Descriptors Divide-and-Conquer Enables Multifaceted and Interpretable Materials Structure–Activity Relationship Analysis

Yue Liu, Linhan Wu, Zhengwei Yang, Xinxin Zou, Zheyi Zou, Yuxiao Lin, Maxim Avdeev, and Siqi Shi*

Machine learning (ML) is increasingly adopted to explore the dependence of properties on descriptors especially for materials with the complicated structure–activity relationships. However, most current ML modeling strategies typically depend on a single combination of descriptors, which leads to inaccurate and unilateral inferences. Here, a descriptors divide-and-conquer method is proposed for machine learning (descriptors-DCML) in which rough set theory (RST) is integrated with materials domain knowledge to select multiple optimal sets of descriptors combinations and thus diverse rule extraction strategies are provided to dig out mechanisms latent in materials data. Its potential utility and applications using the sodium ion energy barrier prediction of NASICION-type solid-state electrolyte compounds with multifaceted influencing factors as an example are demonstrated. A total of 85 NASICION-type samples with 45 descriptors derived from 72 published literature serve as the data foundation for ML modeling. Not only does descriptors-DCML exhibit the energy barrier prediction accuracy of 93.8% but also extract 9 relations mapping essential factors to Na ion energy barrier in which 5 ones conform to existing understanding rule and the rest are waiting for validation. This work paves the way for reducing the complexity of analyzing materials structure–activity relationships and enhancing the interpretability of ML models.

1. Introduction

Data-driven machine learning (ML) is widely employed in materials discovery and design due to its superior ability for the rapid recognition of data patterns and the accurate prediction of materials properties.[1–6] However, pattern detection in ML is subject to prior assumptions regarding the data. Consequently, materials properties and the corresponding influencing mechanisms, which are mostly in close relation to multiple physical/chemical factors,[7–10] may not be effectively and accurately predicted only by an individual ML model or analysis method. The as-established univariate structure–activity relationships are often insufficient for the discovery and design of novel materials. Thus, extensive and interpretable analysis of complex performance-driven mechanisms is in urgent need, particularly for special functional materials.

To characterize various materials properties, multiple physical/chemical descriptors were previously defined by materials

Y. Liu, L. Wu, Z. Yang, X. Zou
School of Computer Engineering and Science
Shanghai University
Shanghai 200444, China
Z. Zou
School of Materials Science and Engineering
Xiangtan University
Xiangtan 411105, China
Y. Lin
School of Physics and Electronic Engineering
Jiangsu Normal University
Xuzhou 221116, China
M. Avdeev
Australian Nuclear Science and Technology Organisation
Sydney 2232, Australia

M. Avdeev
School of Chemistry
The University of Sydney
Sydney 2006, Australia
S. Shi
State Key Laboratory of Materials for Advanced Nuclear Energy & School of Materials Science and Engineering
Shanghai University
Shanghai 200444, China
E-mail: sqshi@shu.edu.cn
S. Shi
Materials Genome Institute
Shanghai University
Shanghai 200444, China

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
FUNCTIONAL
MATERIALS**

www.afm-journal.de

researchers, resulting in a large descriptors space. Traditionally, the descriptors selection for ML model construction is done via the trial-and-error and only a single combination of descriptors related to the target property is focused on, which is only suitable for simple structure–property relationships. In addition, methods such as principal component analysis (PCA)[11] or Lasso regression (LR)[12] are employed to reduce descriptors dimensionality or select descriptors that significantly impact the target properties. However, these methods are purely data-driven and also have limitations in selecting only one optimal descriptors combination. When structure–activity relationships become complex and diverse, they often fall short and a comprehensive and accurate analysis between different descriptors combinations and target property are required. In such a sense, the traditional methods not only limit the scope of structure–activity relationships that ML models can learn, but also waste a valuable opportunity to understand the underlying performance mechanisms.

Rough set theory (RST), mainly used for learning and induction of incomplete data and uncertain knowledge,[13,14] is capable of exploring the relations between factors and indicators via constructing and reducing decision attribute, then obtaining decision rules between the conditional attribute and decision attribute.[15] RST can directly extract relationships and uncover patterns between features within the data itself, without entirely relying on prior labels or target variables to guide the feature extraction process. In contrast, methods like PCA and LR often require prior assumptions and are limited by their linearity assumptions or sensitivity to multicollinearity, which can lead to suboptimal feature selection in complex and non-linear relationships. Furthermore, RST exhibits notable fault tolerance and robustness during attribute reduction, enabling it to effectively identify redundant or irrelevant features, even in the presence of noise and incomplete data. These characteristics make RST well-suited for descriptors-oriented approach, especially when extracting the key influencing descriptors of materials properties. In addition, RST confirmed that there might be multiple optimal subsets with approximate optimization performance in attribute reduction. The resulted multiple descriptors combinations align well with common knowledge in materials science that the same material property may be influenced by various factors or mechanisms. The flexibility and reliability of RST in handling complex data make it appropriate for exploring the structure–activity relationships of materials.

Aiming to expand common single-descriptor-combination based approach in ML model construction, we propose a descriptors divide-and-conquer method for machine learning (descriptors-DCML). Our approach selects multiple optimal sets of descriptors combinations instead of relying on a single descriptors combination to train materials property prediction models. Concretely, core features and feature cluster are defined for ML models. Furthermore, RST and materials domain knowledge are introduced to distinguish multiple influencing mechanisms latent in materials data by recognizing core features and feature clusters. Based on feature clusters, high-precision ML models are constructed to establish relationships between features combinations and materials properties. Finally, interpretable schemes are matched for ML models to provide a multifaceted analysis of materials structure–activity relationships. The efficacy of descriptors-DCML is demonstrated on six materials

datasets across various material systems. Descriptors-DCML can extract multifaceted and interpretable structure–activity relationships related to the target properties while ensuring the prediction accuracy of ML models. The theory and techniques lay a foundation for unveiling structure–activity relationships of materials and can also be expanded to other materials researches.

2. Method

As shown in **Figure 1**, descriptors-DCML employs NCOR-PSO, a feature selection method proposed in our previous work,[7] to construct the feature decision table between different feature combinations and their predictive performance on ML models. In the following process of core features recognition, key-descriptors-related materials domain knowledge is combined with the core computation algorithm of rough set to recognize a set of key descriptors that affect the predictive performance of ML models. After fixing core features, feature clusters are obtained by NCOR-PSO to construct multiple high-precision ML models. Finally, rules are extracted from the learning results of the constructed ML models and thus materials domain knowledge related to influencing mechanisms is integrated to provide a multifaceted and interpretable analysis for the complex structure–activity relationships.

2.1. The Definition of Core Features and Feature Cluster

Because diverse physical/chemical factors impact multiple mechanisms toward specific materials properties,[16] descriptors-DCML distinguishes influencing mechanisms or factors latent in materials data via identification of core features and feature cluster. The definition of core features and feature cluster can be seen in Definitions 1 and 2.

Definition 1. *Core features: the features or minimum features combination in the candidate feature set that has the strongest correlation with the predictive performance of ML models.*

Definition 2. *Feature cluster: the features combination including core features that can be used to construct ML models with higher predictive performance.*

In addition, considering that the predictive performance of ML models is not only affected by the distribution and characteristics of materials data, but also constrained by materials domain knowledge, code features can be further divided into domain-core features and data-core features, which can be seen in Definitions 3 and 4. Data-core features reflect the key information at the data level via reduction of the feature decision table. And domain-core features are based on expert experience in specific domains, extracting features closely related to the target property.

