Journal of Energy Chemistry 78 (2023) 232–239

Contents lists available at ScienceDirect

Journal of Energy Chemistry

journal homepage: www.elsevier.com/locate/jechem

http://www.journals.elsevier.com/journal-of-energy-chemistry/

An integrated machine learning model for accurate and robust prediction of superconducting critical temperature

Jingzi Zhang [a,b,1], Ke Zhang [a,b,1], Shaomeng Xu [d,e,1], Yi Li [a,b,1], Chengquan Zhong [a,b], Mengkun Zhao [a,b], Hua-Jun Qiu [a,b], Mingyang Qin [e], X.-D. Xiang [e,*], Kailong Hu [a,b,c,*], Xi Lin [a,b,c,*]

[a] School of Materials Science and Engineering, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China
[b] Blockchain Development and Research Institute, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China
[c] State Key Laboratory of Advanced Welding and Joining, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China
[d] School of Materials Science and Engineering, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China
[e] Department of Materials Science and Engineering & Department of Physics, Southern University of Science and Technology, Shenzhen 518055, Guangdong, China

ARTICLE INFO

Article history:
Received 20 September 2022
Revised 24 November 2022
Accepted 27 November 2022
Available online 5 December 2022

Keywords:
Superconductors
Integrated machine learning
Superconducting critical temperature

ABSTRACT

Discovering new superconductors via traditional trial-and-error experimental approaches is apparently a time-consuming process, and the correlations between the critical temperature (T_c) and material features are still obscure. The rise of machine learning (ML) technology provides new opportunities to speed up inefficient exploration processes, and could potentially uncover new hints on the unclear correlations. In this work, we utilize open-source materials data, ML models, and data mining methods to explore the correlation between the chemical features and T_c values of superconducting materials. To further improve the prediction accuracy, a new model is created by integrating three basic algorithms, showing an enhanced accuracy with the coefficient of determination (R^2) score of 95.9 % and root mean square error (RMSE) of 6.3 K. The average marginal contributions of material features towards T_c values are estimated to determine the importance of various features during prediction processes. The results suggest that the range thermal conductivity plays a critical role in T_c prediction among all element features. Furthermore, the integrated ML model is utilized to screen out potential twenty superconducting materials with T_c values beyond 50.0 K. This study provides insights towards T_c prediction to accelerate the exploration of potential high-T_c superconductors.

1. Introduction

Development of high-performance superconductors has attracted tremendous attentions in condensed matter physics and for the purpose of the emerging quantum computation [1–3]. The critical temperature (T_c) is an essential factor to evaluate the potential applications of the superconducting materials [4]. Generally, superconductors with high T_c values are explored and screened by enormous experiments and computational methods [5,6]. The trial-and-error experimentation for searching new superconductors usually requires ultralow temperature and extremely high pressure [7,8]. In addition, the density functional theory (DFT) based computation processes are generally time-consuming as well as costly [9,10]. Therefore, the traditional experimentation and computation limited the rapid progress of high-T_c superconductors screening and their potential commercialization. In the past decades, the data-driven scientific developments and machine learning (ML) models have enabled alternative opportunities to address the major challenges faced by new superconductors exploration. For instance, the ML models have been achieved remarkable prediction results for perovskite materials [11–13], electrochemical catalysts [14,15], thermoelectric materials [16], and polymers [17]. Moreover, the T_c values of superconductors also predicted by advanced ML models, which has obtained more fresh perspectives and accelerated the exploration of potential superconductors.

Recently, ML-assisted approaches have been widely used to efficiently predict the superconducting properties of promising superconductors [18–21]. Owolabi et al. [18] used support vector regressor (SVR) to directly investigate correlation between the lattice parameters with T_c values through computational intelligence technology. Stanev et al. [19] developed ML schemes to simulate T_c

* Corresponding authors.
E-mail addresses: xiangxd@sustech.edu.cn (X.-D. Xiang), hukailong@hit.edu.cn (K. Hu), linxi@hit.edu.cn (X. Lin).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jechem.2022.11.047
2095-4956/

values of more than 12,000 known superconductors and achieved 88% accuracy using Random Forest (RF) model. Yet, there has been rarely extensive exploration of the superconductors' database mining, such as data cleaning preprocessing, feature construction in details, and delivering the interpretable mathematical formulas with ML features. On other hand, linear models are usually not competent for complex regressions, while the tree models perform better [22]. Diverse tree models have differences in their efficiency and performance due to leaf depth and segmentation rules [23]. For instance, the RF algorithms are more time-consuming as compared to xgboost ones, while RF tends to be more stable and can avoid overfitting [24,25]. Therefore, the complexity and generalization of the algorithm need to be considered to reach a balance between the accuracy of prediction and the number of datasets. Several studies have also considered the integrated ML model as an effective method to improve the prediction accuracy. Wang et al. applied the integrated stacking approach, which used the output of multiple baseline models to enhance the performance of band gap regression [26]. Chen et al. developed the integrated ML model which contained three submodules to realize accurate prediction of the concentrations of surface particulate matter with an aerodynamic diameter < 2.5 μm [27]. The integration of multistep ML models is a promising approach to deal with the longstanding problems associated with basic ML algorithms.

In this work, an integrated ML model was designed to accurately predict T_c values. Through the data acquisition and cleaning, 13,138 data was retained from initial 33,000 pieces of selected superconductors data, which was set as the dataset for ML model training and T_c prediction. The integrated ML model was established based on the correlation between superconducting properties and T_c values to screen out the potential superconductors with high T_c values. More specifically, three basic algorithms, (i.e., gradient boosting decision tree (GBDT) [28], extra tree (ET) [29], and light gradient boosting machine (LGB) [30]), were integrated as a new ML model for the high accuracy prediction. The integrated ML model exhibited a coefficient determination (R^2) score of 95.9% for T_c prediction in comparison to other basic algorithm-assisted ML models (92.9% for GBDT, 93.0% for ET, and, 93.4% for LGB). Meanwhile, the interpretable mathematical formulas were built to guide the correlation between T_c values and important features. Twenty materials with predicted T_c values over 50.0 K were further screened out via the integrated ML model. This work provided a new insight to accelerate the exploration of potential superconductors with high T_c values.

