ADVANCED THEORY AND SIMULATIONS
www.advtheorysimul.com

Multi-Layer Feature Selection Incorporating Weighted Score-Based Expert Knowledge toward Modeling Materials with Targeted Properties

*Yue Liu, Jun-Ming Wu, Maxim Avdeev, and Si-Qi Shi**

Selecting proper descriptors or features is one of the central problems in exploring structure–activity relationships of materials using machine learning models. The current feature selection algorithms usually require tedious hyperparameter tuning and do not actively consider the prior knowledge of domain experts about the features. Here, this work proposes a data-driven multi-layer feature selection method incorporating domain expert knowledge named DML-FS$_{dek}$, which is automated, with users entering training data without manual tuning of the hyperparameters. The domain expert knowledge is quantified by means of weighted scoring and integrated into the selection process to eliminate the risk of crucial features being removed. The test studies on ten material properties datasets demonstrate the potential of the approach to automatically search for a reduced feature set with lower root mean square errors than those for the initial feature set. Essentially, the most relevant material features, the number of which is much smaller than that in the original feature set, are automatically selected to establish a closer and more accurate structure–activity relationship for the materials of interest. As a result, the method represents the targeted properties of materials with a smaller and more interpretable set of features while ensuring equal or better prediction accuracy.

1. Introduction

Machine learning (ML) is rapidly gaining popularity as an approach to accelerate the design and development of advanced materials. It has been used for material properties prediction and optimization,[1–3] new materials discovery,[4] improvement of parameters for computational studies,[5] and has proved its efficiency and accuracy. The universal procedure of ML in material properties prediction is schematically illustrated in **Figure 1**. The representations of a material dataset, called "descriptors" or "features," not only uniquely define each material in the input dataset but also correlate with its target properties.[6–8] One of the critical aspects of constructing a machine learning model is to select appropriate descriptors to reflect material properties.[9] Ideally, only the relevant features are picked and redundant and irrelevant features are discarded as they reduce the prediction performance of the ML model and increase computational complexity.

For instance, in ion conductivity prediction of lithium battery materials,[10] crystal enthalpy does not correlate with ion conductivity and thus should be regarded as irrelevant attribute for an ML model. A more complicated example is lattice constant prediction.[11] When three features such as composition, average coordination number, and atomic valence are present in the original feature set simultaneously, all of them strongly correlate with the target property, that is, lattice constant. However, it is also known that the average coordination number can be predicted for a given composition via the known atomic radii and valence. In this case, the average coordination number can be eliminated as a redundant attribute. In summary, in predicting

Prof. Y. Liu, Dr. J.-M. Wu
School of Computer Engineering and Science, Shanghai Institute for Advanced Communication and Data Science
Shanghai University
Shanghai 200444, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adts.201900215

DOI: 10.1002/adts.201900215

Prof. M. Avdeev
Australian Nuclear Science and Technology Organisation
Locked Bag 2001
Kirrawee DC NSW 2232, Australia

Prof. M. Avdeev
School of Chemistry
The University of Sydney
Sydney 2006, Australia

Prof. S.-Q. Shi
School of Materials Science and Engineering, Materials Genome Institute
Shanghai University
Shanghai 200444, China
E-mail: sqshi@shu.edu.cn

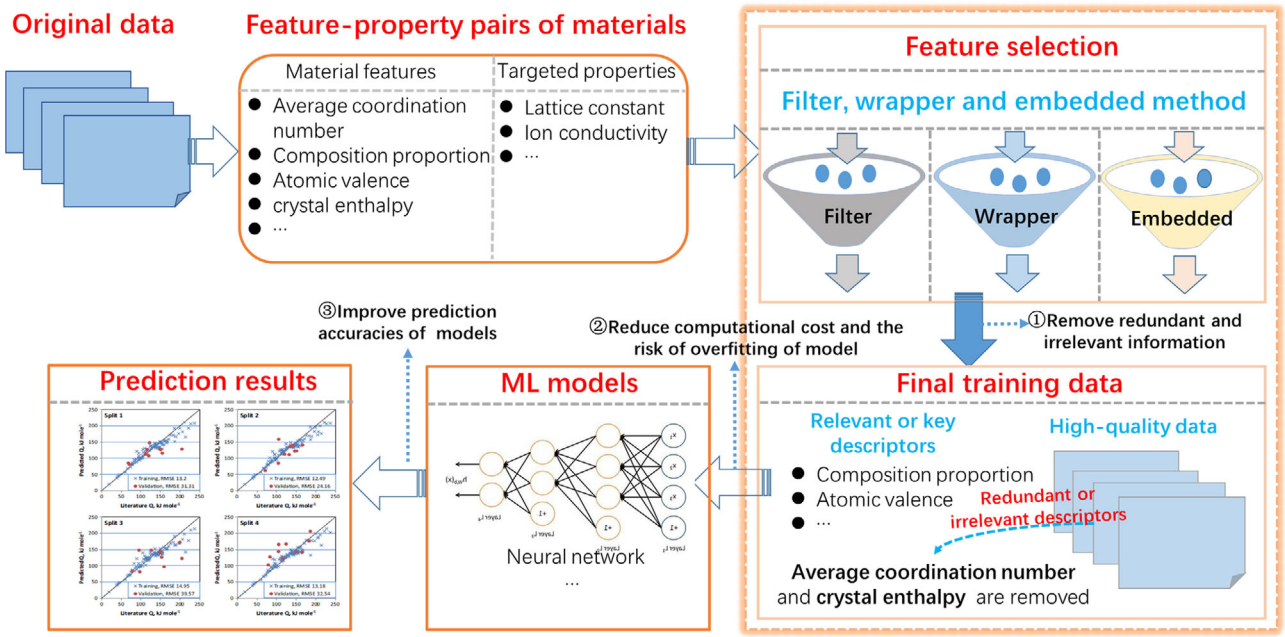


Figure 1. The procedure of combining machine learning with feature selection (FS) in materials properties prediction. FS removes redundant and irrelevant information present in the original dataset, reduces computational cost and the risk of overfitting for the prediction model, and improves its prediction accuracy.

material properties, complex and unclear correlations exist not only between the features and target properties but also among features themselves and it is essential to identify and eliminate the irrelevant and redundant features and retain only the representative features of the original feature set.

Feature selection (FS) is one of the most important steps in the machine learning process for constructing quantitative structure-activity relationships,[12–14] as identification and ranking of the most relevant features greatly affect the computational speed and predictive ability and interpretability of the model.[12,15] Recently, in the machine learning study of the thermodynamic stability of perovskite oxides, Morgan et al. applied three different FS methods, that is, stability selection, recursive feature elimination, and univariate feature selection based on mutual information.[16] Based on the basic principles of these algorithms, we will refer them in this work as Wrapper, Embedded, and Filter methods, respectively. By applying these methods, the authors were able to reduce the initial 791 features to 70 and construct an ML model without significant overfitting. In order to avoid overfitting, Zhang et al.[17] employed L1 (Wrapper) method for feature selection before model training. After a series of parameter adjustment and model selection, the model with good generalization performance was achieved.

In summary, many researchers in the field of computational materials science have begun to adopt a variety of FS algorithms to quantify the relevance of material descriptors to the properties of interest. Nevertheless, given the diversity of the available FS methods, domain experts usually face a problem of choosing the appropriate methods. Additionally, even if a FS method that is suitable for handling a certain type of domain problem is determined, the hyperparameters and strategies involved require manual setting and adjustment, which is usually time-consuming and labor-intensive. For example, for the filter methods, users usually manually define the number of selected features and the filtering threshold, the wrapper methods need manual input of the subset search strategy to generate a candidate feature subset. The embedded methods employ the machine-learning algorithms, such as Lasso and gradient boosting regression (GBR), to measure importance of the features, in which the hyperparameters of the algorithm (e.g., the number and max-depth of decision trees in GBR) also need to be manually searched and optimized to achieve better performance. Consequently, domain experts may face difficulties in selecting easy-to-use but accurate methods, as parameter adjustment requires practical experience in using machine learning. On the opposite side, the prior knowledge of domain experts on the importance or relevance of the features is typically ignored in the FS process, even though domain experts may know in advance which features are more important, which leads to the reduction of the model development efficiency and its predictive ability.

Although such information can be introduced into some machine learning models such as support vector machines (SVMs) to indirectly help select features,[18] in general domain experts' prior knowledge is rarely incorporated into the selection procedure of material features. Therefore, it would be very useful to develop an automatic feature selection approach combining with domain expert knowledge.

Herein, we propose a multi-layer approach utilizing the intrinsic characteristics of data to evaluate the importance of features from different perspectives to eliminate the irrelevant and redundant features from the original training set. The whole process is automated and does not require the user to have experience in feature selection. Moreover, the integration of domain expertise prevents key features from being ignored. In **Table 1**,

Table 1. The comparison of the filter, wrapper, embedded, and our approach.

Items	The filter	The wrapper or embedded	Our approach
Processing speed	Slow	Fast	Slow
Prediction accuracy	Low	High	High
Incorporates domain expert knowledge	No	No	Yes
Requires setting parameters/Automated	Yes/No	Yes/No	No/Yes

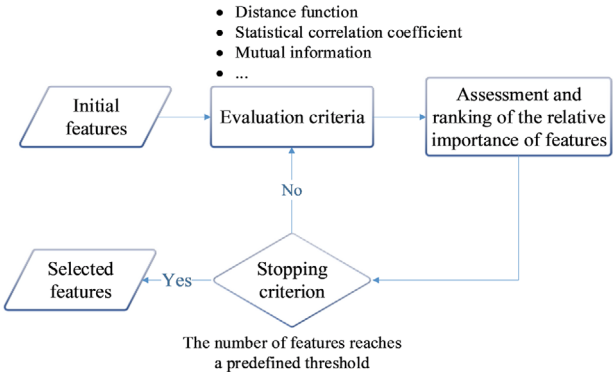


Figure 2. The flow chart of the filter.

we compare the filter, wrapper, and embedded methods and our approach. As we demonstrate below, our method has advantages except for slow processing speed, that can be mitigated in the future through parallelization. The results discussed below in detail indicate that the proposed method can successfully replace time-consuming trial-and-error process of model hyperparameters and provide equal or better prediction performance with a smaller in size and interpretable feature set.