Definition 3. *Domain-core features: the core features obtained from key-descriptors-related materials domain knowledge.*

Definition 4. *Data-core features: the core features derived through the reduction of the feature decision table.*

Note that different ML models may own different core features and feature clusters for the same materials property prediction task. To this end, descriptors-DCML can distinguish and
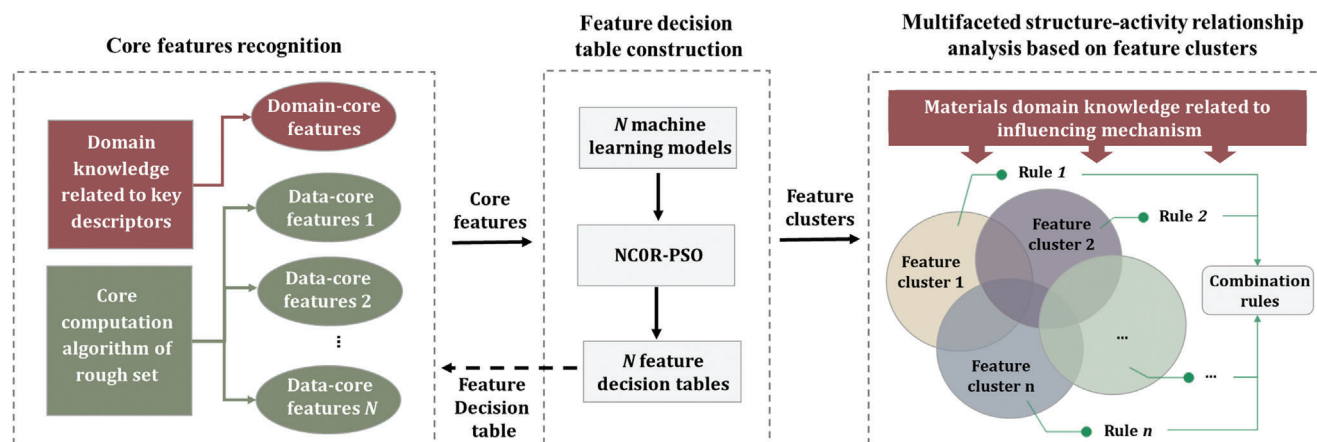
Figure 1. Overall schematic of descriptors-DCML. In the stage of core feature recognition, RST is employed to resolve features into data-core features and domain-core features are collected by material experts. Then, the stage of feature decision table construction is designed to analyze the insight of features combinations, thus constructing feature decision tables for previous stage. Finally, multifaceted structure–activity relationships analysis based on feature clusters are employed to extract rules.

learn diverse structure–activity relationships in the same materials dataset by selecting different feature clusters for different ML models. Core features can ensure that high-precision ML models can be constructed based on feature clusters, thus ensuring the accuracy of the learning results of ML models to a certain extent.

2.2. The Acquisition Approach of Core Features

In general, material experts possess the domain knowledge related to influencing mechanisms of materials properties when constructing datasets for ML modeling, particularly some key descriptors that can determine the performance of the target materials property to a greater extent compared to other descriptors. The importance of these key descriptors has been validated in previous studies, e.g., the high correlation between bottleneck size and the ion transport performance of solid-state electrolytes.[17–19] Hence, domain-core features collected by materials domain knowledge can to some extent ensure that ML models do not exhibit deficient predictive performance due to the lack of key descriptors. Note that key descriptors given by material experts may involve multiple candidate descriptors representing the same physical concept. However, the high correlation between descriptors has negative effects on the predictive performance of ML models. Therefore, it is necessary to divide domain-core features into multiple independent sub-domain-core features. Each sub-domain-core feature may contain one or multiple descriptors that represent the same physical concept. Finally, ML models are constructed based on each sub-domain-core feature, and the key-domain-core feature is the sub-domain-core feature which has the optimal predictive performance.

RST[15] is proposed to dispose the inaccurate and uncertain knowledge. The decision table is an important knowledge representation system[20,21] in RST and most decision problems can be expressed by its form. Note that NCOR-PSO evaluates the quality of features combinations via their predictive performance on ML models. Multiple executions of NCOR-PSO can provide diverse features combinations and their predictive performance of

ML models due to the uncertainty of particle swarm optimization (PSO)[22] algorithm. To this end, based on the expression form of the decision table, we further propose the feature decision table to represent the impact of diverse features combinations on the predictive performance of ML models and its definition is as follows:

Definition 5. *Feature decision table: Given a decision table $S = \,<U, R, V, f>$, where U represents a set of entire feature sets; $R = C \cup D$, $C \cap D = \varnothing$, C means the set of candidate features, D is the set of performance evaluation metrics of ML models; $V = \{V_c, V_D\}$ and $V_c = \{0, 1\}$ means the value set of C, 1 and 0 represent whether each feature set x in U contains a certain feature of C, V_D represents the value set of D; f: $U \times R \to V$ represents an information function that specifies the value of C and D on each object x (feature subset) in U; S is regard as a feature decision table.*

Attribute reduction is a common operation on the decision table in RST to remove unimportant or redundant conditional attributes while maintaining the classification ability of the decision table. The set of remaining conditional attributes is referred to as the reduction of the decision table. Note that there are more than one reduction of a decision table and the intersection of all reductions is called core. Feature decision table is a special type of decision table, so it is accessible to execute attribute reduction and compute core on it. Note that the feature sets obtained by NCOR-PSO may contain redundant features that have little impact on the predictive performance of ML models but contribute to a large diversity of the feature decision table, thus making it tough to find core features. Hence, we propose core features seeking method based on the simplest feature decision table (CFS-SFDT), of which details can be seen in Algorithm 1.

Algorithm 1 consists of three main parts: Initialization (line 1–2), updating feature decision table (line 3–15), and finding positive-data-core features (line 16–19). We observe that the steps of Initialization (lines 1–2) are not at all time-consuming. It can be carried out in $O(1)$. The steps of updating feature decision table (line 3–15) are decided by the time complexity of calculate the dependence which is $O(Klmv)$. The steps of finding

Algorithm 1 Core features seeking method based on the simplest feature decision table

Input: $S = <U, R, V, f>$, $R = C \cup D$, $C = \{C_i | i = 1, 2, \cdots, m\}$, $x \subset U$, $V_x^{C_i} \in x$, $|D| = 1$, $V_D = \{V_D^i | i = 1, 2, \cdots, l\}$, $K = \{k_i | i = 1, 2, \cdots, t\}$, $k_i < k_{i+1}$.
Output: The Set of Core Features $NF \subseteq C$.

1:     $X = \{X_i | i = 1, 2, \cdots, l\}$, where X_i is the set of feature subsets x with the same value on D;
2:     $P = \{P_i | i = 1, 2, \cdots, l\}$, where $P_i = \{p_i^j | j = 1, 2, \cdots, m\}$ is the occurrence frequency of each candidate feature in the X_i;
3:     for k_i in K:
4:        Calculate the dependence $\gamma_C(D)$ of D on C based on S;
5:        for i in $\{1, 2, \cdots, l\}$:
6:           for j in $\{1, 2, \cdots, m\}$:
7:              for x in X_i:
8:                 if $V_x^{C_j} < p_i^j$ and $V_x^{C_j} = 1$ then
9:                    $V_x^{C_j} = 0$, $k_{best} = k_{i-1}$
10:       Obtain the updated feature decision table S';
11:       Calculate the dependence $\gamma_C(D)$ of D on C based on S;
12:       if $\gamma_C'(D) < \gamma_C(D)$ then
13:          Exit
14:       Else
15:          Set $S = S'$
16:       Find $POS_C(D)$ based on S;
17:       Let $NF = \varnothing$;
18:       for C_i in C:
19:          if $POS_{C-\{C_i\}}(D) \neq POS_C(D)$ then $NF = NF \cup \{C_i\}$
20:       Output NF and k_{best}

positive-data-core features decided by the time complexity of find $POS_C(D)POS_C(D)$ and find core feature, which is $O(q)$. Thus, the time complexity is $O(Klmv)$.

As shown in Algorithm 1, CFS-SFDT contains two parts: 1) reduce the original feature decision table to obtain the simplest feature decision table; 2) compute core based on the simplest feature decision table. Concretely, CFS-SFDT divides the domain U of the feature decision table S into l groups according to the values of D, and calculates the occurrence frequency of each candidate feature in each group. Then, the candidate thresholds k_i are iteratively extracted from the set of candidate thresholds K, and the feature decision table S is reduced according to k_i to obtain a new feature decision table S'. The value C_j in the subset x is set as 0 according to the value C_j of each group being 1 as well as its occurrence frequency being lower than k_i, to reduce the diversity of the feature decision table. Specially, to ensure the compatibility of S' and S, it is necessary to calculate the dependability $\gamma_C(D)$ and $\gamma_C'(D)$ of C on D in S' and S, respectively. If $\gamma_C'(D) < \gamma_C(D)$, it means that the reduction of feature decision table in this iteration reduced the classification ability of S, then, the iteration stops at this point and the simplest feature decision table and the best threshold k_{best} are S' and k_{i-1} of the previous iteration, respectively. The calculation formula of dependability $\gamma_C(D)$ is as follows:

$$\gamma_C(Q) = \frac{card\left(POS_C(D)\right)}{card(U)} \quad (1)$$

$$POS_c(D) = \cup_{X\varepsilon\frac{U}{D}}\underline{R}(X) \quad (2)$$

$$\underline{R}(X) = \cup\left\{Y\varepsilon\frac{U}{c} : Y\varepsilon X\right\} \quad (3)$$