2. Method

The suitable ML model is the key to predict T_c values of superconductors efficiently and accurately. In this work, eleven algorithms were used in ML models for T_c predictions, including the GBDT [28], ET [29], LGB [30], Linear Model (LR) [31], Lasso [32], K-nearest neighbor (KNN) [33], support vector regression (SVR) [34], decision tree (DT) [35], RF [36], eXtreme gradient boosting (XGB) [37], and multi-layer perceptron (MLP) [38]. All ML models and data preprocessing are developed by the powerful scikit-learn library [39] and the officially released Python modules [40]. The integrated ML model is applied from stacking the outputs of selected basic models to generate higher accurate predictions. The LGB, ET, and GBDT models used for the prediction of T_c values and evaluation criteria are presented as follows. The LR, Lasso, KNN, SVM, DT, RF, XGB, and MLP models are introduced in Supporting Information. The hyperparameters of all ML models have been provided in Table S1.

Normalization preprocessing was employed for all data during training and prediction processes. The total dataset was randomly shuffled and divided into the training set and test set with the ratio of 8:2. The RMSE and R^2 scores were recorded as the evaluation criteria for the T_c prediction via regression models.

$$\mathrm{RMSE} = \sqrt[2]{\frac{1}{n}\sum\nolimits_i^n (f(x_i) - y_i)^2} \tag{1}$$

$$R^2 = 1 - \frac{\sum_i^n (y_i - f(x_i))^2}{\sum_i^n (y_i - \bar{y}))^2} \tag{2}$$

where, $f(x_i)$ is the predicted value of the model; y_i is the true value; $\bar{y}$ is the mean value; RMSE can be regarded as the prediction error, and R^2 can be approximately regarded as the accuracy of regression fitting.

2.1. Light gradient boosting

LGB model is the gradient decision promotion that the samples are divided from top to bottom to establish cart tree as a weak learner, and each sample will fall on the corresponding leaf node [30]. It uses a variety of strategies such as histogram optimization, memory optimization, leaf-wise, and sequential access gradient [26]. These benefits are greater decrease in the loss function than the level-wise growth method. Therefore, LGB always provides betters prediction accuracy than other gradient boosting tree models [41].

2.2. Extra tree

ET model implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [29]. ET regression model differ from classic decision trees in the way they are built. When looking for the best split to separate the samples of a node into two groups, random splits are drawn for each of the max_features randomly selected features and the best split among those is chosen [42]. To reduce memory consumption, the complexity and size of the trees should be controlled by limiting the fully grown and unpruned trees [43].

2.3. Gradient boosting decision tree

GBDT model builds an additive model in a forward stage-wise fashion. It allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function [28]. This model is optimized by boosting tree using additive model and forward stagewise algorithm. In training, the negative gradient of the loss function is used to fit the approximate value of loss in each iteration. Therefore, error term generated in the training process is continuously reduced [44].

2.4. Integrated machine learning

The strategy of integrating several ML models is an efficient approach to improve the prediction accuracy and applicability [45]. The integrated model usually combines multiple basic ones to achieve a better generalization effect. Integrated model was stacked the outputs of multiple basic models to improve the prediction accuracy, but the prediction results of the integrated model depend on the performance of each selected basic model used for stacking [46]. And the calculation time were also affected by the number of the selected basic models [47]. Thus, the stacking integrated mode requires that the basic model itself has a high accuracy and an appropriate calculation time in this work. In the 11 basic models, the LGB, ET and GBDT models were selected in integrated ML model, because of they given the superior performance and saved the calculation time.

3. Results and discussion

The whole workflow is shown in Fig. 1. The chemical compositions and T_c values of appropriate superconductors were obtained from the SuperCon database [48], and followed by data cleaning processes [49]. Subsequently, the feature engineering was carried out to transform the elements properties to superconductivity compounds features via open-source data toolkits and packages. 115 relevant features were selected by using the recursive feature elimination (RFE) method [50]. After feature engineering, eleven basic ML models were built with the assistance of 115 features. Furthermore, these basic ML models were used to predict T_c values, and the average marginal contributions of features were estimated by the shapley additive explanations (SHAP) values [51,52]. To improve the prediction accuracy, the integrated ML model was established by integrating three basic models with high prediction performances. Finally, several materials with potential high T_c values were screened out from the MP database [53] via the integrated ML model.

3.1. Data preparation and feature engineering

33,000 pieces of superconductors data extracted from the SuperCon were set as the dataset, which contained T_c values and chemical compositions of reported materials [48]. For data cleaning, the compounds with T_c values of 0 K or unclear records were deleted. The repeated data information was also removed. The truncated averaging method was used to revise the T_c values of one compound with several records in different reports. The controversial and abnormal data points were further screened through cleaning rules of visual display and literature investigation (Table S2). After data cleaning, 13,138 superconductors were retained in the dataset. The T_c distribution of various superconductors was shown in Fig. S1. The superconductivity data used to generate the results in this work can be downloaded https://github.com/zhanzghang/Integrated-ML-model-Superconductors-data.