The remainder of the paper is structured as follows: Section 2 briefly reviews the existing feature selection methods (Filter, Wrapper, and Embedded) and their working principles. Section 3 describes the details of the method proposed in this work. Section 4 demonstrates the effectiveness and feasibility of the method for ten material property datasets. Finally, the conclusions of this study are given in Section 5.

2. Preliminaries

The overall goal of feature selection (FS) is to identify a subset of features that contains the most representative information on the properties of interest in the original data. In recent years, the feature selection (FS) process has attracted widespread attention due to its importance for further analysis and understanding the data. Thus, several approaches have been presented, which can be generally grouped into three categories: filter methods, wrapper methods, and embedded methods.[19]

Figure 2 presents the working procedure of the filters. Filter methods[20,21] attempt to use evaluation criteria based on statistical theory and information theory such as distance function,[22] statistical correlation coefficient,[23] mutual information,[24] etc.,

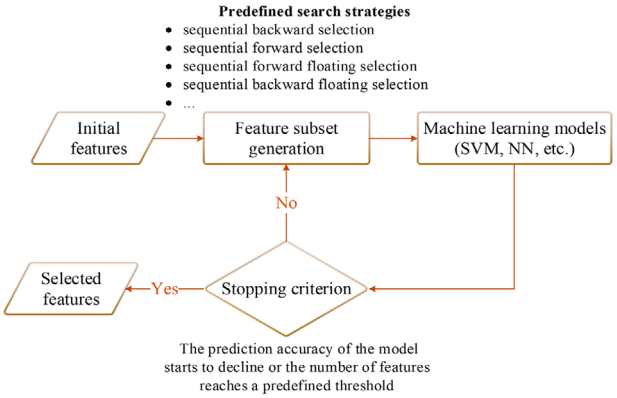


Figure 3. The flow chart of the wrapper.

to assess the relevance of the features and rank them according to their importance. Then the features with high scores are used in the ML model. The advantages of the filter approach are its simplicity and efficiency. Nevertheless, a common disadvantage is that selection process is decoupled from the classifier used to further build the predictor and ignores the effects of a selected feature subset on the performance of the ML model that in general leads to its lower prediction accuracy.[25]

In contrast to the filter approach, wrapper methods[25,26] use the prediction performance of a machine-learning model (e.g., support vector machine (SVM), neural network (NN)) as criteria to evaluate the quality of the candidate feature subset. The workflow of wrappers is schematically illustrated in **Figure 3**. First, the wrapper generates an initial candidate feature subset based on the predefined search strategies, such as sequential backward selection, sequential forward selection, sequential forward floating selection, or sequential backward floating selection, and then a ML model is trained and tested to estimate the candidate feature subset. This process is performed iteratively until the selected feature subset meets the specified requirement. The better results and higher prediction accuracy of the wrappers are achieved at the cost of computational time and complexity.

Embedded methods[27,28] utilize types of ML models, such as linear model (Linear, Lasso, or Ridge regression), support vector machine (SVM) and random forest, to guide the feature selection process and define a criterion depending on a class of regression or classification function.

In general, filter methods are faster than the wrapper and embedded methods in terms of processing speed but produce inferior results because they are independent of the specific ML algorithms.

3. The Principles of the Proposed DML-FS$_{dek}$ Method

To our knowledge, many non-machine learning experts often rely on either a tedious trial-and-error process or personal biased experience in choosing feature selection (FS) methods. Moreover, current FS algorithms tend to ignore the prior knowledge of domain experts about what features are more relevant which may lead to removal of some crucial features. Hence, we develop a

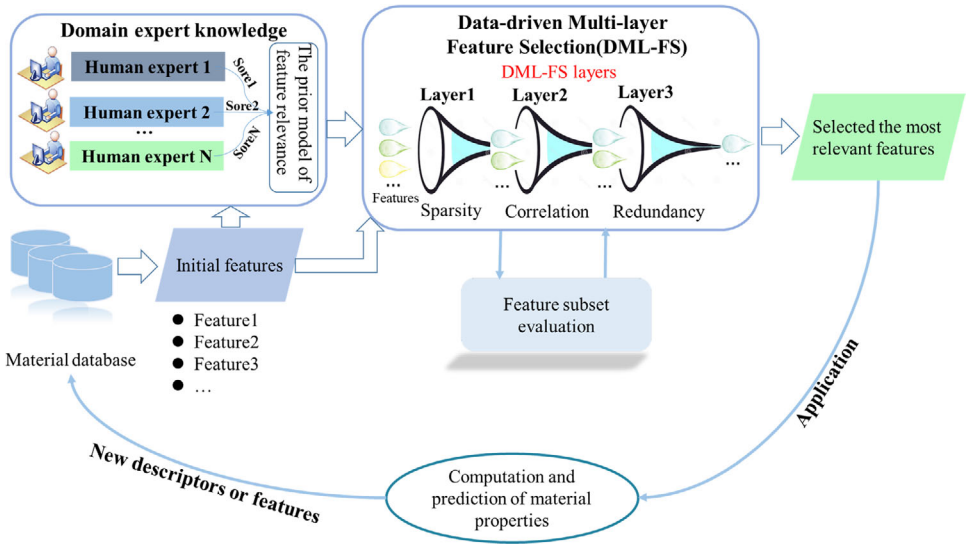


Figure 4. The framework of our proposed approach.

novel multi-layer feature selection method, DML-FS$_{dek}$, that incorporates domain expert knowledge.

The overall framework of our approach is illustrated in **Figure 4**. First, based on the domain expert knowledge applied to the initial features in the material database, an a priori model of feature relevance is constructed and used to drive the process of DML-FS$_{dek}$. Then, in DML-FS$_{dek}$, the problems of sparsity, irrelevance, and redundancy of the input data can be hierarchically addressed by three corresponding processing layers, which ultimately ensures that selected features are highly differentiated and highly correlated with the target attributes. To sum up, the key idea behind DML-FS$_{dek}$ is that statistical analysis method and information theory (see Section 3.1.1) are employed to analyze the relationship between features and target attributes to remove the most interrelated (redundant) features or the least relevant to the target attributes. Further, in feature subset evaluation (see Section 3.1.2), each layer generates an initial subset of features (candidate feature subset) using a default initial filtering threshold ε, and then constructs a specific ML model to evaluate the subset. If it is better than the previous subset of features, then the threshold is adaptively updated and a new subset of the features generated. This whole process is iteratively performed until a subset of features that meets the specific requirements is found. Finally, the best subset of features is picked considering domain expert knowledge and used for subsequent prediction of material properties.

The design of the DML-FS$_{dek}$ three-layer structure, quantitative representation of domain expert knowledge, and strategies to combine them are covered in the following subsections.

3.1. Data-Driven Multi-Layer Feature Selection

3.1.1. The Trigger Conditions Design for DML-FS$_{dek}$ Layers

The proposed DML-FS$_{dek}$ includes the three processing layers: sparsity evaluation, correlation evaluation, and redundancy evaluation, which analyze the importance of features from different perspectives depending on the characteristics of the data. Thus, the trigger conditions for DML-FS$_{dek}$ layers are designed as follows.

$$\text{Multi\_Layer}(X) = \begin{cases} \text{Layer}_1(X), & \text{if } \exists x_i \in X \text{ and } x_i.\text{sparsity} <= \varepsilon \\ \text{Layer}_2(X), & \text{if } \exists x_i \in X \text{ and } x_i.\text{correlation} <= \varphi \\ \text{Layer}_3(X), & \text{if } \exists x_i \in X \text{ and } x_i.\text{redundancy} >= \gamma \end{cases} \quad (1)$$

where X indicates the input features, x_i indicates the i^{th} feature. ε, φ, γ are the sparsity threshold, correlation threshold, and redundancy threshold, respectively. $\text{Layer}_1(X)$, $\text{Layer}_2(X)$, and$\text{Layer}_3(X)$ will be defined in the following Equations (2), (3), and (9), respectively. In the first layer ($\text{Layer}_1(X)$), in order to address the problem of sparsity in discrete and continuous variables, numerical statistical (NS) method and variance score (VS) are adopted for preprocessing. For a continuous variable, if its variance is close to zero, it indicates that the variable fluctuates in a small range and thus its correlation with the target attribute cannot be precisely assessed and the variable should be ignored. Similarly, if the fraction of a certain value of a discrete variable exceeds 95% of the total number of samples, this also implies that the discrete variable is sparse and can be disregarded. Thus, in this layer, the numerical type of each feature is first determined and then corresponding evaluation criteria are used to calculate the sparsity value. In summary, if the sparsity value of features meets the sparsity threshold, the feature will be discarded, otherwise, retained. The calculation of NS and VS of each feature is defined as follows:

$$\text{Layer}_1(X) = \text{Sparsity}(x_i) = \begin{cases} NS(x_i) = \dfrac{1}{n}\sum_{i=1}^{n} x_i, & \text{if } x_i.\text{type} = \text{bool} \\ VS(x_i) = \dfrac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x_i})^2, & \text{otherwise} \end{cases} \quad (2)$$

where n represents the total number of samples, x_i denotes the i^{th} eigenvector, and $\overline{x_i}$ represents the mean value of the i^{th} eigenvalue.