where $card(\cdot)$ is employed to calculate the number of objects in certain set; $\frac{U}{D}$ and $\frac{U}{C}$ represent the obtained equivalence classes based on the value of C and D, and each equivalence class contains an indistinguishable number of objects x; $\underline{R}(X)$ represents X-positive domain. According to the knowledge C, the set of entire elements in U can be classified as X definitely; $POS_C(D)$ is the C-positive domain of D, which is the set of objects in $\frac{U}{D}$ that can be correctly classified using the knowledge expressed by the classification $\frac{U}{C}$. The dependability $\gamma_C(D) \in [0, 1]$, $\gamma_C(D) = 1$ indicates that the classification result of $\frac{U}{D}$ depends entirely on C; $\gamma_C(D) = 0$ indicates that the classification result of $\frac{U}{D}$ does not depend entirely on C.
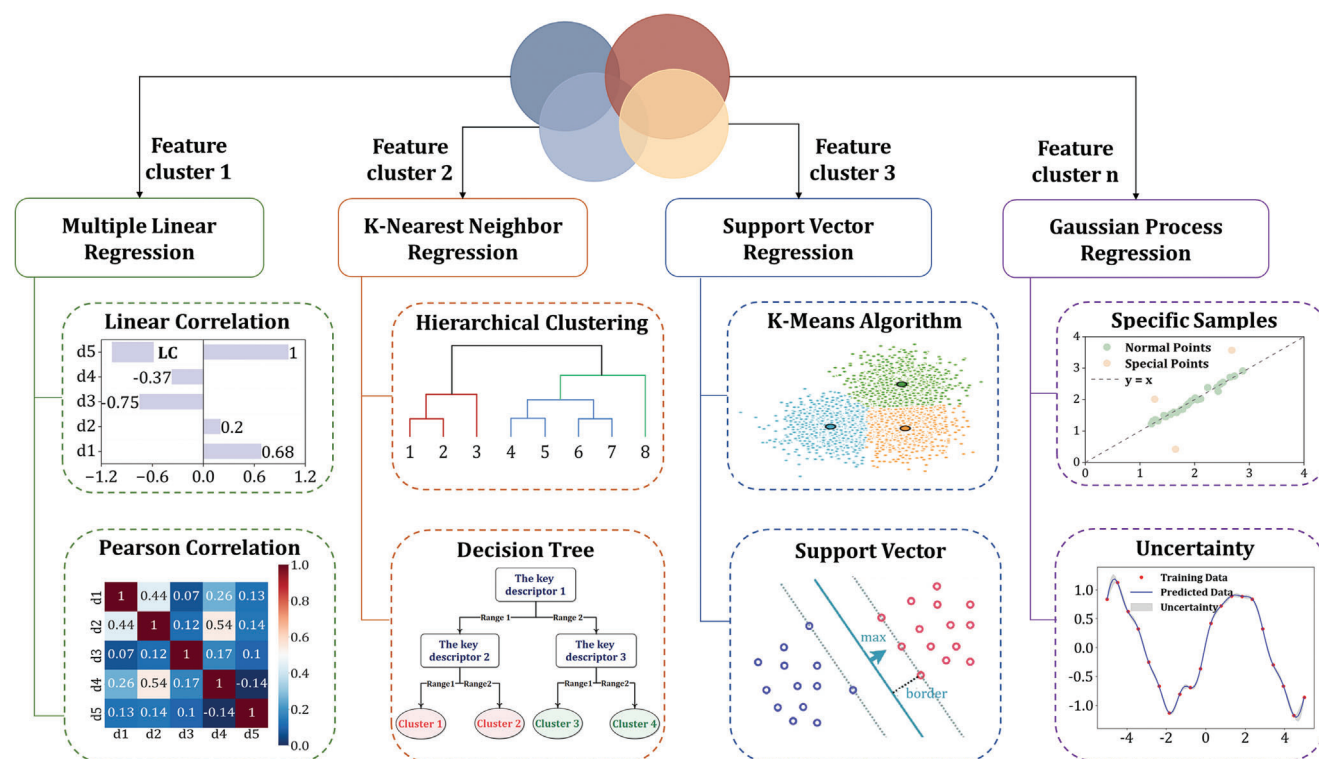
The direct core computation algorithm based on positive region[20] is employed to compute the core of the simplest feature decision table. However, data-core features obtained by CFS-SFDT can only indicate that they can determine the predictive performance of ML models to a certain extent, but cannot conclude whether they should be retained or not. Therefore, we further propose positive-data-core features, which are the data-core features that have a positive impact on the predictive performance of ML models. Positive-data-core features can be obtained by selecting data-core feature whose frequency of appearance in the optimal group of the feature decision table is greater than a given threshold. Note that both positive-data-core features and domain-core features can ensure the predictive performance of ML models, thus combined-core features can be constructed to further enhance the core features while maintaining low correlation between descriptors. Combined-core features are the combination of key-domain-core features and non-domain-core features, where non-domain-core features are those in the positive-data-core features that are not part of the domain-core features.

2.3. Multifaceted Structure–Activity Relationship Analysis Based on Feature Clusters

As aforementioned, core features can ensure the prediction accuracy of ML models because they contain the key information required by ML modeling. However, merely employing core features to construct ML models loses the interaction between other features and core features. Here, we propose multifaceted structure–activity relationship analysis based on feature clusters (MRA-FC). MRA-FC constructs feature clusters with higher predictive performance by incorporating the interaction between other features and core features while retaining combined-core features, thus realizing the construction of multiple high-precision ML models and the extraction of relationships to distinguish and explain the complex influencing mechanisms latent in different feature spaces.

Note that different ML models can learn different structure–activity relationships because of their different modeling principles. Hence, MRA-FC can provide a multifaceted and interpretable analysis of materials structure–activity relationships by extracting rules from prediction results. We select several

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
FUNCTIONAL
MATERIALS**

www.afm-journal.de

Figure 2. The framework of MRA-FC. Multiple high-precision ML models are constructed based on feature clusters. According to the ML modeling principles, corresponding interpretable schemes are matched to realize the multifaceted analysis of materials structure–activity relationships.

models that are commonly employed in materials science but different in learning mechanisms and demonstrate their rule extraction strategies. Four ML models are multiple linear regression (MLR)[23] that learns linear correlation between conditional features and target property, K-nearest neighbor (KNN)[24] that learns the similarity between data samples, support vector regression (SVR)[25] that learns the optimal hyperplane in the high-dimensional feature space and Gaussian process regression (GPR)[26] that learns the probability distribution of data samples.

As shown in **Figure 2**, we provide rule extraction strategies for each ML model based on its modeling mechanism. For MLR model, we combine Pearson correlation coefficient (PCC) with linear coefficient (LC) to analyze the correlation between features and the target property. For KNN model, the decision tree that represents the target property can be obtained by using hierarchical clustering algorithm.[27] Note that SVR and GPR are both complex nonlinear models. The rule extraction of SVR is achieved by analyzing the difference between support vectors and other samples in different clusters guided by materials domain knowledge, thus to identify anomalous or novel structure–activity relationships. GPR is capable of outputting not only an estimate of the target performance of an unknown sample (expectation) but also the uncertainty of that prediction (variance). Therefore, the rule extraction of GPR can be achieved by analyzing materials samples with large prediction errors and high uncertainty.

Descriptors-DCML employs MRA-FC to conduct structure–activity relationship analysis, facilitating an effective trade-off between prediction accuracy and interpretability in ML models. The core features, selected through RST and materials domain knowl-

edge, inherently offer a high level of interpretability while maintaining the prediction accuracy of ML models. Feature clusters are constructed based on core features, which can further enhance the predictive performance of ML models. Subsequently, MRA-FC is utilized to perform a comprehensive analysis of the model's learning results, extracting structure–activity relationships between multiple factors and target properties. Our approach not only ensures the robustness of prediction accuracy but also allows for the extraction of interpretable insights into materials structure–activity relationships, thereby achieving an optimal balance between interpretability and accuracy in ML models.

3. Experiments

3.1. Datasets

To ensure the broad applicability of descriptors-DCML within the field of materials science, we conducted experiments on six material datasets, which demonstrates substantial diversity across multiple dimensions. The characteristics of the six datasets are summarized in **Table 1**. These datasets, as derived from published literature,[7,28–32] have a wide range of descriptors sources and encompass a variety of typical material systems, including alloys, polymers, and inorganic compounds. These material systems are not only of significant importance in materials science research but also widely used in practical engineering and technological fields, highlighting the representativeness and practical relevance of the datasets. Furthermore, six datasets exhibit notable diversity in the number of samples and

ADVANCED
SCIENCE NEWS

www.advancedsciencenews.com

ADVANCED
FUNCTIONAL
MATERIALS

www.afm-journal.de

Table 1. Statistical summary of the six datasets utilized in the experiments (* "Ratio" represents the proportion of the number of features to the sample size. "#" reflects the number).