Furthermore, the feature engineering was performed to transform elements properties to compound features. According to the published literatures [54,55], online websites [56,57], and open-source material property sets [58,59], the characteristics of compound elements were selected by following three aspects. Firstly, the stoichiometric characteristics were expressed by one-hot coding, which depended on the proportion of the relevant elements in the compound. Secondly, the physical and chemical properties were selected based on prior knowledge and expressed. Thirdly, electronic characteristics were introduced through the electronic composition information and calculated proportion of electrons in *s*, *p*, *d*, and *f* layers of element. To construct reliable compound features, the relevant properties of various elements were gathered, which can be queried from the WebElements [56] or downloaded from the Mendeleev [57] and Magpie [59] modules. Remarkably, the above properties are specific for a single element instead of a compound. Therefore, some analytical and statistical functions were encoded to transform the element properties to compound features. Furthermore, the calculation methods were used to expand the numbers of compound features, which could efficiently avoid the overfitting [60]. Table S3 listed the calculation methods [25] for feature extraction from the chemical formula in this work. Overall, 239 initial features were generated after the first-round feature engineering.

Next, some feature selections and preprocessing are performed to avoid over-fitting during predictions. Typically, it is not recommended to use all initial features for modeling, because they often contain redundant components or highly correlated feature-pairs, which may result in negative effects, such as the non-convergence and overfitting issues [61]. The RFE method was used to remove the features that were irrelevant and undistinguishable. Based on the feature engineering, 115 features were extracted from 239 initial features and used for further superconductors data modeling.

3.2. Eleven basic models and integrated ML model

The integrated ML model training process is shown in Fig. 2. In the first layer, LGB, ET, and GBDT models were used as basic models to make test predictions on the data by 5-fold cross validation respectively. Their prediction output of each basic models in the first layer were trained as the inputs of the next layer. In the second layer, the model is usually known as meta-learner that is used to weight the prediction result of each basic model and generate the final prediction. In this work, relatively simple linear model Lasso has been selected as a meta-learner because of its stable performance, so the combined weights between different models can be automatically obtained [62]. The T_c prediction was carried out by using eleven ML models based on the test dataset. The fivefold cross-validation method was used to achieve accurate and stable

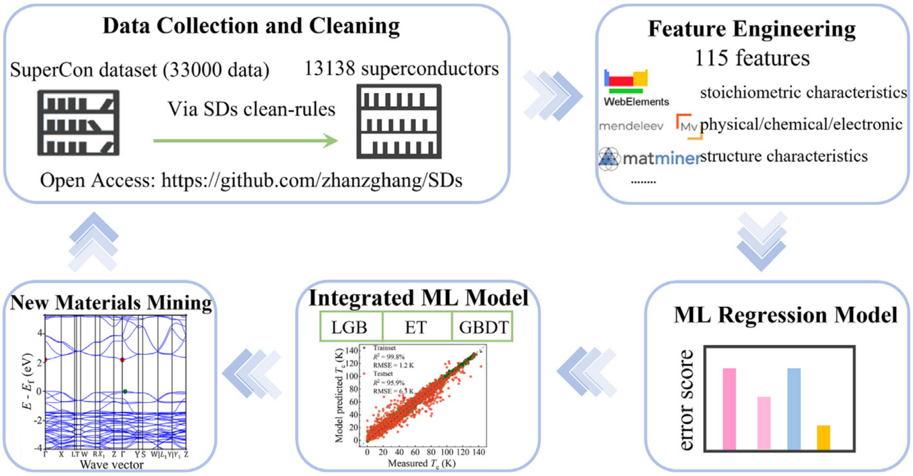

Fig. 1. The workflow of the integrated model-based ML methods for accurate T_c prediction and new superconductor materials mining.

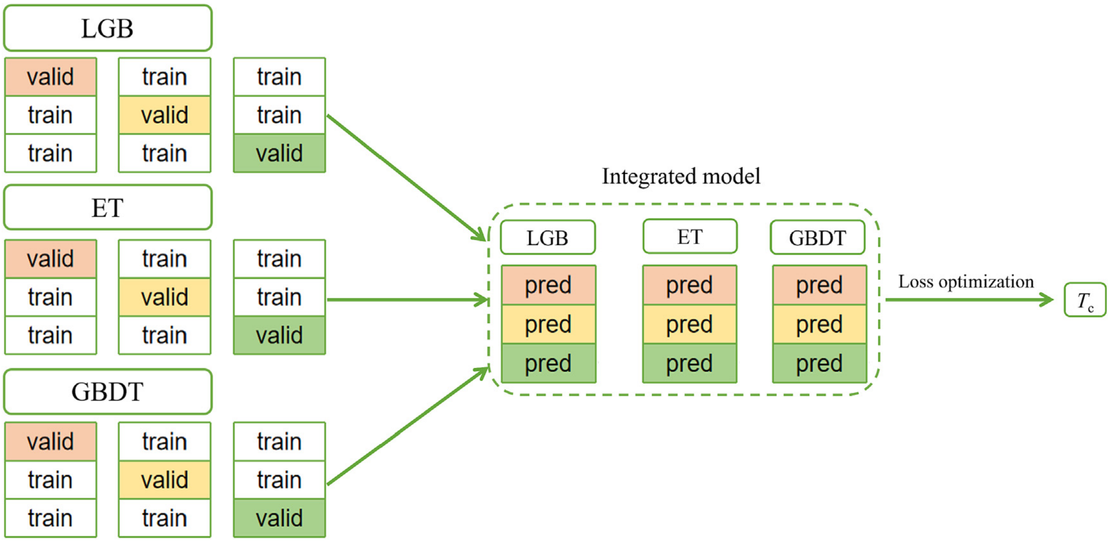

Fig. 2. Integrated regression ML schematic diagram. The "valid", "train", and "pred" represent the test set, train set, and prediction values of regression models.

Table 1
The T_c prediction performances via basic model-based ML methods.