After the sparse features are removed, the second layer ($\text{Layer}_2(X)$) is used to eliminate irrelevant features only weakly correlated to the target attribute. The Mutual information (MI) and Pearson correlation coefficient (PCC) are chosen to measure the correlation between the features themselves and between the features and target attributes, respectively. From Equation (3), it follows that the corresponding correlation evaluation methods are selected according to different trigger conditions. If the number n of samples is less than or equal to k_1 and the target attribute y is discrete, the MI is used to calculate the correlation value, otherwise, the PCC. Finally, if the correlation between the features and attributes is lower than the correlation threshold, the feature will be discarded, otherwise, retained.

$$\text{Layer}_2(X) = \text{Correlation}(x_i, y) = \begin{cases} \text{MI}(x_i, y), & \text{if } n <= k_1 \text{ or } y.\text{type} = \text{bool} \\ |\text{PCC}(x_i, y)|, & \text{otherwise} \end{cases} \tag{3}$$

$$\text{PCC}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \tag{4}$$

where x_i is the i^{th} feature, y is the target attribute, n is the amount of training data, k_1 is the threshold of data size, Cov() is the covariance and Var() is the variance.

$$\text{H}(Y) = -\sum_{y} \text{P}(y) \log(\text{P}(y)) \tag{5}$$

$$\text{H}(Y|X) = -\sum_{X}\sum_{Y} \text{P}(x, y) \log(\text{P}(y|x)) \tag{6}$$

$$\text{MI}(Y,X) = H(Y) - H(Y|X) \tag{7}$$

In Equation (3), MI evaluates how much information each feature can provide. Equations (4)–(6) are the steps for calculating MI. Equation (5) calculates the uncertainty (information content) in all classes Y. Equation (6) implies that by observing a variable X, the uncertainty in the output Y is reduced. The decrease in uncertainty is given as Equation (7). That gives the mutual information (MI) between Y and X meaning that if X and Y are independent then MI is going to be zero, otherwise they are interdependent. Similarly, PCC measures the correlation between two variables as defined in Equation (4).

The irrelevant features are eliminated in the Layer_2, thus there is strong correlation between the remaining features and target attributes. However, in the subset of features obtained, some of the features may still be correlated with the other features. Hence, in the third layer ($\text{Layer}_3(X)$), DCC (Distance correlation coefficient calculated by Equation (8)) and PCC are used to evaluate the redundancy among features. As shown in Equation (9), if the number n of samples is less than or equal to k_1 or the number d for the features is less than or equal to k_2, the DCC is used to calculate the correlation coefficient (redundancy) among features, otherwise, the PCC is calculated. If the redundancy value is greater than the redundancy filtering threshold, one of the two features will be removed, otherwise, the two features will be both retained.

$$\text{DCC}(u, v) = \frac{\hat{d}\,\text{Cov}(u, v)}{\sqrt{\hat{d}\,\text{Var}(u, u)\,\hat{d}\,\text{Var}(v, v)}} \tag{8}$$

where $\hat{d}\,\text{Cov}()$ is distance covariance and $\hat{d}\,\text{Var}()$ is distance variance.

$$\text{Layer}_3(X) = \text{Redundancy}(x_i, x_j) = \begin{cases} \text{DCC}(x_i, x_j), & \text{if } n \le k_1 \text{ or } d \le k_2 \\ |\text{PCC}(x_i, x_j)|, & \text{otherwise} \end{cases} \tag{9}$$

3.1.2. Feature Subset Evaluation Based on Machine Learning Model

The merits of the feature subsets obtained in each layer are evaluated by testing them in ML model and only the best set is passed onto the next layers. **Figure 5** presents the procedure of feature subset evaluation. First, a candidate feature subset is generated; next, the subset evaluation step is performed, which estimates the quality of the current feature set. In this step, the learning model is determined by the user according to specific learning problems, such as support vector machines (SVMs), neural network (NN), decision tree, etc. In addition, evaluation criteria are also adaptively chosen based on different learning problems, such as root mean square error (*RMSE*), mean absolute percentage error (*MAPE*), etc. These steps are performed iteratively, until a stopping criterion is met, which happens either when the results begin to deteriorate or the number of features reaches a predetermined threshold.

3.2. The Weighted Scoring for Domain Expert Knowledge

When the $\text{DML-FS}_{\text{dek}}$ is applied for machine learning in practice, the importance of the features depends, among other things, on the domain knowledge of the user. We quantify that aspect by the score ***s*** that includes the importance score (weight) of the feature given by the user and rating weight of the user himself. The importance score (weight) ***su***, ***sp*** of the feature given by the user is described by

$$\text{su, sp} = \begin{cases} 0 \\ 0.5 \\ 1 \end{cases} \tag{10}$$

where ***su*** represents the current user, ***sp*** represents experienced user, ***su, sp*** $= 0$ indicates that the user thinks the feature is not important, ***su, sp*** $= 0.5$ indicates that the user is uncertain about the importance of the feature, and ***su, sp*** $= 1$ indicates that the user considers the feature is crucial. The rating weight ***d*** of the user is described by

$$d = \begin{cases} 1 \\ 1.5 \\ 2 \end{cases} \tag{11}$$

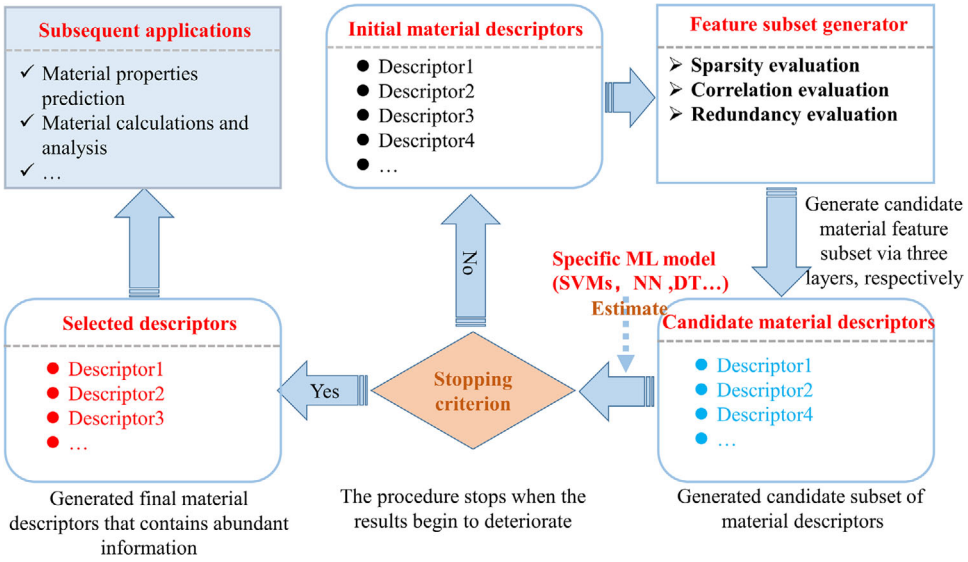


Figure 5. Evaluation procedure for the candidate feature subset of materials.

where $d = 2$ indicates that the user is a material expert, $d = 1.5$ indicates that the user is a computer expert, and $d = 1$ indicates that the user is neither a material expert nor a computer expert. The rating weight is determined according to the users' expertise in the problem domain. The higher the weight is, the more significant the suggestion of this user is. For example, for a problem in the material domain, the material experts have the most abundant domain knowledge and experience, their judgments are the most authoritative, and thus the weight d is given the higher value of 2. For a computer expert, who may also have some background in material science with the rapid development of machine learning in materials discovery and materials properties prediction, consequently, the weight d is given to 1.5. Finally, for the users who are neither material experts nor computer experts, the judgment may not be well founded and thus the weight d is given to 1.

When analyzing data using our method, any kind of human expert can theoretically participate. However, in order to obtain good prediction results, a great deal of domain expertise with high professional level should be acquired. The greater the amount of domain expert knowledge is and the higher the professional level is, the more suitable it is to obtain good results.

3.3. Collaborative Feature Selection between DML-FS and Domain Expert Knowledge

When we perform the feature selection procedure utilizing the proposed DML-FS layers, the importance score ***sa*** of each feature can be described as follows:

$$sa = \begin{cases} 0 \\ 1 \end{cases} \tag{12}$$

where 0 means removing the feature and 1 means retaining the feature. This means that when the DML-FS layer considers that a feature has little or no effect on the result, the feature is scored as 0, otherwise, the feature is scored as 1.

On the other hand, based on the presentation of domain expert knowledge proposed above, the expert experience score s for each feature can be described quantitatively as follows:

$$s = su * \sqrt{1 - \frac{c_2(m-n)^2}{c_1 m^2}} + \frac{\sum sp * d}{\sum d} * \left(1 - \sqrt{1 - \frac{c_2(m-n)^2}{c_1 m^2}}\right) \tag{13}$$

Where m is the total number of people who had scored the feature, n is the number of people who gave the feature the same score as the current user in the m people and $\frac{\sum sp*d}{\sum d}$ is the historical users' score which is the weighted average of the scores of all historical users. The current user's score ***su*** is given a weight $\sqrt{1 - \frac{c_2(m-n)^2}{c_1 m^2}}$ and the historical users' score is given another weight $(1 - \sqrt{1 - \frac{c_2(m-n)^2}{c_1 m^2}})$, where c_1 and c_2 are unknown constants and $c_1 \leq c_2$. The specific values of c_1 and c_2 can be determined based on experimental data and experts' experience to distinguish the credibility of current user' score from historical users' scores. Here, we consider that the parameters c_1 and c_2 are set to 9 and 8, respectively, because in the expert experience score ***s***, the weight of the current user' score (less than 1/2) should be less than the weight of the historical users' scores (more than 1/2) (the minority is subordinate to the majority). Therefore, we can see from the Equation 13 that the range of s is between 0 and 1. Using the weight $\sqrt{1 - \frac{c_2(m-n)^2}{c_1 m^2}}$, we can measure the credibility of historical users' scores. When n is unchanged and m increases, the weight of the current user' score will become smaller, but the

degree of reduction will be reduced. This ensures the dominance of the current user. However, when m does not change and n becomes smaller, the weight of the current user' score will decrease, and the degree of reduction will increase. This reflects the trustworthiness of historical users.