	Materials category	# of samples	# of features	Ratio*	Target property	Descriptors source	Refs.
MD 1	NASICON-type SSEs	85	45	0.529	Energy barrier	ICSD database, experiments, literature, and BVSE-based calculations	[7]
MD 2	Cubic Li-argyrodites	51	32	0.627	Activation energy	Experiments, BVSE-based calculations	[28]
MD 3	Ni-based single crystal superalloys	1024	41	0.04	Creep rupture life	Experiments, literature and patents	[29]
MD 4	Nanocomposite solid polymer	160	5	0.031	Ionic conductivity	Experiments	[30]
MD 5	Disordered perovskite oxides	1016	81	0.079	Bandgap	ICSD database	[31]
MD 6	Ceramic material	119	234	1.966	Bandgap	DFT calculations, literature	[32]

feature-to-sample ratios. Specifically, the datasets not only include small-scale datasets (e.g., MD2) but also have large-scale datasets with a large number of descriptors (e.g., MD6, the number of features > 230) or a larger number of samples (e.g., MD3, MD5, the number of samples > 1000). Meanwhile, the number of samples span three distinct ranges: $[10^1, 10^2)$, $[10^2, 10^3)$ and $[10^3, \infty)$, enabling an in-depth analysis of the potential impact of varying sample sizes on model performance. The feature-to-sample ratios are distributed across four intervals: $[0, 0.25)$, $[0.25, 0.55)$, $[0.55, 1)$ and $[1, \infty)$, highlighting the influence of data density and complexity on the performance of descriptors-DCML. By conducting experiments on these diverse datasets, which are representative in real-world applications, the robustness and efficacy of descriptors-DCML can be comprehensively evaluated.

3.2. Experimental Setup

3.2.1. Model Selection and Parameter Setup

MLR, SVR, KNN and GPR models are selected to perform modeling task for the structure–activity relationships of energy barrier. Thereinto, grid search is employed to optimize the hyperparameters of ML models. The hyperparameters of NCOR-PSO method are set as follows: the number of populations is 50, Max = 50, c_1 = c_2 = 2, T = 0.6; the stopping condition is that the number of iterations reaches Max, or the populations output the same $g$best for 5 consecutive times and their values on the objective functions $H_1(x)$ and $H_2(x)$ are not greater than 0.1 and 0.05; k_1 and k_2 are set to 200 and 0.9, respectively.

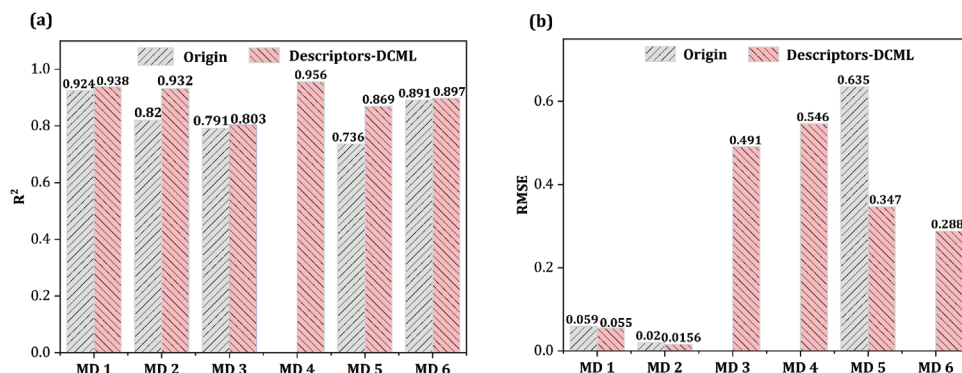#### 3.2.2. Performance Evaluation

To fully evaluate the effectiveness of core features and feature cluster in ML modeling, the experimental dataset is divided into training, validation and test set in the ratio of 2:1:1 and randomly spilt 10 times. The training set is used for feature selection based on NCOR-PSO, which is executed 200 times on the training set to obtain 200 different features combinations and their predictive performance. Subsequently, the feature decision table can

be constructed. The validation set is used to evaluate the predictive performance of features combinations obtained by the training set. Note that the feature decision table can only solve the discrete problem based on Algorithm 1. Therefore, K-means algorithm (K = 4) is used to cluster three performance evaluation metrics of ML models, and the class labels are regarded as the discrete variables representing the predictive performance of ML models; the candidate threshold value interval used for the reduction is [0.2, 0.6] and is taken continuously in steps of 0.01; whether the data-core features are retained in the modeling depends on their occurrence in the optimal grouping of the feature decision table with a frequency of not less than 0.3, and if they meet the criterion, they are retained; if not, they are eliminated. The test set is used to evaluate the predictive performance of core features and feature clusters obtained by the descriptors-DCML.

When evaluating the efficacy of multifaceted cluster-based constructive relationship analysis method, based on the whole dataset, core features and feature clusters recognition strategies in the descriptors-DCML are used to construct ML models in different feature spaces; then, the rule extraction strategies of MLR, KNN, SVR, GPR models are executed, and the rationality of rules are evaluated with the guidance of materials domain knowledge.

3.3. Results

Descriptors-DCML is evaluated on six material datasets based on the experimental setups described in Section 3.2. We collect and compare the prediction accuracy reported in the original reference,[7,28–32] as shown in **Figure 3**. The results clearly show that our method outperforms the prediction accuracy reported in the original reference,[7,28–32] demonstrating its broad applicability on various material systems. It is noteworthy that for the MD4 dataset, the original reference[30] presented only scatter plots of the prediction results, without specifying any evaluation metrics such as R^2 or RMSE. Consequently, the original performance metrics for MD4 dataset are not displayed in Figure 3. However, our method achieves a prediction accuracy of 95.6% on the MD4 dataset, which is widely regarded as an excellent

**ADVANCED
SCIENCE NEWS**
www.advancedsciencenews.com

**ADVANCED
FUNCTIONAL
MATERIALS**
www.afm-journal.de

Figure 3. Prediction accuracy comparison between origin and descriptors-DCML. "Origin" and "Descriptors-DCML" represent the prediction accuracy from original reference and our proposed method, respectively. a,b) Denote the performance comparisons on the R^2 and RMSE metrics, respectively.
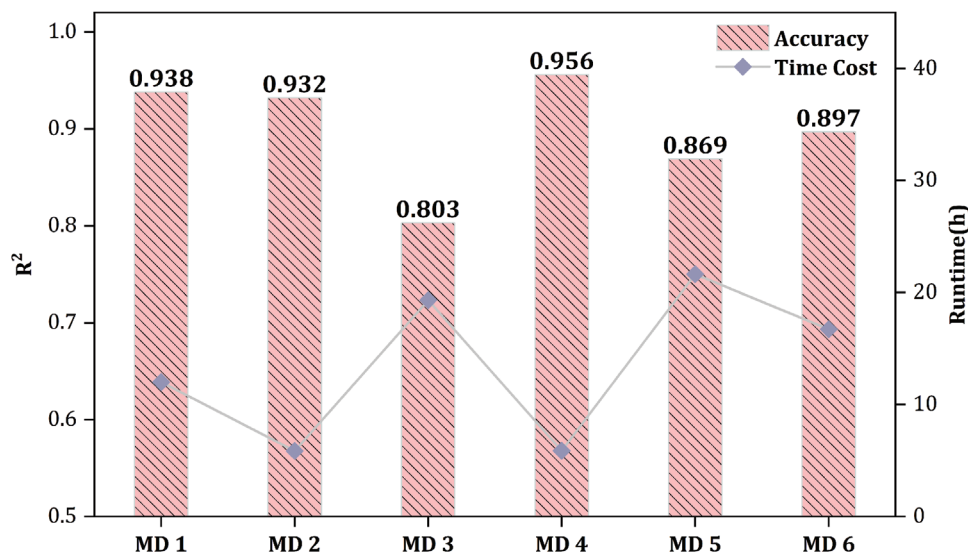
performance within the ML community. The original references[29,32] associated with MD3 and MD6 dataset did not report the performance on RMSE metric, so only the performance with respect to the R^2 metric is presented in Figure 3a.

We further perform a comprehensive analysis of the runtime of descriptors-DCML, as shown in **Figure 4**. Although our method requires a certain computational investment during the training process, this is well-justified by its ability to ensure high prediction accuracy and interpretability. The prediction accuracy of descriptors-DCML across diverse material datasets clearly demonstrates its robustness and reliability in a variety of material systems, thus validating the efficacy of integrating RST with materials domain knowledge to select core features and feature clusters for ML models. Due to space limitations in this paper, a detailed analysis of the intermediate results of descriptors-DCML is presented using the MD1 dataset as a representative example. The M1 dataset involves 85 NASICON-type compounds. Each compound in the dataset contains 45 descriptors and 1 target property (energy barrier). The detailed information of 45 descriptors can be found in Table S1 (Supporting Information). Further-

more, the detailed analysis for the remaining five datasets are presented in the Supporting Information.