Model	R^2	RMSE (K)
LR	68.2%	17.8
Lasso	73.4%	16.0
SVR	77.3%	14.9
KNN	91.6%	9.0
LGB	93.4%	8.1
RF	92.5%	8.4
ET	93.0%	8.2
GBDT	92.9%	8.3
XGB	87.8%	10.9
DT	91.0%	9.5
MLP	91.3%	9.5

prediction results. LGB model showed a R^2 score of 93.4% and RMSE of 8.1 K, showing the higher prediction accuracy in comparison to other models. In addition, the ET and GBDT models reached R^2 scores of 93.0% and 92.9%, respectively, which were the second and third highest performances. The comparison of T_c prediction results of eleven models is shown in Table 1.

The average marginal contributions of material features towards T_c values were estimated. LGB model was selected for the estimation by using SHAP, due to its high prediction performance (R^2 > 93%). Fig. 3 showed the SHAP values derived from the LGB model, where the features were ordered based on the degree of influence on model output. The degree of influence was defined by the mean absolute SHAP value of all the point in the dataset. The correlation between T_c values and material features could be evaluated by the SHAP value. The explication of each feature was specified in Table S4. The SHAP with a positive value indicates the positive correlation with T_c values, and vice versa. It is apparent that the range thermal conductivity (the range value of thermal conductivity among elements in the composition) was the most important feature with the highest SHAP value, followed by the avg_dev Gs volume (average deviation of DFT ground state magnetic moment among elements in the composition) and mean N unfilled (mean of number of unfilled valence orbitals among elements in the composition). Meanwhile, the rang thermal conductivity feature with high values (i. e., red area) exhibited a positive correlation with T_c values, while the negative correlation was recorded for the data points with low thermal conductivity values. The thermal conductivity of superconductors typically offers important insights of electron–phonon couplings, which behaves as the basis of the famous Bardeen-Cooper-Schrieffer (BCS) theory. In this regime, the condensation of the Cooper pairs relies heavily on the interplay between electrons and phonons [63,64]. The mechanisms of the superconductors with high-T_c properties still remain elusive, but the ML model provided potential insights to explore new superconductors when the superconductivity aspects have not been utilized in feature engineering. Similarly, the nelement feature (the number of elements in the composition) with high values (i.e., res area) exhibited a positive correlation with T_c values. Doping played a significant role in optimizing T_c values [65], which is marked the various number of elements in compounds. The addition of cations was discussed as doping mark that can lead to strong electronic density. For example, T_c value for the $Y_1Ba_2Cu_{2.7}Zn_{0.3}O_7$ material is about 23 K. And with the increasing of doping elements number, the $Y_1Ba_2Cu_{2.85}Fe_{0.024}Zn_{0.126}O_7$ has the higher T_c value that is about 46.6 K. In addition, the four (i. e., mean N unfilled (mean of number of unfilled valence orbitals among elements in the composition), avg_dev N unfilled (average deviation of unfilled valence orbitals among elements in the composition), avg_dev Np valence (average deviation of valence *d*-orbitals among elements in the composition), and avg_dev Nd valence (average deviation of number of valence *d*-orbitals among elements in the composition)) of top 20 important features were related to electrons of the atom's extranuclear orbital, and other five features (i.e., avg_dev Gs volume (average deviation of DFT ground state magnetic moment among elements in the composition), entropy atomic volume (weight entropy of atomic volume among elements in the composition), std_dev atom radius (standard deviation of atomic radius among elements in the composition), std_dev atomic column (standard deviation of atomic column among elements in the composition), and std_dev metallic radius (standard deviation of metallic radius among elements in the composition)) were related to the atomic size. It indicated that the T_c values were majorly affected by the electron distribution and valence state of the constituent elements.

Symbolic regression (SR) [66] analysis using a genetic algorithm was performed via gplearn. SR was used to search for a formula that can be generated using normalized features from SHAP value ranking (Fig. 3). The small RMSE represented the high accuracy result, which is suitable for guiding the relationship between T_c values and features by the mathematical formulas (Table S5). The RMSE and R^2 score for SR analysis in different features number

