation analysis showed the importance of each selected feature to the model's prediction, which is critical in physical insights. It reveal the importance of molecular saturation, rotatable bonds, and drug-likeness in making predictions, and also check the significance of the chosen features by comparing them with the results of the previous univariate analysis.

In order to achieve additional superior robustness and generalization the top selected model Decision Tree was further subjected to an optimization process, which allows for reliable extrapolation to unexplored regions of chemical space. This model is than applied to unseen data to predict the seeebck coefficient of i-TE materials used in TECs. Which predict various matrix-ion donor pairs with high seebeck coefficients, especially for the most systems of cellulose-base, and PVA-base along with gelatin ionogel, PAM hydrogel systems. These predictions again reveal the effectiveness of the machine learning framework as a screening and design tool. Finally, this research effort makes a substantial impact to the design and optimization of efficient TECs for future energy-harvesting technologies.

References

[1] International Energy Agency, "World energy outlook 2022," International Energy Agency, Paris, France, 2022. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2022

[2] H. Jouhara, N. Khordehgah, S. Almahmoud, B. Delpech, A. Chauhan, and S. A. Tassou, "Waste heat recovery technologies and applications," *Thermal Science and Engineering Progress*, vol. 6, pp. 268–289, Jun. 2018, doi: 10.1016/j.tsep.2018.04.017.

[3] M. T. Alam, T. I. Anowar, Md. Ashiquzzaman, F. Dhali, and S. Niloy, "Electricity generation from exhaust steam of a steam turbine generator," *Journal of Electrical and Power System Engineering*, vol. 10, no. 1, pp. 33–43, Jan.–Apr. 2024.

[4] K.-T. Lee, D.-S. Lee, W.-H. Chen, D. Luo, Y.-K. Park, and A. Bandala, "An overview of commercialization and marketization of thermoelectric generators for low-temperature waste heat recovery," *Renewable and Sustainable Energy Reviews*, vol. 186, p. 113874, Oct. 20, 2023.

[5] W. Li and S. Wang, "Principles of low-grade heat harvesting," in *Low-Grade Thermal Energy Harvesting: Advances in Materials, Devices, and Emerging Applications*, Woodhead Publishing Series in Electronic and Optical Materials, 2022, pp. 1–10.

[6] D. Zhao, A. Würger and X. Crispin, "Ionic thermoelectric materials and devices," J. Energy Chem, vol. 61, p. 88–103, 2021.

[7] C.-G. Han, X. Qian and Q. e. a. Li, "Giant thermopower of ionic gelatin near room temperature," Science, vol. 368, no. 6495, p. 1091–1098, 2020.

[8] C. Dagdeviren, Z. Li, Z. L. Wang, "Energy harvesting from the animal/human body for self-powered electronics," Annu. Rev. Biomed. Eng, vol. 19, p. 85–108, 2017.

[9] Zhou et al., "Novel porous thermosensitive gel electrolytes for wearable thermo-electrochemical cells," Cell Reports Physical Science, vol. 449, no. 137775, 2022.

[10] T. Wang, C. Zhang, H. Snoussi, and G. Zhang, "Machine learning approaches for thermoelectric materials research," *Advanced Functional Materials*, vol. 30, no. 12, p. 1906041, 2020, doi: 10.1002/adfm.201906041.

[11] H. Wang et al., "Data-driven screening of ionic thermoelectric materials," Energy Environ. Sci, vol. 15, p. 2923–2934, 2022.

[12] B. T. Huang et al, "Thermoelectricity and thermodiffusion in charged colloids," Journal of Chemical Physics (J. Chem. Phys.), vol. 143, no. 5, p. 054902, 2015.

[13] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, "Machine learning for molecular and materials science," Nature, vol. 559, p. 547–555, 2018.

[14] S. Wu et al., "Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm," npj Comput. Mater, vol. 5, p. 66, 2019.

[15] U. S. Vaitesswar et al, "Machine learning based feature engineering for thermoelectric materials by design," Digit. Discov, vol. 3, p. 210–220, 2024.

[16] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell, vol. 2, no. 1, pp. 56-57, 2020.

[17] Zhou et al., "Novel porous thermosensitive gel electrolytes for wearable thermo-electrochemical cells," Cell Reports Physical Science, vol. 449, no. 137775, 2022.

[18] R. Wu, B. Liu et al., “High-Thermopower Thermogalvanic Ionic Hydrogel for Efficient Low-Grade Heat Energy Harvesting in Electronic Devices,” ACS, vol. 7, no. 19, 2025.

[19] H. Cheng, X. He, Z. Fan, and J. Ouyang., "Flexible Quasi-Solid State Ionogels with Remarkable Seebeck Coefficient and High Thermoelectric Properties," Adv. Energy Mater, vol. 9, no. 32, p. 1901085, 2019.

[20] Wang S, Li Y, Yu M et al, "High-performance cryo-temperature ionic thermoelectric liquid cell developed through a eutectic solvent strategy," Nature Communication, vol. 15, p. 1172, 2024.

[21] D. Zhang, Y. Zhou, et al, "Highly antifreezing thermogalvanic hydrogels for," Nano Lett, vol. 23, p. 11272–11279, 2023. [19] Qian X, Ma Z, Huang Q et al, "Thermodynamics of ionic thermoelectrics for low-heat harvesting," ACS Energy Lett, vol. 9, p. 679–706, 2024.

[22] Q. Chen, Z. Meng, X. Liu, Q. Jin, and R. Su, “Decision variants for the automatic determination of optimal feature subset in RF-RFE,” Genes, vol. 9, no. 6, p. 301, Jun. 2018, doi: 10.3390/genes9060301.

[23] T. Emura, S. Matsui, and H.-Y. Chen, “compound.Cox: Univariate feature selection and compound covariate for predicting survival,” Computer Methods and Programs in Biomedicine, vol. 168, pp. 21–37, Jan. 2019, doi: 10.1016/j.cmpb.2018.10.020.

[24] R. Muthukrishnan and R. Rohini, “LASSO: A feature selection technique in predictive modeling for machine learning,” in *Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 2016, pp. 18–20.

[25] Pedregosa F, Varoquaux G, Gramfort A et al, "Scikit-learn: machine learning in Python," J. Mach. Learn. Res., vol. 12, p. 2825–2830, 2011.

[26] Bergstra J, Komer B et al, "Hyperopt: a Python library for model selection and hyperparameter optimization," Comput. Sci. Discov, vol. 8, no. 1, p. 014008, 2015.

[27] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, “Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development,” Clinical and Translational Science, Oct. 2024, doi: 10.1111/cts.70056.

[28] N. J. Gogtay and U. M. Thatte, “Principles of correlation analysis,” Journal of the Association of Physicians of India, vol. 65, Mar. 2017.