3.3.1. The Recognition of Core Features

The Recognition of Domain-Core Features: Ion transport performance is directly described by ionic conductivity. However, measuring ionic conductivity via experiments is very time-consuming. As a result, the available ionic conductivity data of NASICON-type SSE materials are quite scarce. Activation energy is another physical quantity to represent the ion transport performance, which has negative correlation with ionic conductivity.[17] However, it is often measured by experiments at the same time as ionic conductivity. Energy barrier is an important component of activation energy, representing the local energy required for ion migration. At present, many low-cost calculation methods are proposed to obtain energy barrier value of ions in SSE materials,[33,34] and energy barrier has been used as one of the important indicators for rapid screening of superconducting materials. Therefore, the proven materials domain knowledge about

Figure 4. The time cost and prediction accuracy of descriptors-DCML on six material datasets.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
FUNCTIONAL
MATERIALS**

www.afm-journal.de

Table 2. Predictive performance of ML models based on key-domain-core features (with standard deviation).

Model	RMSE$_{CV}$	RMSE$_{Test}$	R^2_{Test}
MLR	0.0820 ± 0.0046	0.0973 ± 0.0102	0.8139 ± 0.0539
SVR	0.0836 ± 0.0058	0.0962 ± 0.0093	0.8179 ± 0.0536
KNN	0.0816 ± 0.0070	0.0915 ± 0.0151	0.8350 ± 0.0605
GPR	0.1291 ± 0.0244	0.1352 ± 0.0464	0.6283 ± 0.0464

ionic conductivity or activation energy structure–activity relationships can also be used to analyze the structure–activity relationships of energy barrier.

Bottlenecks represent the smallest cross-sectional area in the ion transport channel. Many studies have explored the influence of bottlenecks on activation energy of NASICON-type compounds. For instance, Sai et al.[35] studied the relationship between bottlenecks and the ion migration ability of Na$^+$ in NASICON-type compounds and found that activation energy decreased obviously with the increase of bottleneck size. Martinez et al.[19] studied the relationship between activation energy and the bottleneck of LiMM$'$(PO$_4$)$_4$ type compounds by X-ray diffraction method, and found that activation energy dropped sharply when the bottleneck size reached 2.04Å, while almost unchanged above 2.04Å. Guin et al.[17] extensively investigated and analyzed the crystal structure, composition diversity and ionic conductivity of NASICON-type compounds, and found that the linear correlation between the T1 bottleneck size defined by Losilla et al.[36] and activation energy was as high as 0.56515. Therefore, descriptors representing bottleneck concepts will be used as the domain-core features, including sub-domain-core features BT1, the combination of BT1 and BT2, Min$_{BT}$ and RT.

Further, key-domain-core features are obtained based on the predictive performance of each sub-domain-core feature. **Table 2** shows the average predictive performance of MLR, SVR, KNN and GPR models based on key-domain-core features for the total training set (training set + validation set) and the test set under 10 data divisions. CV RMSE is the 10-fold cross-validation RMSE on the total training set, Test RMSE and Test R^2 are the average RMSE and R^2 on the test set. As shown in Table 2, accurate and stable energy barrier prediction can be achieved by applying only the key-domain-core features to construct ML models, indicating that the key-domain-core features indeed contain key information of energy barrier prediction and can ensure the predictive performance of ML models.
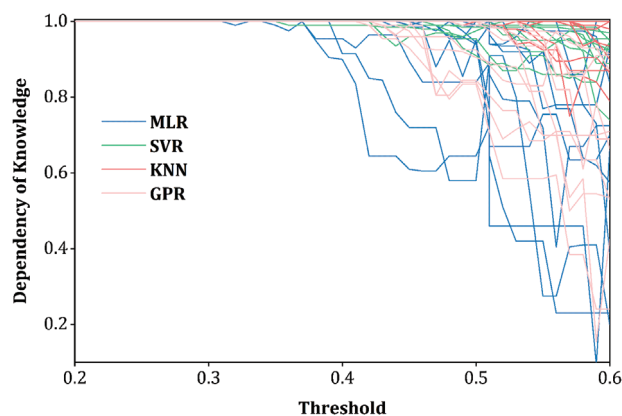
The Recognition of Data-Core Features: According to Algorithm 1, the first step to recognize data-core features is to construct the feature decision table. We execute NCOR-PSO on the NASICON-type compounds dataset 200 times to obtain different sets of features combinations and their predictive performance. Then K-Means is employed to cluster the predictive performance of each ML model. Finally, the feature decision table is formed by coupling 200 sets of features combinations with class labels of the predictive performance of ML models.

It is necessary to perform the reduction operation for the feature decision table. The change of knowledge dependency during the reduction process is shown in **Figure 5**. Overall, as the threshold increases, the knowledge dependence in the feature decision table remains at 1 and then gradually decreases. This

indicates that redundant information does exist in the feature decision table, and removing it can achieve the reduction without decreasing the classification ability of the feature decision table. However, removing too much information will decrease the classification ability of the feature decision table. Therefore, according to Algorithm 1, we use maximum value within threshold value range that makes the knowledge dependency equal to 1 to preform reduction on the feature decision table.
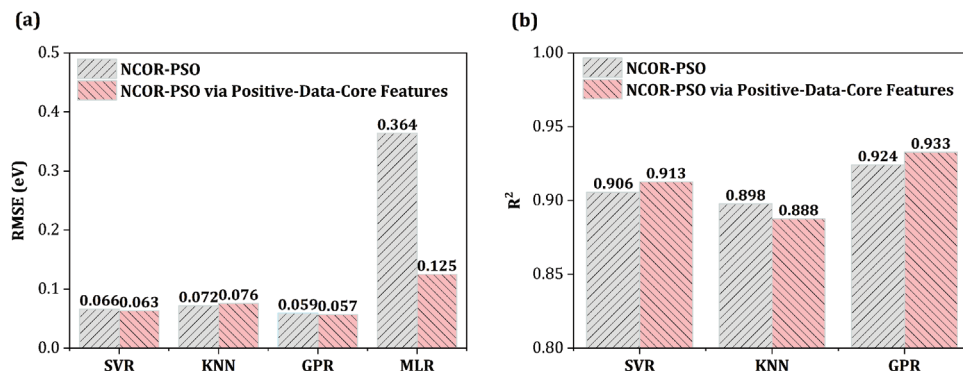
After the feature decision table is reduced, data-core features can be obtained using the direct core computation algorithm[20] based on the positive region. The positive-data-core features can be recognized by removing the data-core feature whose occurrence frequency in the optimal class of the feature decision table is less than 0.3. By fixing the positive-data-core features, feature clusters can be obtained by executing NCOR-PSO. As shown in **Figure 6**, under 10 kinds of data divisions, the predictive performance of SVR, KNN and GPR models based on feature clusters are approximate to the features combinations directly obtained by NCOR-PSO. The predictive performance of the MLR model is found to be deficient and is not suitable for assessment via R^2, so only RMSE is reported. The MLR model based on NCOR-PSO has the largest RMSE on the test set, and feature clusters can obviously decrease the RMSE of the test set. Based on the above analysis, the positive-data-core features can indeed ensure or improve the predictive performance of ML models to a certain extent, indicating that these descriptors contain the key information of energy barrier prediction.

The Construction of Combined-Core Features: **Figure 7** shows the predictive performance of ML models based on feature clusters obtained by combined-core features. The predictive performance of the MLR model utilizing NCOR-PSO is found to be

Figure 5. Threshold sensitivity analysis of knowledge dependence of feature decision table.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
FUNCTIONAL
MATERIALS**

www.afm-journal.de

Figure 6. Predictive performance of ML models based on feature clusters obtained by positive data-core features. a,b) Denote the performance comparisons on the RMSE and R^2 metrics, respectively.

deficient and is not suitable for assessment via R^2, so only RMSE is reported. Under 10 kinds of data divisions, the predictive performance of MLR, KNN, SVR and GPR models are all improved than features combinations directly obtained by NCOR-PSO, and the RMSE of MLR model is greatly decreased. It can be inferred that the combination of domain-core features and positive-data-core features can establish high-precision ML models. The accurate prediction of the sodium ion energy barrier enables the rapid screening of high-performance electrolytes, thereby accelerating the design and discovery of new advancing materials.

3.3.2. Multifaceted Structure–Activity Relationship Analysis

In this section, we use multifaceted structure–activity relationship analysis based on feature cluster (MRA-FC) proposed in Section 2.3 to extract rules from ML models (MLR, KNN, SVR, GPR). ML models are constructed via MRA-FC on the whole NASICON-type SSE compounds dataset and the extracted results by MRA-FC are summarized in **Figure 8**.