Finally, we embed the domain expert knowledge into the three DML-FS layers. We define the comprehensive importance score of the feature (*FCIS*) as follows:

$$FCIS = sa + s \tag{14}$$

where sa represents the feature importance score obtained through DML-FS layer, s represents the experience score which is rated by domain experts. Therefore, if *FCIS* of the feature is greater than 0.5, the feature would be retained, removed otherwise. The *FCIS* contains the following six core ideas:

1) The current user's experience, historical users' experience, and the result of the DML-FS layer are needed to jointly determine whether to remove the feature;
2) The expert experience score considers the expert's domain issues and establishes weights to divide the differences in each field, which can improve the credibility of the scores;
3) The weighted sum of the current user's score and the historical users' score is 1, which can integrate the experience of the current user with the experience of the historical user;
4) The DML-FS layers result in the fact that the feature should be retained, and the feature will be retained;
5) The current user and historical users consider that the feature is very important, then the feature will not be removed;
6) The current user believes that this feature is unrelated, however, based on the experiences of historical users, this feature has the potential to be retained.

4. Experimental Section

In this section, we evaluate our proposed method on ten material properties datasets. We first introduce the datasets. Then we introduce the parameter setups of the experiments. Finally, we provide the analysis and discussions of experimental results.

4.1. Experimental Datasets

In order to validate the performance of the proposed method, this paper collected ten groups of material properties data sets from the published references or online resources. The brief information on the data sets is shown in **Table 2**, which involves the macro and micro properties of materials. Since all ten datasets have been publicly released, the authenticity and reliability of the data can be guaranteed. Moreover, the ten datasets also cover various sizes and dimensions of samples, which can sufficiently verify the adaptability of the method in different data scenarios. Datasets 1, 2, 7, 8, and 9 have a small amount of data and a low dimension. The dimension of dataset 3 is also relatively low, but the amount of data is relatively large. The original feature sets of datasets 1, 2, 3, 6, 8, and 10 are summarized by domain experts' experiences, so the space for optimization by feature selection method is limited. Datasets 4 and 5 have large data volume and high dimension, and there are many sparse, irrelevant, and redundant features in the two datasets, which are suitable for feature selection algorithm testing.

Table 2. Statistics of the material datasets with targeted properties. N is the number of examples, F is the number of material descriptors, TP is the targeted properties of materials.

Datasets	N	F	TP
1[10]	128	6	The ionic conductivity of Lithium superionic conductors
2[11]	161	6	Lattice constant
3[29]	1302	5	Crystal enthalpy
4[30]	5619	47	The density of organic materials
5[30]	669	18	The viscosity of organic material
6[31]	77	27	Creep fracture life of Ni-based single crystal superalloy
7[32]	160	5	The ionic conductivity of nanocomposite solid polymer electrolyte system
8[33]	117	6	The oxide ionic conductivities in ABO_3 perovskites
9[34]	136	16	Lattice misfit of Ni-based single crystal superalloy
10[35]	9	7	The onset temperature (T_g) of Ge_xSe_{1-x} glass transition

4.2. Experimental Setups

The DML-FS_{dek} is used to perform feature selection experiments on the collected data sets, in which the subset of features obtained from each layer serves as the input of the next layer, and support vector regression (SVR) is employed to evaluate the merit of the feature set. The details of the process are as follows. In the first layer, the sparsity filtering threshold is initially set as 0.01, and the threshold is automatically adjusted with until the prediction accuracy of the model is no longer improved. In the second layer, the correlation filtering threshold is initially set as 0.4, and the threshold is automatically adjusted until the prediction accuracy of the model is no longer improved. Finally, in the third layer, the redundancy filtering threshold is initially set as 0.88, and the threshold is automatically adjusted until the prediction accuracy of the model is no longer improved. To evaluate the generalization performance of the constructed machine learning model on unseen data and reduce the risk of overfitting, fivefold cross-validation is used to assess the comprehensive predictive power of the learning model. In addition, to achieve better prediction performance of the models, some specific search strategies such as random search and grid search, can be used to find the optimal model parameters from the hyperparameter space. Herein, grid search is used to optimize the hyperparameters of the model.

All the algorithms covered in this paper are implemented in Python and call the scikit-learn toolkit.[36]

Table 3. Expert scoring table of features for each group of material dataset (a sample or template).

Domain experts		Feature1	Feature2	Feature3	Feature4	...	Feature*N*
Index	Scoring weight						
1	2	1	1	1	1	...	1
2	2	1	1	0.	1	...	1
3	1.5	0.5	0.5	1	1	...	1
4	1	0.5	0.5	0.5	1	...	1
5	1	0.5	0.5	1	1	....	1
6	2	0.5	0.5	1	1	...	0.5
7	2	1	1	1	1	...	1

4.3. Experimental Results and Analysis

Referring to the scoring rules in Section 3.2, we have invited seven material experts from different fields to score the features contained in ten groups of datasets, and designed an expert scoring table of features shown in **Table 3** for each group of dataset (which has been uploaded to https://github.com/wujunming1/material-attribute-datasets). The expert experience score of each feature on the ten material properties data sets can be calculated through Equation (13) based on the score records of seven experts from different fields, as shown in **Figure 6**. In order to speed up the convergence of our model and eliminate the influence of dimensionality on the prediction accuracy, the Max–Min normalization processing is first performed on all datasets. For comparison, we also conducted the material properties prediction without employing any feature selection method. For all the ten datasets support vector regression (SVR) is used as a predictor. *RMSE* and *MAPE* are used to evaluate the prediction accuracy of the model. The results are listed in **Table 4**. The prediction accuracies of the models on data sets 2, 3, and 4 are relatively high, which indicates that the original features well reflect the distribution of data. The data sets 1, 2, 3, 7, 8, 9, and 10 have fewer original features and limited feature selection space. In addition, when the original feature set is used, the predictive ability of the model on datasets 1, 5, 6, and 8 is relatively poor, and their *RMSE*s even two orders of magnitude higher than those for the dataset 3. Therefore, there is still room for improvement in the prediction accuracy of the model on the four data sets.

Finally, the results and analysis of the sparsity, correlation, redundancy evaluation on the ten data sets are described and analyzed in the following subsections.

4.3.1. Results and Analysis of Sparsity Evaluation

Sparsity evaluation is conducted on the original data, and the results are listed in **Table 5**. It can be observed that the number of features of datasets 1, 2, 3, 8, 9, and 10 has no change, indicating that there are no sparse features in the six data sets. The number of features of dataset 4 is reduced significantly from the original 47 to 34, among which the thirteen features including Y25, Y26, Y22, Y21, Y20, Y19, Y16, Y13, Y12, Y11, Y9, Y8, and Y6, are below the sparsity filtering threshold and thus are eliminated. The number of features of datasets 5 and 6 is reduced from the original 18 (27) to 17 (25), respectively. Among them, only the feature Y5 in dataset 5 is screened out because its calculated variance is close to zero, and only the features Y (mass fraction of the Y element) and a_2time (The second stage aging treatment time) in dataset 6 are also removed because their sparsity evaluation values do not meet the threshold requirement. On dataset 7, the number of features is decreased from a total of five to three, among which the feature X1 and X2 are eliminated. The feature subset evaluation on the selected features shows that the prediction accuracies on the datasets 4, 5, 6, and 7 are slightly improved, which indicates that the sparse features have little influence on the prediction accuracy of the model. Furthermore, when the collaborative selection based on sparsity evaluation layer and domain expert knowledge is performed on the datasets 4, 5, and 7, the number of selected features is different from that of DML-FS layer, which may indicate that some of the features that domain experts consider to be crucial for system modeling are retained. For instance, for dataset 4, the number of features selected through the pure DML-FS layer is 34, while the DML-FS with domain expert knowledge (DML-FS_{dek}) has picked out 38 features in which four key features including Y25, Y22, Y19, and Y16 that domain experts consider important are retained (high expert experience scores, see Figure 6). Moreover, we can observe that the *RMSE* for the two selected feature sets (0.0611 and 0.0581) maintain at the same level. As for dataset 5, the removed feature Y5 is also retained due to the integration of domain expert knowledge.

4.3.2. Results and Analysis of Correlation Evaluation

The results of the correlation evaluation of the remaining features from the first layer are listed in **Table 6**. As can be seen, no features on the datasets 2, 6, 8, and 10 are removed, indicating that there is no irrelevant information in the four datasets. On the other hand, some of the features are discarded due to the weak correlation between features and target attribute in the other six data sets. For dataset 1, the features X3 and X6 are discarded. The feature Specimen Thickness on dataset 9 is removed. As for the dataset 3, the feature X5 is removed. And for the dataset 7, the feature X5 is eliminated. In addition, the number of features is decreased from 34 (17) to 31 (15) for the datasets 4 and 5, respectively. To be specific, the three features including Y18, Y10, and Y4 for the dataset 4, are deleted because their correlations with target property (density of organic materials) are all lower than the correlation filtering threshold. Moreover, the two features G and Y23 are also removed for the dataset 5. Finally, compared with the previous feature subset, the selected features are evaluated via a subset evaluation procedure in this layer, and it can be concluded that the data sets 1, 3, 4, and 5 reach improvement in the prediction accuracy, which is mainly reflected in the decrease of *RMSE*. In particular, the prediction accuracy is improved significantly for the dataset 1 and its *RMSE* is decreased by 19.2%. The above results indicate that uncorrelated features have an impact on the prediction performance of the model. Furthermore, when the collaborative selection based on correlation evaluation layer and domain expert knowledge is performed, the key feature X3 is preserved for the dataset 1 and for the dataset 5 the key feature G is also retained as the result of domain expert knowledge integration.