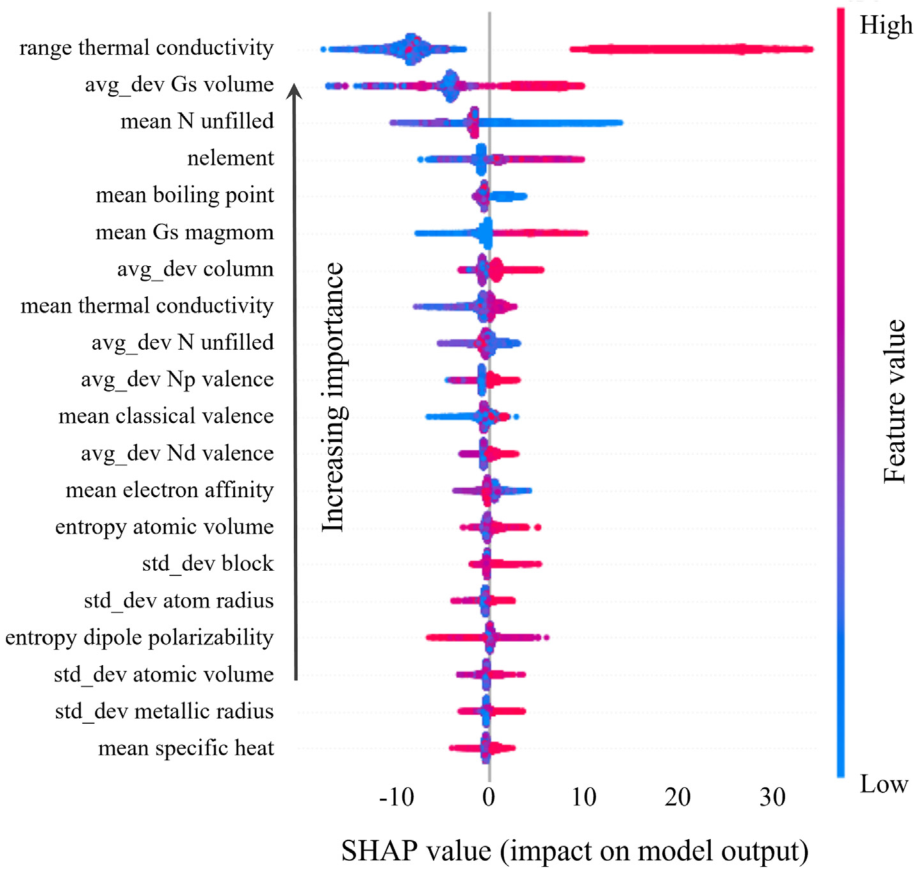

Fig. 3. SHAP plot summarizing 20 features for every point in the dataset, in order of increasing importance (i.e., the sum of SHAP value magnitudes). The color corresponds to the value of each input feature and can demonstrate positive or negative correlation with T_c values. Red and blue color mean the values of listed feature on each data point, respectively.

Table 2
Comparison of T_c prediction performances of basic model- and integrated model-based ML methods.

Model	R^2	RMSE (K)
GBDT	92.9%	8.3
ET	93.0%	8.2
LGB	93.4%	8.1
Integrated ML model	95.9%	6.3

are demonstrated in Fig. S2. When the features number was set as eight it reached the lowest RMSE value of 15.7 K during SR analysis. SR analysis showed that the mathematical formula, generated from the range thermal conductivity, avg_dev Gs volume, mean N unfilled, nelement, mean builing point, mean Gs magmom, avg_dev column, and mean thermal conductivity, has a more accurate correlation with the T_c values than other feature combinations.

It should be noted that the stacking mode requires that the basic models itself has a high accuracy [21], thus the LGB, ET and

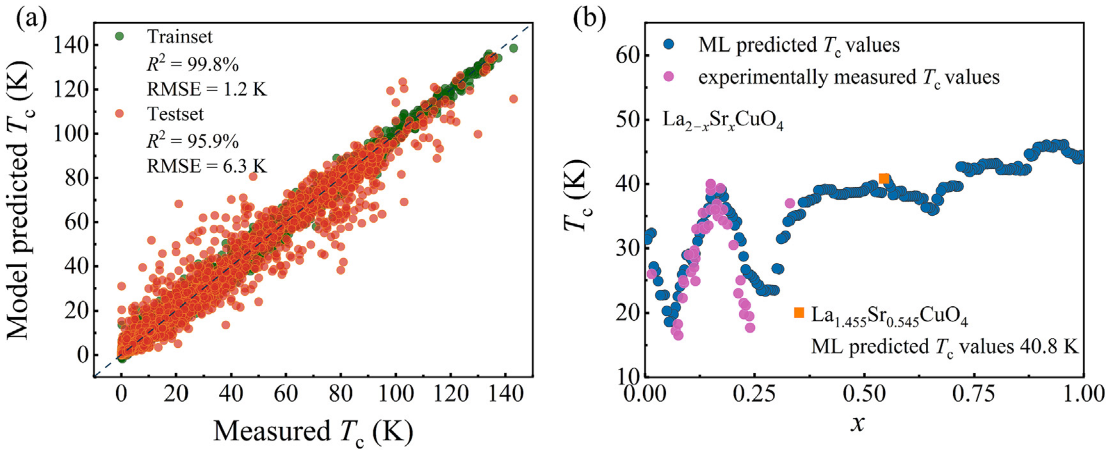

Fig. 4. (a) Comparison of predicted and experimentally measured T_c values. (b) Variation of the T_c values of the $La_{2-x}Sr_xCuO_4$. Purple and blue dots represented the Experimentally measured and ML predicted T_c values, respectively. The orange dot represents the ML predicted T_c values when the × equals to 0.545.

Table 3
Comparison of experimentally measured T_c with the values predicted by different SuperCon ML models. $Bi_{1.66}Pb_{0.34}Sr_2Ca_2Cr_xCu_{3-x}O_{10}$ samples were out of 13,138 data in this work.

Formulas	Experimentally measured T_c (K)	Predicted T_c (K) (This work)	Predicted T_c (K) [20]	Predicted T_c (K) [19]
$Bi_{1.66}Pb_{0.34}Sr_2Ca_2Cr_xCu_{3-x}O_{10}$				
x = 0	105.4	103.9	103.5	106.7
x = 0.005	108.5	103.1	103.4	92.9
x = 0.010	101.2	102.6	102.1	90.8
x = 0.015	/	101.9	101.0	88.3
x = 0.020	/	101.3	101.8	87.7
x = 0.025	/	100.8	101.1	87.4
$Tl_{1-x}Hg_xBa_2Ca_2Cu_3O_8$				
x = 0.1	132	133.4	131.3	112.4
x = 0.2	/	134.5	134.2	113.1
x = 0.3	133	134.8	134.7	114.3
x = 0.4	/	135.3	134.0	115.6
$La_{2-x}Sr_xCuO_4$				
x = 0.15	40	39.6	35.1	34.2
x = 0.545	/	40.8	27.1	18.6