The specific content of 9 relations extracted by ML models are as follows:

Relation 1:Larger unit cell size causes lower energy barrier.
Relation 2:Higher temperature causes lower energy barrier.
Relation 3:Larger channel size causes lower energy barrier.
Relation 4:Larger effective electronegativity of element in X site causes lower energy barrier.

Relation 5:Smaller effective electronegativity and larger ionic radius of element in M site cause lower energy barrier.
Relation 6:E_Na(3) can effectively differentiate the quality of compound, and larger E_Na(3) causes lower energy barrier.
Relation 7:Descriptors V_Na(1)O$_6$ and Min_BT are in the last layer of prediction decision tree shown in Figure 10, denoting that they can predict the energy barrier more accurately.
Relation 8:Larger avg_R_X and avg_R_M cause lower energy barrier.
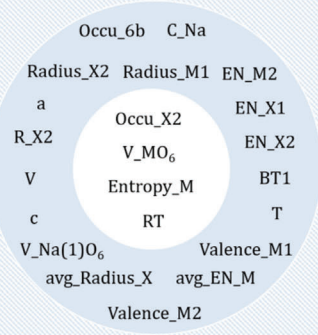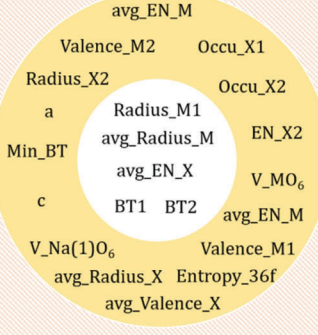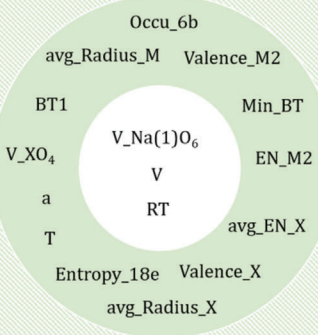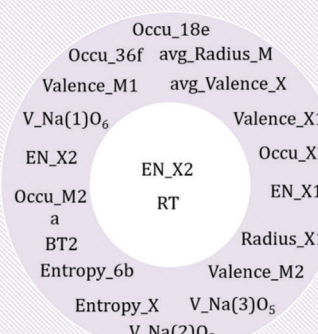Relation 9:Smaller avg_R_X and avg_R_M cause larger a, V, V_XO_4, $V\_Na(1)O_6$, BT1, Min_BT, RT and lower energy barrier.

The optimal feature cluster of the MLR model contains 22 descriptors. By analyzing the LC and PCC values (seen in Figure S1, Supporting Information) between each descriptor and energy barrier, 13 descriptors with the same LC and PCC symbols are saved, as shown in **Figure 9**. Based on that, 5 relations are obtained and Relations 1, 3, 4, 5 conform to materials domain knowledge. The reason to obtain Relation 1 may be that most compounds collected in the dataset have positive coefficient of thermal expansion, whereas in fact NASICON-type compounds also have negative coefficient and zero coefficient of thermal expansion.

For the interpretation of KNN model, a complete clustering hierarchical tree is obtained (seen in Figure S2, Supporting Information) by hierarchical clustering algorithm.[27] Based on that,
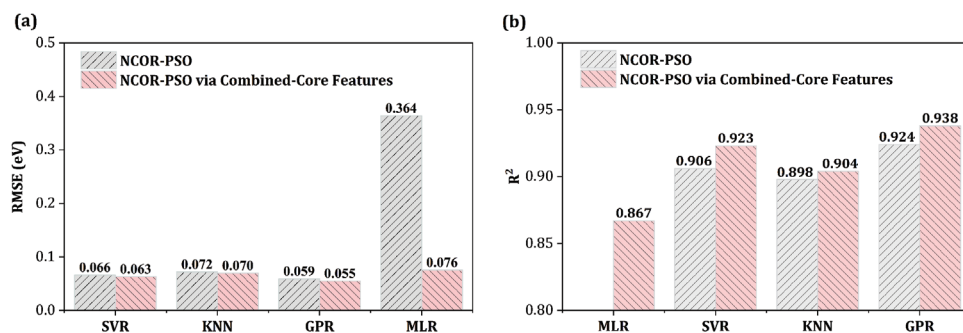
Figure 7. Predictive performance of ML models based on feature clusters obtained by combined-core features. A,B) Denote the performance comparisons on the RMSE and R^2 metrics, respectively.

Figure 8. Summary of extracted results obtained by MRA-FC. The summary includes core features and feature cluster used to construct high-precision ML models, interpretation methods, the extracted relations, and the literature that allows for the validation of relations.
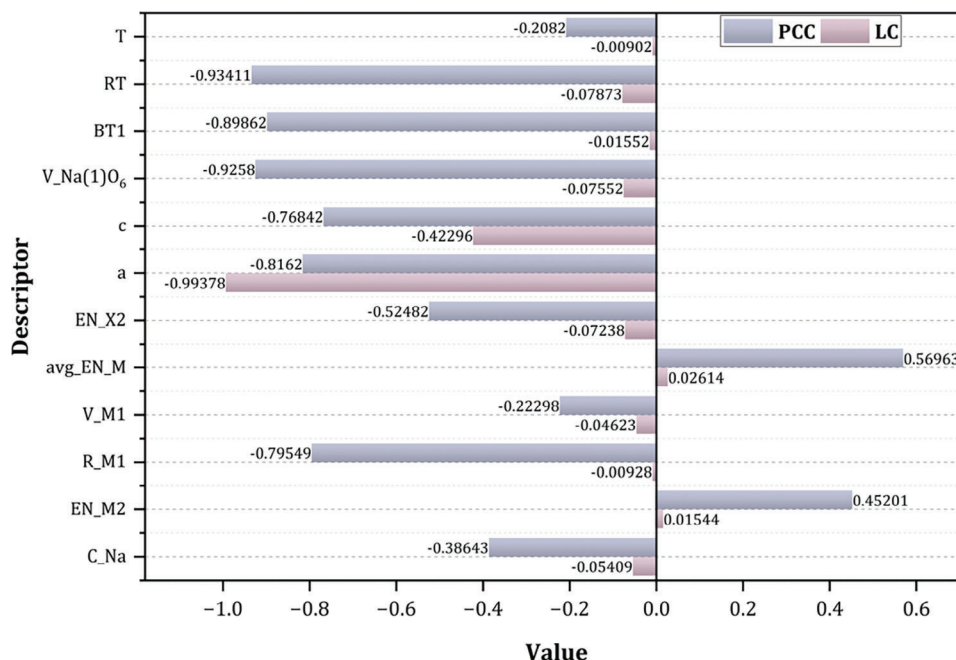
ADVANCED
SCIENCE NEWS