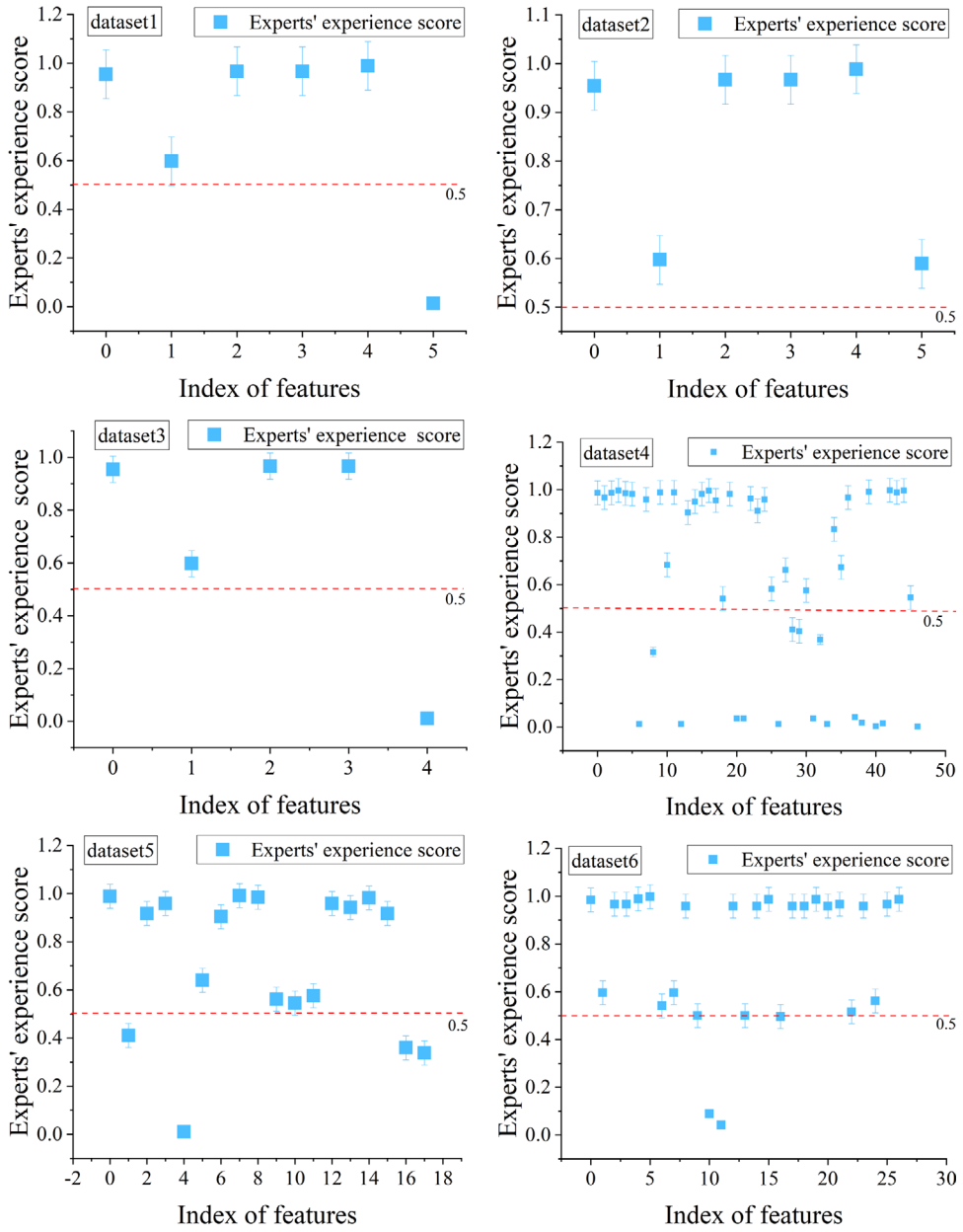


Figure 6. Expert experience scores for each feature in ten material properties data sets. The red dotted line (0.5) represents the dividing line of the experts' experience score. When the experts' experience score for a feature is greater than 0.5, the feature is considered so important by domain experts that it cannot be removed from the model.

4.4. Results and Analysis of Redundancy Evaluation

Redundancy assessment is conducted for the output feature subset from the previous layer to eliminate redundant features and the results are listed in **Table 7**. The number of features for the datasets 1, 2, 3, and 7 do not change, which indicates that the features in the four datasets are independent of each other or very weakly correlated. However, for the datasets 4, 5, 6, 8, 9, and 10, the number of features is decreased. To be specific, the number of features is reduced from 31 to 20 for the dataset 4, among which the eleven features including A, X3, X4, X5, X10, X15, X16, X17, X22, Y1, and Y24 are removed due to their correlation with other features. The size of the feature subset is reduced from 15 to 11 for the dataset 5, in which the four features including A, X11, X13, and Y17 are removed. For the dataset 6, only the feature B (mass fraction of B element) is discarded. For the dataset 8, the feature X3 is removed. Regarding dataset 9, the four features of Ni, Al, W, Ti are removed. The features x, $\langle r \rangle$ and K on dataset 10 are deleted because of their high correlation with other features. Finally, as shown in Table 7, the *RMSE*s of the prediction models for the six data sets are decreased slightly, indicating that redundant features have little impact on the prediction performance. Furthermore, when the collaborative selection based on redundancy evaluation layer and domain expert knowledge is

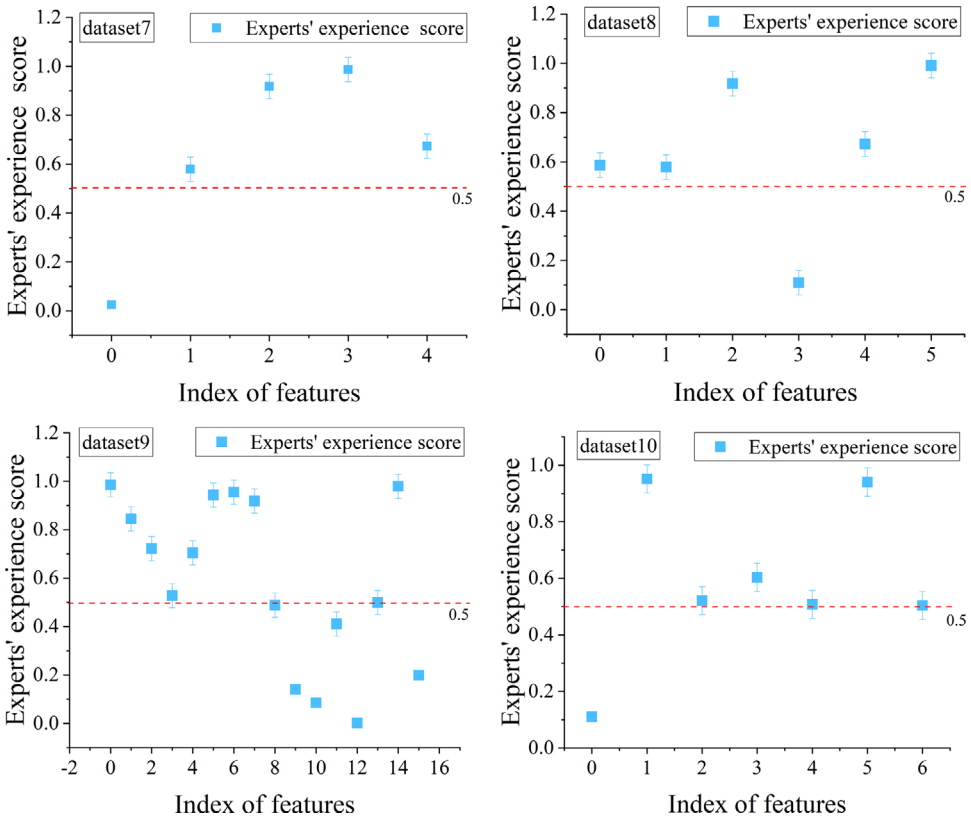


Figure 6. *Continued*

Table 4. Prediction accuracies of the models using the original feature set.

Dataset	Number of features	Prediction accuracy (fivefold cross validation)	
		RMSE	*MAPE*
1	6	0.1611	0.1310
2	6	0.0820	0.0695
3	5	0.0373	0.0265
4	47	0.0620	0.0510
5	18	0.1964	0.1482
6	27	0.1679	0.1446
7	5	0.1145	0.0809
8	6	0.1807	0.1458
9	16	0.1206	0.0956
10	7	0.1289	0.1239

Table 5. Number of features and prediction accuracies after sparsity evaluation.

	DML-FS				DML-FS$_{dek}$	
Dataset	Number of Input features	Number of remaining features	*RMSE*	*MAPE*	Number of remaining features	*RMSE*
1	6	6	0.1611	0.1310	6	0.1600
2	6	6	0.0819	0.0695	6	0.0819
3	5	5	0.0373	0.0265	5	0.0373
4	47	34	0.0611	0.0500	38	0.0581
5	18	17	0.1953	0.1461	18	0.1964
6	27	25	0.1672	0.1436	25	0.1672
7	5	3	0.1140	0.0805	4	0.1138
8	6	6	0.1807	0.1458	6	0.1807
9	16	16	0.1206	0.0956	16	0.1206
10	7	7	0.1289	0.1239	7	0.1289

performed, the size of the feature subset selected is different from that of DML-FS, which may imply that some of the features that domain experts consider to be crucial for system modeling are retained. For instance, for the dataset 4 the number of features selected using only DML-FS layer is 20, while the DML-FS with domain expert knowledge retained 30 features in which the six key features including A, X4, X15, X16, X17, and Y24 are considered important by domain experts. Moreover, we can observe that the *RMSE*s for the two selected feature sets (0.0513 and 0.0545) remain at the same level. For the dataset 5, the two features including X11 and X13 are retained due to high expert experience scores (see Figure 6, index = 10, index = 12) and the prediction accuracy is maintained at the same level (0.1789 and 0.1786). The features Ni and Al have a great influence on the lattice misfit of Ni-based single-crystal superalloy, and domain experts also give these two features high expert experience scores (see Figure 6, index = 0, index = 1). Thus, our approach ultimately

Table 6. Number of features and prediction accuracies after correlation evaluation.