GBDT models were selected in integrated ML model in this work. The integrated ML model results were listed in Table 2, which demonstrated a higher R^2 score of 95.9% in comparison to those of GBDT (92.9%), ET (93.0%), LGB (93.4%) models and eleven-mentioned model (94.6%, Table S6). Prediction results showed that the integrated model-based ML method obtained the optimal value in the ensemble stage and plot the scatter diagram of train and test set (Fig. 4a). To further verify the accuracy of integrated ML model, T_c values of the $La_{2-x}Sr_xCu_1O_4$ were predicted with variation of the component x ($0 < x < 1$) (Fig. 4b). When the x ranged from 0.0 to 0.33, the experimentally measured T_c values corresponds well with the predicted T_c values, which indicated an excellent prediction accuracy. However, there is no recorded T_c data for $La_{2-x}Sr_xCuO_4$ materials with x beyond 0.33. When doping content of Sr is between 0.00 and 1.00, the crystal structure can be shown as the orthorhombic or tetragonal structure by X-ray diffraction patterns [67]. Noticeably, the T_c values of $La_{2-x}Sr_xCuO_4$ with x ranging between 0.33 and 1.0 were also predicted through the integrated model-based ML method. When the x equals to 0.545, the ML predicted T_c values were 40.8 K (Fig. 4b). This provides reasonable information for further investigations of $La_{2-x}Sr_xCuO_4$ materials. Furthermore, T_c values of the other 3 superconductor materials with different component were predicted via the integrated ML model, and the prediction results were compared with the T_c values obtained by experiments and various ML models (Table 3). The comparison of model performances in terms of dataset size, the number of features, and R^2 score were listed in Table S7. The predicted T_c values via integrated model corresponds well with the recorded values in both reported other ML models [19,20]. And based on the SuperCon database, $Hg_{0.66}Pb_{0.43}Ba_2Ca_{1.98}Cu_{2.9}O_{8.4}$ currently showed the highest T_c value of 143 K that the predicted T_c is 138.6 K by the integrated ML model. The $Tl_{1-x}Hg_xBa_2Ca_2Cu_3O_8$ compounds were predicted to reach a high T_c value of 135.3 K while the x = 0.4, which is quite close to the highest recorded value of 143 K for $Hg_{0.66}Pb_{0.43}Ba_2Ca_{1.98}Cu_{2.9}O_{8.4}$. It is expected that more high-T_c superconductors with different composition can be mined from the known superconductors.

3.3. New materials mining

The goal of modeling is to predict unreported and potential superconductor materials with high T_c values. 9000 compounds were extracted from the MP database for the new superconductor materials mining. Then 115 features were added to each data by following the method used in feature engineering. In order to explore superconductors with T_c values beyond 50.0 K, the integrated model-based ML method was used to predict potential materials in selected compounds. In addition, the common cuprate- and iron-based compounds were eliminated, aiming to find unconventional and high-performance superconductors. The predicted values showed that the T_c values of 20 compounds were higher than 50.0 K and the Ba_4AgAuO_6 reached the highest value of 70.8 K. The corresponding MP-id, chemical formula, and crystal system were listed in Table 4. In terms of the crystal system, the proportion of cubic, orthogonal, tetragonal, monoclinic, and trigonal materials were 52.6%, 21.3%, 15.7%, 2.6%, and 7.8%, respectively. In the view point of element composition, Ag was observed in every predicted compound, and O and F were existed in 44.7% and 47.3% of compounds.

The MP provided the electrical characteristics which estimated the additional insight with the probable connection between these candidates. The energy band and density of states (DOS) diagrams were illustrated for visual analysis of the predicted superconductor materials with highest T_c values. Taking Ba_4AgAuO_6 and $KAgCO_3$ as examples to shown in Fig. 5, the energy bands neared the Fermi level (E_F) that appeared the flat bands. These bands caused a large rise in the DOS and can result in a significant increase in T_c values.

Table 4
Predicted superconductor materials with T_c beyond 50 K via integrated model-based ML method (common Cu/Fe-based materials were excluded).

MP - id	Chemical formula	Predicted of T_c values (K)	Crystal system
mp-556896	Ba_4AgAuO_6	70.8	Orthorhombic
mp-1239304	$Ba_2YAg_3O_8$	69.8	Tetragonal
mp-8666	$CsAgO$	67.6	Tetragonal
mp-683972	$Cs_5Ag_4C_8IN_8$	67.0	Cubic
mp-572510	K_3AgO_2	65.3	Orthorhombic
mp-1096933	$CsAgO_2$	63.0	Orthorhombic
mp-19378	$CrAgO_2$	63.0	Trigonal
mp-553907	Rb_3AgO_2	61.7	Orthorhombic
mp-997052	$RbAgO_2$	61.2	Orthorhombic
mp-3074	$KAgO$	61.0	Tetragonal
mp-541966	$KAgCO_3$	61.0	Orthorhombic
mp-8603	$RbAgO$	60.8	Tetragonal
mp-557862	$BaAg_2(HgO_2)_2$	58.7	Tetragonal
mp-997088	$KAgO_2$	58.5	Orthorhombic
mp-643123	$K_5Ag(NO)_2$	58.4	Monoclinic
mp-1114287	K_2TaAgF_6	58.2	Cubic
mp-1112462	K_2AgIrF_6	57.7	Cubic
mp-6855	$K_2NaAg_3(CN)_6$	56.5	Trigonal
mp-1253888	$Ba_2AlAg_3O_8$	55.3	Tetragonal
mp-976229	$KAgO_3$	55.2	Cubic

Note: When the compounds have more than one crystal structures, only the most common one is retained.