www.advancedsciencenews.com

ADVANCED
FUNCTIONAL
MATERIALS

www.afm-journal.de

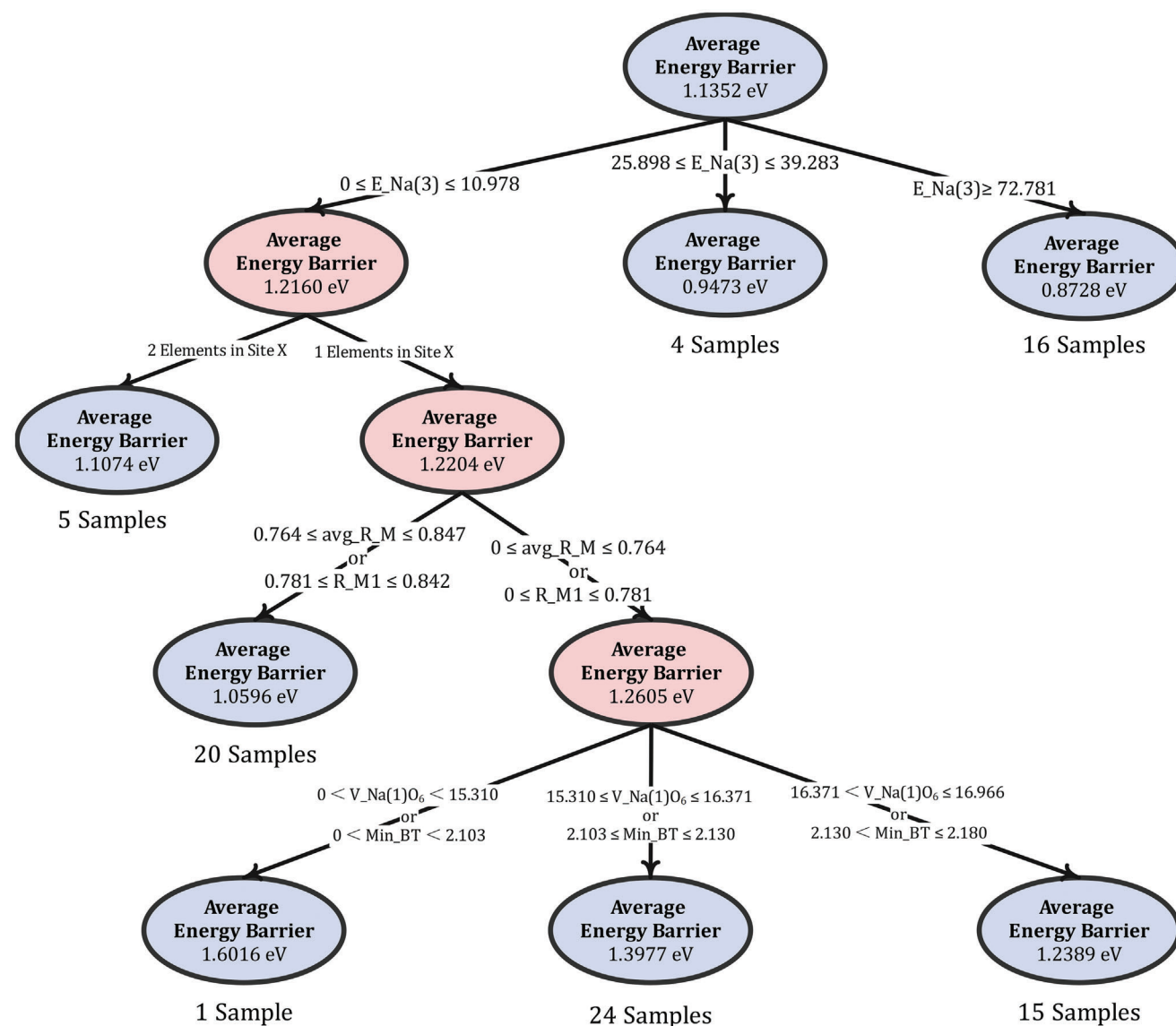**Figure 9.** The LC and PCC values of remaining 13 descriptors trained by MLR model.

the energy barrier prediction decision tree can be constructed as shown in **Figure 10**. 5 relations are extracted by analyzing the descriptor values of leaf nodes and the splitting conditions of each layer of the decision tree. Relation 6 reflects ion transport mechanisms. Zou et al.[37] pointed out that in NASICON-type compounds, Na^+ tends to preferentially occupy the $6b$ site which has lower energy. When the amount of Na^+ is greater than the maximum number that $6b$ site can be occupied, the excessive Na^+ will occupy the $18e$ site. If the occupancy of the $36f$ site becomes more accessible, it indicates that the energy barrier encountered by the ions during migration is decreased. There is no relevant domain knowledge to directly support Relation 7. By communicating with domain experts in our research group, one possible explanation for Relation 7 is that Na^+ tends to preferentially occupy the Na(1) site and the polyhedron volume of $Na(1)O_6$ has a stable influence on the energy barrier compared with the polyhedron centered at other sites, Therefore, it plays a greater role in influencing the energy barrier, as does Min_{BT}.

Support vector are samples that are more difficult to predict accurately compared with other samples. They may represent abnormal samples or novel structure–activity relationships in the dataset. According to the interpretable strategy proposed in Section 2.3, all samples are mapped to high-dimensional descriptor space for clustering, as shown in **Figure 11**. By analyzing the differences in energy barrier values and descriptor values of samples in different clusters, 2 relations are extracted. There is no relevant domain knowledge to explain these two rules, but it may be due to the fact that most synthesis experiments focus more on element doping at the X site. In addition, we compare the differences in descriptor value between support vectors and other samples within the same cluster. Table S2 (Supporting Information) shows the detailed information of these 8 support vectors. By communicating with domain experts in our research group,

we find that the samples corresponding to these 8 support vectors do not contain specific structure–activity relationships. This is more likely due to the limited number of compounds in the dataset and the inhomogeneous distribution of compounds in the chemical element space. For example, the Co, Mo, and V elements contained in compounds $Na_{4.0002}Co_3Mo_{22.332}O_{72}$ and $Na_6Zr_{12}V_{2.394}P_{15.606}O_{72}$ as support vectors appear only three times in the dataset.

GPR model trained on the optimal feature cluster is used to predict energy barrier and the uncertainty is calculated based on the mean value $\bar{y}$ and standard deviation σ of the predicted distribution. The compound with high uncertainty may be an abnormal sample or represents an unexplored structure–activity relationship. To quantify the uncertainty of predicted values, the fluctuation range at 95.4% confidence level is provided, as shown in **Figure 12**. As with the results of SVR model, the difference between compounds with large prediction errors and high uncertainties and others in energy barrier may be caused by the inhomogeneous distribution of compounds in the chemical element space.

Cubic Li-argyrodites and NASICON-type compounds are both classified as inorganic non-metal materials and are widely regarded as promising candidates for high-performance SSEs. Due to their exceptional ionic conductivity and chemical stability, these materials have demonstrated considerable potential in energy storage systems. Specifically, the cubic Li-argyrodites correspond to the MD2 dataset, with activation energy serving as the target property. Since energy barrier can be considered as an approximation of activation energy, it is conceivable that the structure–activity relationships of these two datasets may exhibit similarities. To explore this hypothesis further, we employ multifaceted structure–activity relationship analysis method proposed in Section 2.3 to dig out relevant rules from the high-precision

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
FUNCTIONAL
MATERIALS**

www.afm-journal.de

Figure 10. The energy barrier prediction decision tree. Material samples are divided into different leaf nodes of the decision tree based on the range of feature values.

ML models trained on the MD2 dataset. A detailed analysis of the results and corresponding visualizations can be found in the Figure S3 (Supporting Information). The structure–activity relationships of activation energy extracted from MD2 dataset are as follows:

Relation 1: Larger channel size causes lower energy barrier.
Relation 2: Smaller average ionic radius of the anions causes lower activation energy.
Relation 3: Descriptors $Sconf_c$ can effectively differentiate the quality of SSEs, and larger $Sconf_c$ causes lower activation energy.
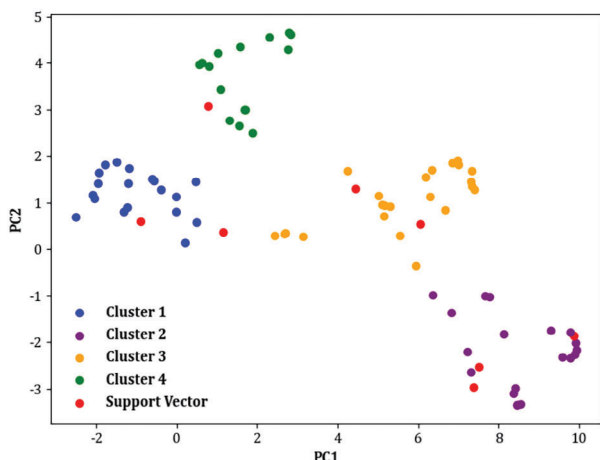
Relation 1 is consistent with Relation 3 extracted from the NASICON-type dataset, revealing a distinct negative correlation between channel size and ion transport properties. This rela-

tion demonstrates broad applicability in SSEs materials, further substantiating the capability of our method to effectively capture such fundamental structure–activity relationships. Relation 2 is a novel finding, first introduced in study.[38] Relation 3 has been previously explored and validated in study.[28] The extraction of these three relations further demonstrates the high flexibility of our approach, indicating that it not only effectively extracts common rules but also unveils new rules beyond the traditional framework of structure–activity relationships.

4. Discussion and Conclusion

Leveraging ML models based on limited descriptors combinations fails to explore complex materials structure–activity relationships. In this study, we introduce RST into the field of materials science for exploring diverse mechanisms latent in

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
FUNCTIONAL
MATERIALS**

www.afm-journal.de

Figure 11. The clustering of all samples in a high-dimensional descriptor space. The red dots represent different support vectors.

materials data and propose a descriptors divide-and-conquer method for machine learning (descriptors-DCML) to unveil multifaceted structure–activity relationships latent in materials data. Our approach integrates RST with materials domain knowledge to identify core features and feature clusters for different ML models and thus rule extraction strategies are employed to dig out complex influencing mechanisms of materials. Experiment on six materials datasets across different material systems show that descriptors-DCML not only exhibits high prediction accuracy of material properties but also demonstrates robust capabilities in rule extraction and interpretable analysis of structure–activity relationships. Our work provides a high-precision ML modeling approach and interpretable structure–activity relationships extraction strategies for AI-driven materials research, which hold significant potential for advancing high-throughput materials screening and enhancing real-time prediction systems for material properties, thereby contributing to the advancement of materials science.
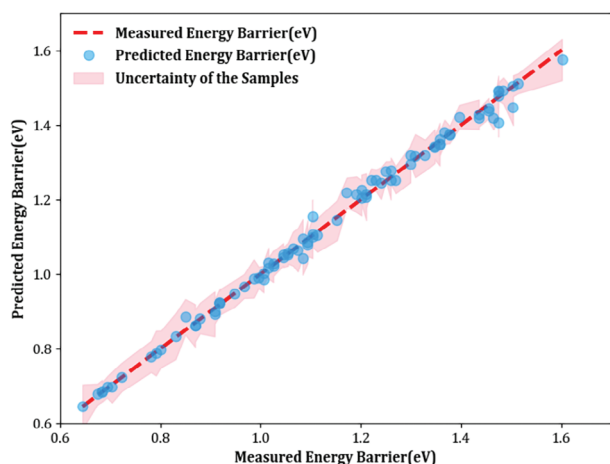
Figure 12. The predicted energy barrier values and uncertainties of material samples trained by GPR model.