	DML-FS				DML-FS$_{dek}$	
Dataset	Number of Input features	Number of remaining features	*RMSE*	*MAPE*	Number of remaining features	*RMSE*
1	6	4	0.1302	0.0467	5	0.1301
2	6	6	0.0819	0.0695	6	0.0819
3	5	4	0.0364	0.0263	4	0.0364
4	34	31	0.0565	0.0379	35	0.0573
5	17	15	0.1886	0.1459	17	0.1880
6	25	25	0.1672	0.1436	25	0.1672
7	3	2	0.1137	0.0776	3	0.1136
8	6	6	0.1807	0.1458	6	0.1807
9	16	15	0.1152	0.1113	15	0.1152
10	7	7	0.1289	0.1239	7	0.1289

Table 7. Number of features and prediction accuracy after redundancy evaluation.

	DML-FS				DML-FS$_{dek}$	
Dataset	Number of Input features	Number of remaining features	*RMSE*	*MAPE*	Number of remaining features	*RMSE*
1	4	4	0.1302	0.0467	5	0.1301
2	6	6	0.0819	0.0695	6	0.0819
3	4	4	0.0364	0.0263	4	0.0364
4	31	20	0.0513	0.0359	30	0.0545
5	15	11	0.1789	0.1443	15	0.1786
6	25	24	0.1663	0.1414	24	0.1663
7	2	2	0.1137	0.0776	3	0.1136
8	6	5	0.1428	0.1076	5	0.1428
9	15	11	0.1134	0.0897	13	0.1104
10	7	4	0.1235	0.1146	5	0.1214

retains these two important features, which is consistent with accepted knowledge of the physical and chemical domains.

In order to verify the superiority of our model in predicting performance and interpretability of materials, we compared our method with two existing sparsity methods (Lasso, Elastic net) on ten groups of materials properties datasets collected. The experimental results are shown in **Table 8**. For these ten datasets, our method is lower than Lasso and Elastic net in terms of *RMSE*, which shows a better prediction performance. The number of features selected by our method is more than that of the sparsity methods. This is because the introduction of domain expert knowledge allows the domain experts to consider important features to be retained, which is consistent with the accepted domain knowledge of materials physical-chemistry. Especially, for dataset 2, Lasso and Elastic net selected only one feature, while the DML-FS$_{dek}$ selected six features due to the combination of domain expertise, which not only improved the material interpretability but also improved the predictive accuracy by 40% (*RMSE* is decreased

Table 8. Comparison between Lasso and Elastic Net methods and ours on ten groups of materials properties dataset.

	Feature selection models						
		Lasso		Elastic net		Our method	
Dataset	Initial	Selected	*RMSE*	Selected	*RMSE*	Selected	*RMSE*
1	6	2	0.1448	3	0.1373	**5**	0.1301
2	6	1	0.2166	1	0.2166	**6**	**0.1301**
3	5	4	0.0422	5	0.0383	4	0.0364
4	47	36	0.0640	44	0.0624	30	0.0545
5	18	9	0.1941	9	0.1826	**15**	**0.1786**
6	27	18	0.1678	20	0.1688	**24**	0.1663
7	5	2	0.1437	2	0.1437	**3**	**0.1136**
8	6	4	0.1546	3	0.1657	**5**	0.1428
9	16	7	0.1242	8	0.1315	**13**	0.1104
10	7	4	0.2511	4	0.2475	**5**	**0.1214**

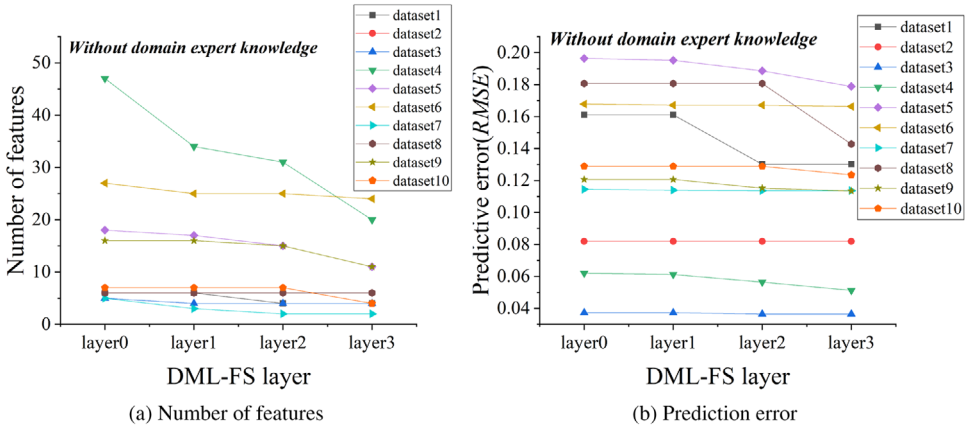
***Initial**: Number of initial features; ***Selected**: Number of final selected features

Table 9. Highly correlated pairs of features on dataset 6.

	Features	Highly correlated features	Correlation coefficient
1	C	B	0.9353
2	B	a_2T	0.8692
3	Ni	Co	−0.8487
4	B	Nb	0.8431
5	Co	a_2T	−0.8073

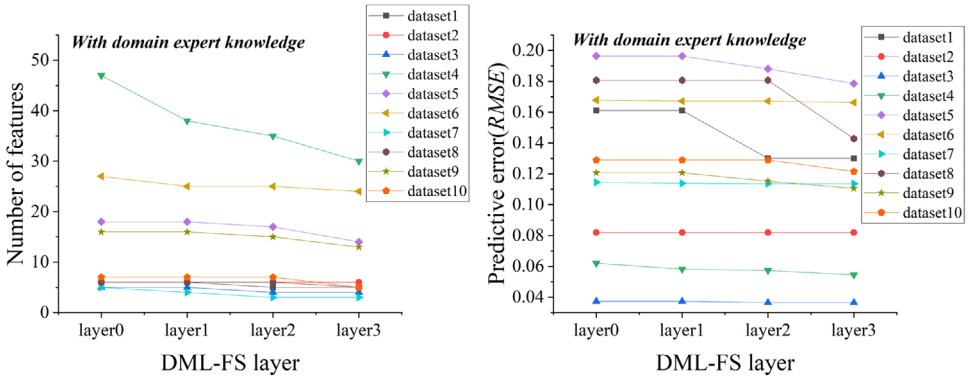
by about 40%). Similarly, for dataset 10, Lasso and Elastic net selected four features, while our method selected five features due to the assessment of the expert experience scores, resulting in a 52% improvement in prediction accuracy (*RMSE* is decreased by approximately 52%). Except for dataset 3 and 4, our method selected more features in the other eight datasets than the two sparse methods (the features that domain experts considered important were retained), and the predictive performance of the model was better. In general, compared with the two sparsity methods, the proposed method can improve the predicted performance while ensuring that the selected features are coincided with domain expert knowledge (i.e., the materials physics and chemistry information can be interpreted).

Next, taking the dataset 6 as an example, sparsity, correlation, and redundancy evaluation were conducted for all its features. The sparsity of the features is shown in **Figure 10**. We can observe that the variances of the two features Y (mass fraction of Y element) and a_2time (The second stage aging treatment time) are 0 and 0.0174, respectively, lower than the updated filtering threshold of 0.02. Combining the experts' experience score in Figure 6 with the proposed collaboration strategy in Section 3.3, we have calculated that the FCIS*s* for these two features are less than 0.5. Thus, the two features were discarded in this procedure. Then, correlation evaluation was further conducted for the remaining features and the result is shown in **Figure 11**. We can observe that the correlations between all the retained features and target



(a) Number of features (b) Prediction error

Figure 7. Size and prediction accuracy of the subset of output features in each layer of DML-FS on ten datasets (DML-FS without domain expert knowledge). a) Number of features. b) Prediction error.



(a) Number of features (b) Prediction error

Figure 8. Size and prediction accuracy of the subset of output features in each layer of DML-FS on ten datasets (DML-FS with domain expert knowledge).

attribute are all strong, exceeding the updated correlation filtering threshold of 0.3. Therefore, in this layer, no features were removed. Finally, we conducted redundancy analysis on the remaining features from the previous layer. **Figure 12** presents the correlation relationship between features and top 5 highly correlated pairs of features are listed in **Table 9**. It can be observed that high correlation or interdependence between features C and B, B and a_2T, Ni and Co, with their correlation coefficients of 0.9353, 0.8692, −0.8487, respectively. Further, the correlation coefficient between the C and the B is greater than the updated redundancy filtering threshold of 0.88, so feature C is retained, but the feature B is discarded (its FCIS is also below 0.5).

Finally, we accumulated statistics on the prediction accuracy for the above ten testing models, as shown in **Figure 7**b. It can be clearly observed that after passing through the three layers of DML-FS, the prediction error has decreased, particularly for the datasets 1 and 5. Moreover, Figure 7a shows the size of the output feature subset from each DML-FS layer. It can be observed that the 17 features were filtered out for the dataset 4. Analysis of the removed features revealed that there is a large amount of sparse, irrelevant and redundant information in the dataset. In contrast, for the datasets 1, 2, 3, and 8, their initial sets features were already pre-processed by domain experts and thus they had little sparsity or irrelevant or redundant information. Similarly, **Figure 8** presents the number of selected features and the prediction accuracy after the introduction of domain expert knowledge into each DML-FS layer. After the integration of experts' experience the risk of important features being removed is mitigated (the number of selected features by each layer is different from DML-FS layers without experts' input), and the prediction performance is equal to or better than that of pure DML-FS layer. From the results on consumed computational time for each DML-FS layer, shown in **Figure 9**, we can observe that it mainly depends on the size of the data set and none of the layers has a dominant effect on the process performance.

5. Conclusion

A novel data-driven multi-layer feature selection mechanism integrating domain expert knowledge is proposed and tested. The proposed method can eliminate sparse, irrelevant, and redundant information from the original feature set using three layers of sparsity evaluation, correlation evaluation, and redundancy evaluation. The whole process is automatic and does not require the user to have professional knowledge about feature selection.