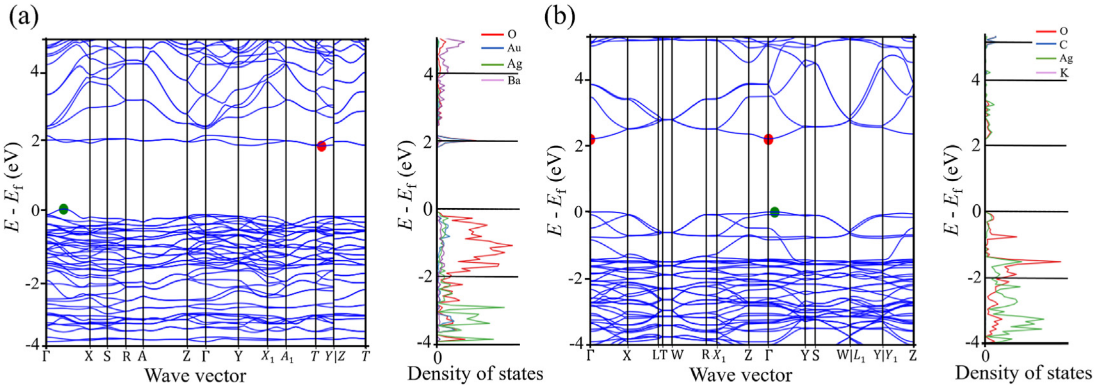

Fig. 5. Energy band and DOS diagrams of predicted superconductor materials. (a) Ba_4AgAuO_6 compound (70.8 K). (b) $KAgCO_3$ compound (61.0 K).

Fig. 5 indicated the peaks in the DOS elicited by van Hove singularities (VHS) can be significantly increased the T_c values when the VHS were sufficiently closer to E_F [68]. And the experiments have been proved that the high T_c superconductors VHS were near the E_F [69,70]. The ML model explored these band structure characteristics when the no explicit information about the electronic band structure were included in these features. Therefore, T_c values can be increased when the distance between E_F and its adjacent VHS was shorten by changed or doped material composition.

4. Conclusions

In summary, the integrated ML model was developed to predict potential high-T_c superconductors. 13,138 pieces of superconducting data were collected from SuperCon database as the dataset for the element features engineering in ML model training. A 95.9% prediction accuracy of R^2 score was reached via the integrated ML model cross-verification. The range thermal conductivity feature with high values exhibited a positive correlation forward T_c value. Twenty new superconductors with T_c values over 50.0 K were predicted by using the integrated ML model. Moreover, the work successfully extracted effective features mathematical formulas to estimate T_c values. This work provides new insights for improving the prediction accuracy of T_c values, and further explores new type ML methods to screen out potential high-T_c superconductors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors appreciate financial supports from the Fund of Science and Technology on Reactor Fuel and Materials Laboratory (JCKYS2019201074), the Affiliated Hospital of Putian University, the Shenzhen Fundamental Research Program (JCYJ20220531095404009), the Shenzhen Knowledge Innovation Plan-Fundamental Research (Discipline Distribution) (JCYJ20180507184623297), and the Major Science and Technology Infrastructure Project of Material Genome Big-science Facilities Platform supported by Municipal Development and Reform Commission of Shenzhen.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jechem.2022.11.047.

References

[1] P. Kong, V.S. Minkov, M.A. Kuzovnikov, A.P. Drozdov, S.P. Besedin, S. Mozaffari, L. Balicas, F.F. Balakirev, V.B. Prakapenka, Nat. Commun. 12 (2021) 1–9.
[2] M.C. Diamantini, C.A. Trugenberger, V.M. Vinokur, Adv. Quantum. Technol. 4 (2021) 2000135.
[3] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, Phys. Rev. Lett. 127 (2021).
[4] X. Zhou, W.-S. Lee, M. Imada, N. Trivedi, P. Phillips, H.-Y. Kee, P. Törmä, M. Eremets, Nat. Rev. Phys. 3 (2021) 462–465.
[5] J. Paglione, R.L. Greene, Nat. Phys. 6 (2010) 645–658.
[6] J.A. Flores-Livas, L. Boeri, A. Sanna, G. Profeta, R. Arita, M. Eremets, Phys. Rep. 856 (2020) 1–78.
[7] C. Heil, L. Boeri, Phys. Rev. B. 92 (2015).
[8] I.A. Troyan, D.V. Semenok, A.G. Kvashnin, A.V. Sadakov, O.A. Sobolevskiy, V.M. Pudalov, A.G. Ivanova, V.B. Prakapenka, E. Greenberg, A.G. Gavriliuk, Adv. Mater. 33 (2021) 2006832.
[9] C. Pellegrini, H. Glawe, A. Sanna, Phys. Rev. Mater. 3 (2019).
[10] J. Alarco, P. Talbot, I. Mackinnon, J. Phys. Conf. Ser. 1143 (2018).
[11] W. Feng, R. Zhao, X. Wang, B. Xing, Y. Zhang, X. He, L. Zhang, J. Energy Chem. 70 (2022) 1–8.
[12] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Nat. Commun. 9 (2018) 3405.
[13] Q. Tao, T. Lu, Y. Sheng, L. Li, W. Lu, M. Li, J. Energy Chem. 60 (2021) 351–359.
[14] M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A.S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C.-S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S.-C. Lo, A. Ip, Z. Ulissi, E.H. Sargent, Nature. 581 (2020) 178–183.
[15] Y. Kang, L. Li, B. Li, J. Energy Chem. 54 (2021) 72–88.
[16] X. Jia, Y. Deng, X. Bao, H. Yao, S. Li, Z. Li, C. Chen, X. Wang, J. Mao, F. Cao, npj Comput. Mater. 8 (2022) 1–9.
[17] A. Mannodi-Kanakkithodi, G. Pilania, R. Ramprasad, Comput. Mater. Sci. 125 (2016) 123–135.
[18] T.O. Owolabi, K.O. Akande, S.O. Olatunji, J. Supercond. Novel Magn. 28 (2015) 75–81.
[19] V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, npj Comput. Mater. 4 (2018) 29–40.
[20] S. Zeng, Y. Zhao, G. Li, R. Wang, X. Wang, J. Ni, npj Comput. Mater. 5 (2019) 1–7.
[21] J. Zhang, Z. Zhu, X.-D. Xiang, K. Zhang, S. Huang, C. Zhong, H.-J. Qiu, K. Hu, X. Lin, J. Phys. Chem. C. 126 (2022) 8922–8927.
[22] D. Maulud, A.M. Abdulazeez, J. Appl. Sci. Technol. Tre. 1 (2020) 140–147.
[23] Z.L. Liu, P. Kang, Y. Zhu, L. Liu, H. Guo, APL Mater. 8 (2020).
[24] J. Schmidt, M.R. Marques, S. Botti, M.A. Marques, npj Comput. Mater. 5 (2019) 1–36.
[25] Z. Guo, B. Lin, Sol. Energy 228 (2021) 689–699.
[26] T. Wang, K. Zhang, J. Thé, H.Y. Comput, Mater. Sci. 201 (2022).
[27] J. Chen, J. Yin, L. Zang, T. Zhang, M. Zhao, Sci. Total Environ. 697 (2019).
[28] T. Hastie, R. Tibshirani, J. Friedman, Boosting and Additive Trees, second ed., Springer, New York, 2009.
[29] P. Geurts, D. Ernst, L. Wehenkel, Mach. Learn 63 (2006) 3–42.
[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Adv. Neural. Inf. Process. Syst. 30 (2017) 3149–13144.
[31] D.V. Lindley, A.F. Smith, J.R. Stat, Soc. B 34 (1972) 1–18.
[32] J. Ranstam, J. Cook, Br. J. Surg. 105 (2018) 1348.
[33] S.A. Dudani, IEEE Trans. Syst. Man Cybern. 6 (1976) 325–327.