While descriptors-DCML demonstrates advantages in both prediction accuracy and interpretability, there are still areas that necessitate refinement in future research endeavors. Varying dataset sizes, descriptor choice, or specific material systems may introduce potential biases or limitations to the learning results of ML models, and data argumentation or transfer learning can be considered to address these challenges. We validate extracted rules by consulting relevant reference. For the novel rules awaiting validation, there are some approaches that provide insights for the confirmation of novel materials structure–activity relationships, e.g., integrating ML with omics technologies and computational modeling techniques[39–43] (e.g., quantum mechanics). Improving computational efficiency of algorithm is also important, so future work could consider leveraging distributed computing frameworks[44,45] to better accommodate complex computational tasks.

Supporting Information

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

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1] M. A. Shandiz, R. Gauvin, *Comp. Mater. Sci.* **2016**, *117*, 270.

[2] Y. Liu, T. Zhao, W. Ju, S. Shi, *J. Materiomics* **2017**, *3*, 159.

[3] A. D. Sendek, B. Ransom, E. D. Cubuk, L. A. Pellouchoud, J. Nanda, E. J. Reed, *Adv. Energy Mater.* **2022**, *12*, 2200553.

[4] J. Li, M. Zhou, H. H. Wu, L. Wang, J. Zhang, N. Wu, K. Pan, G. Liu, Y. Zhang, J. Han, X. Liu, X. Chen, J. Wan, Q. Zhang, *Adv. Energy Mater.* **2024**, *14*, 2304480.

[5] A. K. Mishra, S. Rajput, M. Karamta, I. Mukhopadhyay, *ACS Omega* **2023**, *8*, 16419.

[6] Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, S. Shi, *J. Materiomics* **2023**, *9*, 798.

[7] Y. Liu, X. Zou, S. Ma, M. Avdeev, S. Shi, *Acta Mater.* **2022**, *238*, 118195.

[8] Y. Xu, Y. Zong, K. Hippalgaonkar, *J. Phys. Commun.* **2020**, *4*, 055015.

[9] Y. Zhang, T. Zhan, Y. Sun, L. Lu, B. Chen, *ChemSusChem* **2024**, *17*, 202301284.

[10] C. Li, Z. Huang, H. Hao, Z. Shen, G. Zhao, B. Xu, H. Liu, *J. Materiomics* **2025**, *1*, 100848.

[11] S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37.

[12] J. Ranstam, J. A. Cook, *Br. J. Surg.* **2018**, *105*, 1348.

[13] Q. Zhou, H. Mo, Y. Deng, *Mathematics* **2020**, *8*, 142.

[14] A. Gaeta, V. Loia, L. Lomasto, F. Orciuoli, *Appl. Intell.* **2023**, *53*, 15993.

[15] B. Walczak, D. L. Massart, *Chemom. Intell. Lab. Syst.* **1999**, *47*, 1.

[16] T. Famprikis, P. Canepa, J. A. Dawson, M. S. Islam, C. Masquelier, *Nat. Mater.* **2019**, *18*, 1278.

[17] M. Guin, F. Tietz, *J. Power Sources* **2015**, *273*, 1056.

[18] D. T. Qui, J. J. Capponi, M. Gondrand, M. Saib, J. C. Joubert, R. D. Shannon, *Solid State Ionics* **1981**, *3*, 219.

[19] A. Martinez-Juarez, C. Pecharromán, J. E. Iglesias, J. M. Rojo, *J. Phys. Chem. B* **1998**, *102*, 372.

[20] T. F. Zhang, J. M. Xiao, X. H. Wang, *Acta Electron. Sin.* **2005**, *33*, 2008.

[21] P. Wang, J. He, Z. Li, *Inf. Sci.* **2023**, *632*, 555.

[22] R. Poli, J. Kennedy, T. Blackwell, *Swarm Intell.* **2007**, *1*, 33.

[23] K. A. Marill, *Acad. Emerg. Med.* **2004**, *11*, 94.

[24] P. Cunningham, S. J. Delany, *ACM Comput. Surv.* **2021**, *54*, 1.

[25] A. J. Smola, B. Schölkopf, *Stat. Comput.* **2004**, *14*, 199.

[26] E. Schulz, M. Speekenbrink, A. Krause, *J. Math. Psychol.* **2018**, *85*, 1.

[27] F. Murtagh, P. Contreras, *WIREs Data Min. Knowl.* **2012**, *2*, 86.

[28] Q. Zhao, M. Avdeev, L. Chen, S. Shi, *Sci. Bull.* **2021**, *66*, 1401.

[29] F. Yang, W. Zhao, Y. Ru, Y. Pei, S. Li, S. Gong, H. Xu, *Mater. Des.* **2023**, *232*, 112174.

[30] M. R. Johan, S. Ibrahim, *Commun. Nonlinear Sci. Numer. Simul.* **2012**, *17*, 329.

[31] A. Ihalage, Y. Hao, *npj Comput. Mater.* **2021**, *7*, 75.

[32] X. Wang, Y. Xu, J. Yang, J. Ni, W. Zhang, W. Zhu, *Comp. Mater. Sci.* **2019**, *169*, 109117.

[33] J. C. Bachman, S. Muy, A. Grimaud, *Chem. Rev.* **2016**, *116*, 140.

[34] Y. Miyajima, Y. Saito, M. Matsuoka, Y. Yamamoto, *Solid State Ionics* **1996**, *84*, 61.

[35] W. Sai, G. B. Chai, N. Srikanth, *Adv. Theory Simul.* **2020**, *3*, 2000048.

[36] E. R. Losilla, M. A. Aranda, S. Bruque, M. A. París, J. Sanz, A. R. West, *Chem. Mater.* **1998**, *10*, 665.

[37] Z. Zou, N. Ma, A. Wang, Y. Ran, T. Song, Y. Jiao, J. Liu, H. Zhou, W. Shi, B. He, D. Wang, Y. Li, M. Avdeev, S. Shi, *Adv. Energy Mater.* **2020**, *10*, 2001486.

[38] Q. Zhao, L. Zhang, B. He, A. Ye, M. Avdeev, L. Chen, S. Shi, *Energy Storage Mater.* **2021**, *40*, 386.

[39] A. V. Singh, P. Bhardwaj, P. Laux, P. Pradeep, M. Busse, A. Luch, A. Hirose, C. J. Osgood, M. W. Stacey, *Front. Toxicol.* **2024**, *6*, 1461587.

[40] A. V. Singh, M. Varma, M. Rai, S. P. Singh, G. Bansod, P. Laux, A. Luch, *Adv. Intell. Syst.* **2024**, *6*, 2300366.

[41] A. V. Singh, A. Shelar, M. Rai, P. Laux, M. Thakur, I. Dosnkyi, G. Santomauro, A. K. Singh, A. Luch, R. Patil, J. Bill, *J. Agric. Food Chem.* **2024**, *72*, 2835.

[42] A. V. Singh, M. H. D. Ansari, D. Rosenkranz, R. S. Maharjan, F. L. Kriegel, K. Gandhi, A. Kanase, R. Singh, P. Laux, A. Luch, *Adv. Healthcare Mater.* **2020**, *9*, 1901862.

[43] A. V. Singh, M. Varma, P. Laux, S. Choudhary, A. K. Datusalia, N. Gupta, A. Luch, A. Gandhi, P. Kulkarni, B. Nath, *Arch. Toxicol.* **2023**, *97*, 963.

[44] X. Sun, Y. He, D. Wu, J. Z. Huang, *Big Data Min. Anal.* **2023**, *6*, 154.

[45] J. Knap, C. E. Spear, O. Borodin, K. W. Leiter, *Nanotechnology* **2015**, *26*, 434004.