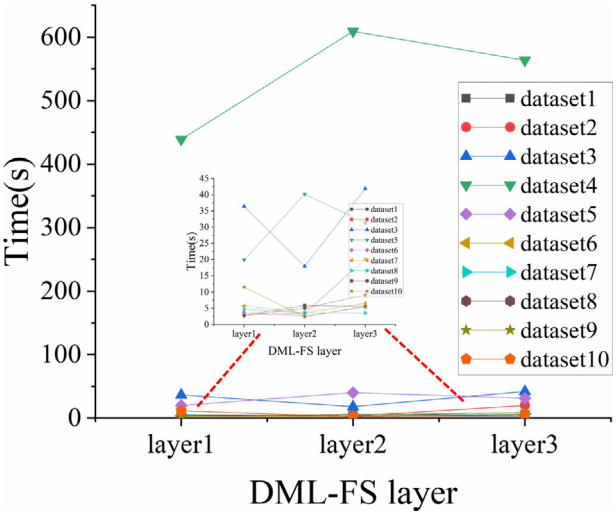


Figure 9. Computation time for each DML-FS layer.

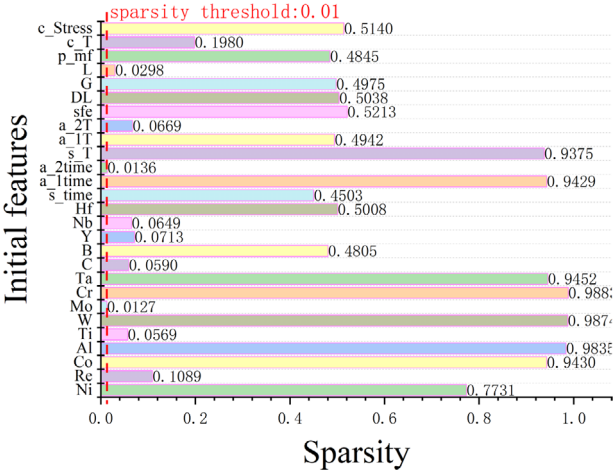


Figure 10. Results of sparsity evaluation on dataset 6. The X-coordinate indicates the material feature, and the Y-coordinate indicates the sparsity value of the feature.

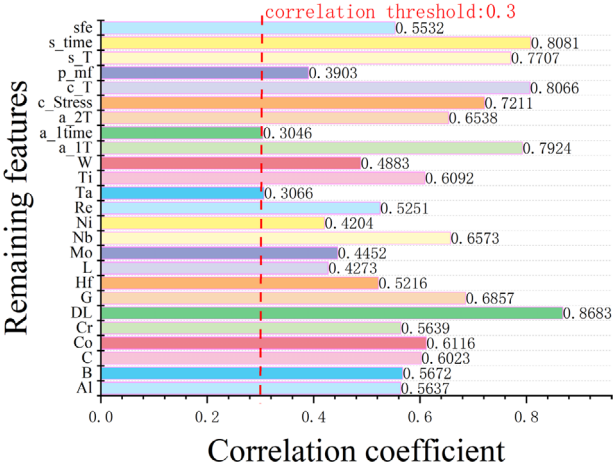


Figure 11. Results of correlation evaluation on dataset 6. The X-coordinate indicates the material feature, and the Y-coordinate indicates the correlation between the features and the material properties.

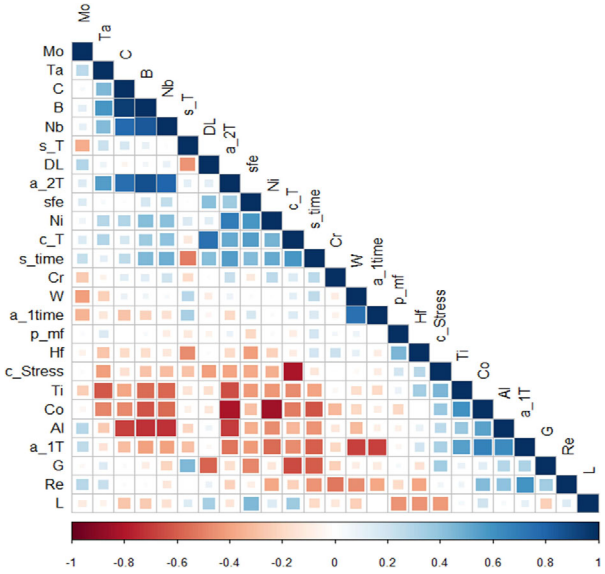


Figure 12. Results of redundancy evaluation for the dataset 6, and the heat map of Pearson correlation coefficient matrix among the remaining features. The red and blue colors indicate positive and negative correlation, respectively. The darker the color is, the stronger the correlation is.

Moreover, we present a method to quantify and integrate domain expert knowledge into the feature selection process. First, the domain expert knowledge of features is quantified as the weights (importance scores) of features and expert experience score of each feature is evaluated by means of weighted average. Second, the feature subset is optimized based on the collaborative strategy between the expert experience score and the DML-FS layers, which reduces the risk of removing features that domain experts consider important. The proposed method was tested on ten groups of material properties datasets. The results show that the mechanism can effectively select the optimal and interpretable feature subset while ensuring the prediction accuracy remains unchanged or even improves.

With the extensive use of DML-FS$_{dek}$ in the field of materials science, additional expert knowledge, that is, the importance score of different experts for features, will be stored and recorded internally to further improve predictive performance of the models. Overall, the DML-FS$_{dek}$ implementation presented herein is expected to enable the feature and correlation analysis of large-scale material properties datasets which was unattainable so far. Notably, as the data volume increases and the expert knowledge obtained is more abundant, there may be uncertainty when the model is used for future predictions, that is, the prediction results may change accordingly. Additionally, when novices use our automatic modeling method, they may produce models that look good with serious flaws and little utility. Therefore, we urgently need to bring in the domain knowledge of experts to further regulate the model so that the model can meet the needs of materials modeling in accuracy and reliability.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (Grant Nos. 2017YFB0701500 and 2017YFB0701600). All the computations were performed on the High Performance Computing Center, Shanghai University.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Y.L., M.A., and S.S. conceived and designed the theoretical calculations and experiments. Y.L., J.W., and S.S. carried out all the theoretical calculations and analyzed the data. Y.L. and J.W. performed and analyzed the experiments. All authors discussed the results and prepared and revised the manuscript.

Keywords

domain expert knowledge, feature selection, machine learning, materials properties prediction

Received: October 31, 2019
Revised: December 17, 2019
Published online: January 15, 2020

[1] D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, P. W. Chung, *Sci. Rep.* **2018**, *8*, 9059.
[2] Y. Liu, T. Zhao, G. Yang, J. Wu, S. Shi, *Comput. Mater. Sci.* **2017**, *140*, 315.
[3] S. F. Fang, M. P. Wang, M. Song, *Mater. Des.* **2009**, *30*, 2460.
[4] J. M. Granda, L. Donina, V. Dragone, D. L. Long, L. Cronin, *Nature* **2018**, *559*, 377.
[5] J. C. Snyder, M. Rupp, K. Hansen, K. R. Müller, K. Burke, *Phys. Rev. Lett.* **2012**, *108*, 253002.
[6] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, *Phys. Rev. B* **2017**, *95*, 144110.
[7] S. S. Young, F. Yuan, M. Zhu, *Mol. Inf.* **2012**, *31*, 707.
[8] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, *Nat. Commun.* **2018**, *9*, 3405.
[9] J. P. Janet, H. J. Kulik, *J. Phys. Chem. A* **2017**, *121*, 8939.
[10] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, A. J. Fisher Craig, H. Moriwake, I. Tanaka, *Adv. Energy Mater.* **2013**, *3*, 980.
[11] L. Chonghe, T. Yihao, Z. Yingzhi, W. Chunmei, W. Ping, *J. Phys. Chem. Solids* **2003**, *64*, 2147.
[12] M. Shahlaei, *Chem. Rev.* **2013**, *113*, 8093.
[13] J. M. Sutter, S. L. Dixon, P. C. Jurs, J. Chem, *Inf. Comput. Sci.* **1995**, *35*, 77.
[14] A. Z. Dudek, T. Arodz, J. Gálvez, *Comb. Chem. High Throughput Screen.* **2006**, *9*, 213.
[15] U. Özge, I. B. Türkşen, *Inf. Sci. (N. Y.)* **2007**, *177*, 449.
[16] W. Li, R. Jacobs, D. Morgan, *Comput. Mater. Sci.* **2018**, *150*, 454.
[17] Z. Qi, N. Zhang, Y. Liu, W. Chen, *Compos. Struct.* **2019**, *212*, 199.
[18] W. Zhang, L. Yu, T. Yoshida, Q. Wang, *Knowledge Inform. Syst.* **2019**, *58*, 371.
[19] G. Chandrashekar, F. Sahin, *Comput. Electr. Eng.* **2014**, *40*, 16.
[20] M. A. Hall, *Proc. Seventeenth Int. Conf. on Machine Learning*, Morgan Kaufmann Publishers Inc. **2000**, pp. 359.
[21] H. Peng, F. Long, C. Ding, *IEEE Trans. Pattern Anal. Machine Intellig.* **2005**, *27*, 1226.
[22] S. W. Card, *Conf. Companion Genetic Evolution. Comput. ACM* **2010**, *1851*, 1854.
[23] R. Caruana, V. R. Sa, *J. Mach. Learn. Res.* **2003**, *3*, 1245.
[24] P. Langley, *Proc. AAAI Fall Symp. Relevance* **1994**, *184*, 245.
[25] F. Lin, D. Liang, C. C. Yeh, J. C. Huang, *Expert Syst. Appl.* **2014**, *41*, 2472.
[26] E.T. Tekin, C. Tas, M. Cebi, *Comput. Biol. Med.* **2015**, *64*, 127.
[27] R. Genuer, J. M. Poggi, C. Tuleau-Malot, *Pattern Recog. Lett.* **2010**, *31*, 2225.
[28] V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo, M. P. Mendes, *Sci. Total Environ.* **2018**, *624*, 661.
[29] M. Salahinejad, T. C. Le, D. A. Winkler, *J. Chem. Inf. Model.* **2013**, *53*, 223.
[30] D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, N. Ferrando, B. Creton, *Energy Fuels* **2012**, *26*, 2416.
[31] Creep data, https://github.com/dlcj/creepdata (accessed: December 2018).
[32] M. R. Johan, S. Ibrahim, *Commun. Nonlinear Sci. Num. Simul.* **2012**, *17*, 329.
[33] L. Xu, L. Wencong, P. Chunrong, S. Qiang, G. Jin, *Comput. Mater. Sci.* **2009**, *46*, 860.
[34] X. Jiang, H. Q. Yin, C. Zhang, R. J. Zhang, K. Q. Zhang, Z. H. Deng, G. Q. Liu, X. H. Qu, *Comput. Mater. Sci.* **2018**, *143*, 295.
[35] Y. Liu, J. Wu, G. Yang, T. Zhao, S. Shi, *Sci. Bull.* **2019**, *64*, 1195.
[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.