[34] D.A. Pisner, D.M. Schnyer, Mach. Learn (2020) 101–121.
[35] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, J. Chemom. 18 (2004) 275–285.
[36] L. Breiman, Mach. Learn 45 (2001) 5–32.
[37] T. Chen, C. Guestrin, Xgboost, A scalable tree boosting system, in: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, New York, 2016, pp. 785–794.
[38] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer, New York, 2009.
[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Mach. Learn. Res 12 (2011) 2825–2830.
[40] W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media Inc., Sebastopol, 2012.
[41] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, Artif. Intell. Rev. 54 (2021) 1937–1967.
[42] L. Abhishek, Optical character recognition using ensemble of SVM, MLP and extra trees classifier, in: International Conference for Emerging Technology, Belgaum, 2020, pp. 1–4.
[43] B.A. Goldstein, E.C. Polley, F.B. Briggs, Stat. Appl. Genet. Mol. Biol. 10 (2011) 1544–6115.
[44] J. Wang, P. Li, R. Ran, Y. Che, Y. Zhou, Appl. Sci. 8 (2018) 689.
[45] S. Džeroski, B. Ženko, Mach. Learn 54 (2004) 255–273.
[46] B. Pavlyshenko, Using stacking approaches for machine learning models, in: IEEE Second International Conference on Data Stream Mining & Processing, Lviv, 2018, pp. 255−258.
[47] M. Tavana, K. Puranam, in: Handbook of research on organizational transformations through big data analytics, IGi Global, Inc., Hershey, 2014, pp. 43–53.
[48] Materials Information Station, http://supercon.nims.go.jp/index_en.html (accessed 20th October 2021).
[49] K. Hamidieh, Comput. Mater. Sci 154 (2018) 346–354.
[50] H. Jeon, S. Oh, Appl. Sci. 10 (2020) 3211.
[51] M.V. García, J.L. Aznarte, Ecol. Inf. 56 (2020).
[52] J. Xiong, T.-Y. Zhang, J. Mater. Sci. Technol. 121 (2022) 99–104.
[53] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, APL Mater. 1 (2013).
[54] X. Jia, S. Li, Z. Zhang, Y. Deng, X. Li, Y. Cao, Y. Yan, J. Mao, J. Yang, Q. Zhang, Mater. Today Phys. 18 (2021).
[55] Y. Zhang, C. Wen, C. Wang, S. Antonov, D. Xue, Y. Bai, Y. Su, Acta Mater. 185 (2020) 528–539.
[56] WebElements, https://www.webelements.com (accessed 28th October 2022).
[57] A.A. Semenova, A.B. Tarasov, E.A. Goodilin, Mendeleev Commun. 29 (2019) 479–485.
[58] L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G.J. Snyder, I. Foster, A. Jain, Comput. Mater. Sci 152 (2018) 60–69.
[59] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, npj Comput. Mater. 2 (2016) 16028.
[60] A.O. Furmanchuk, J.E. Saal, J.W. Doak, G.B. Olson, A. Choudhary, A. Agrawal, J. Comput. Chem. 39 (2018) 191–202.
[61] L. Yu, H. Liu, J. Mach. Learn. Res 5 (2004) 1205–1224.
[62] C. Ju, A. Bibaut, M. van der Laan, J. Appl. Stat. 45 (2018) 2800–2818.
[63] A. Gabovich, A. Voitenko, M. Ausloos, Phys. Rep. 367 (2002) 583–709.
[64] S. Bouscher, Z. Kang, K. Balasubramanian, D. Panna, P. Yu, X. Chen, A. Hayat, J. Phys.: Condens. Matter 32 (2020).
[65] H. Hilgenkamp, C.W. Schneider, R.R. Schulz, B. Götz, A. Schmehl, H. Bielefeldt, J. Mannhart, Physica C 326 (1999) 7–11.
[66] Y. Wang, N. Wagner, J.M. Rondinelli, MRS Commun 9 (2019) 793–805.
[67] J. Zhu, Z. Zhao, D. Xiao, J. Li, X. Yang, Y. Wu, Mater. Chem. Phys. 94 (2005) 257–260.
[68] Y. Quan, W.E. Pickett, Phys. Rev. B 93 (2016).
[69] R. Markiewicz, Int. J. Mod. Phys. 5 (1991) 2037–2071.
[70] J. Labbé, S. Barišić, J. Friedel, Phys. Rev. Lett. 19 (1967) 1039.