Multi-layer Feature Selection Incorporating Weighted Score-based Expert Knowledge toward Modelling Materials with Targeted Properties

Yue Liu[a], Junming Wu[a], Maxim Avdeev[b,c], Siqi Shi[d,e,*]

[a]*School of Computer Engineering and Science, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China*

[b]*Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC NSW 2232, Australia*

[c]*School of Chemistry, The University of Sydney, Sydney 2006, Australia*

[d]*School of Material Science and Engineering, Shanghai University, Shanghai 200444, China*

[e]*Materials Genome Institute, Shanghai University, Shanghai 200444, China*

[*]E-mail: sqshi@shu.edu.cn (Siqi Shi)

1. Specifications Table

Subject area	Feature selection, Material properties prediction, Machine learning
More specific subject area	Material properties datasets and source codes for selection of material features and training of machine-learning predictive model
Type of data	Spreadsheet, Source code, Word document
How data was acquired	Published literatures [1][2][3][4][6][7], Online resources [5]
Data format	Analyzed
Data accessibility	Data is with this article

2. Value of the data

- Provide detailed information on eight different material properties data sets. All datasets can be extended to other data mining and machine learning tasks in the future.
- The detailed information of feature selection and model training results is provided, which enables other researchers to quickly get familiar with the specific execution procedure of the algorithm, and use it to assess the importance of features and the inherent correlation among features.
- Provide with source code for training our proposed models. Users can provide their own material data sets for training and testing.

3. The details of datasets

The datasets used for feature selection and training of material properties prediction model includes 8 csv files with file names of data1, data2, data3, data4, data5, data6, data7, and data8, respectively. The details of each csv file are summarized as follows.

3.1 data1.csv

This dataset is about the prediction of the ionic conductivity of Lithium superionic conductors, in which the total number of samples is 128, the number of material features is 6, and the targeted property is the ionic conductivity of lithium battery. The detailed descriptions of material features and targeted property for dataset 1 are shown in **Table S1**.

Table S1. The detailed descriptions of material features and targeted property for dataset 1.

Material features	**Targeted property**
Symbols in this article	Symbols in this article
X1	Y
X2	
X3	
X4	
X5	
X6	

3.2 data2.csv

This dataset is about the prediction of the lattice constant, in which the total number of samples is 161, the number of material features is 5, and the targeted property is the lattice constant of organic compounds. The detailed descriptions of material features and targeted property for dataset 2 are shown in **Table S2**.

Table S2. The detailed information of material features and targeted property for dataset 2.

Material features	**Targeted property**
Symbols in this article	Symbols in this article
X1	Y
X2	
X3	
X4	
X5	
X6	

3.3 data3.csv

This dataset is about the prediction of the crystal enthalpy, in which the total number of samples is 1302, the number of material features is 5, and the targeted property is the crystal enthalpy of organic compounds. The detailed descriptions of material features and targeted property for dataset 3 are shown in **Table S3**.

Table S3. The detailed information of material features and targeted property for dataset 3.

Material features	**Targeted property**
Symbols in this article	Symbols in this article
X1	Y
X2	
X3	
X4	
X5	

3.4 data4.csv

This dataset is about the prediction of the density of organic materials, in which the total number of samples is 5619, the number of material features is 47, and the targeted property is the density of organic materials. The detailed descriptions of material features and targeted property for dataset 4 are shown in **Table S4**.

Table S4. The detailed information of material features and targeted property for dataset 4.

Material features				**Targeted property**
Symbols in this article				Symbols in this article
T(K)	A	B	C	Y
D	X2	X3	X4	
X5	X6	X7	X8	
X10	X12	X14	X15	

X16	X17	X19	X21
X22	Y1	Y2	Y3
Y4	Y5	Y6	Y7
Y8	Y9	Y10	Y11
Y12	Y13	Y14	Y15
Y16	Y17	Y18	Y19
Y20	Y21	Y22	Y23
Y24	Y25	Y26	

3.5 data5.csv

This dataset is about the prediction of the viscosity of organic material, in which the total number of samples is 669, the number of material features is 18, and the targeted property is the viscosity of organic material. The detailed descriptions of material features and targeted property for dataset 5 are shown in **Table S5**.

Table S5. The detailed information of material features and targeted property for dataset 5.

Material features		**Targeted property**
Symbols in this article		Symbols in this article
T (K)	X18	Viscosity
A	X20	
F	Y1	
G	Y2	
X1	Y5	
X9	Y14	
X10	Y15	
X11	Y17	
X13	Y23	

3.6 data6.csv

This dataset is about the prediction of the creep fracture life, in which the total number of samples is 77, the number of material features is 27, and the targeted property is the creep rupture life of Ni-based single crystal superalloy. The detailed descriptions of material features and targeted property for dataset 6 are shown in **Table S6**.

Table S6. The detailed information of material features and targeted property for dataset 6.

Material features			**Targeted property**
Symbols in this article			Symbols in this article
Ni	Re	Co	Creep rupture life
Al	Ti	W	
Mo	Cr	Ta	
C	B	Y	
Nb	Hf	s_time	
a_1time	a_2time	s_T	
a_1T	a_2T	sfe	

DL	G	L
p_mf	c_T	c_Stress

3.7 data7.csv

This dataset is about the prediction of the ionic conductivity of nanocomposite, in which the total number of samples is 92, the number of material features is 6, and the targeted property is the ionic conductivity of nanocomposite solid polymer electrolyte system. The detailed descriptions of material features and targeted property for dataset 7 are shown in **Table S7**.

Table S7. The detailed information of material features and targeted property for dataset 7.

Material features	**Targeted property**
Symbols in this article	Symbols in this article
PEO	Conductivity
LiPF6	
EC	
CNT	
Temp	

3.8 data8.csv

This dataset is about the prediction of the oxide ionic conductivities, in which the total number of samples is 117, the number of material features is 5, and the targeted property is the oxide ionic conductivities in ABO_3 perovskites. The detailed descriptions of material features and targeted property for dataset 8 are shown in **Table S8**.

Table S8. The detailed information of material features and targeted property for dataset 8.

Material features	**Targeted property**
Symbols in this article	Symbols in this article
X1	Lnr
X2	
X3	
X4	
X5	
X6	

3.9 data9.csv

This dataset is about the prediction of the lattice misfit, in which the total number of samples is 136, the number of material features is 16, and the targeted property is the lattice misfit of Ni-based single crystal superalloy. The detailed descriptions of material features and targeted property for dataset 9 are shown in **Table S9**.

Table S9. The detailed information of material features and targeted property for dataset 9.

Material features	**Targeted property**
Symbols in this article	Symbols in this article
Ni	The lattice misfit of Ni-based single crystal superalloy
Al	

Co	
Cr	
Mo	
Re	
Ru	
Ta	
Nb	
Ti	
W	
Hf	
Ir	
Temperature	
Dendrite Location	
Specimen Thickness	

3.10 data10.csv

This dataset is about the prediction of the onset temperature (T_g), in which the total number of samples is 9, the number of material features is 7, and the targeted property is the onset temperature (T_g) of Ge_xSe_{1-x} glass transition. The detailed descriptions of material features and targeted property for dataset 10 are shown in **Table S10**.

Table S10. The detailed information of material features and targeted property for dataset 10.

Material features	**Targeted property**
Symbols in this article	Symbols in this article
$\boldsymbol{x}$	T_g
$<r>$	
v	
K	
V_0	
U_{0ex}	
b	

4. Code repository

This repository contains source code for training/testing machine learning models on the provided dataset. This package can provide predictions of material properties using the machine learning approaches developed in this work. The code repository is available online at https://github.com/wujunming1/material-attribute-datasets.

5. Reference

[1] ujimura K, Seko A, Koyama Y, et al. Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms [J]. Advanced Energy Materials, 2013, 3(8): 980-985.

[2] Chonghe L, Yihao T, Yingzhi Z, et al. Prediction of lattice constant in perovskites of GdFeO

3 structure [J]. Journal of Physics and Chemistry of Solids, 2003, 64(11): 2147-2156.

[3] Salahinejad M, Le T C, Winkler D A. Capturing the crystal: prediction of enthalpy of sublimation, crystal lattice energy, and melting points of organic compounds [J]. Journal of chemical information and modeling, 2013, 53(1): 223-229.

[4] Saldana D A, Starck L, Mougin P, et al. Prediction of density and viscosity of biofuel compounds using machine learning methods [J]. Energy & Fuels, 2012, 26(4): 2416-2426.

[5] https://github.com/dlcj/creepdata.

[6] Johan M R, Ibrahim S. Optimization of neural network for ionic conductivity of nanocomposite solid polymer electrolyte system (PEO–LiPF6–EC–CNT)[J]. Communications in Nonlinear Science and Numerical Simulation, 2012, 17(1): 329-340.

[7] Xu L, Wencong L, Chunrong P, et al. Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO3 perovskites[J]. Computational Materials Science, 2009, 46(4): 860-868.