Journal of Materiomics 11 (2025) 101066

Contents lists available at ScienceDirect

Journal of Materiomics

journal homepage: www.journals.elsevier.com/journal-of-materiomics/

Research paper

Domain knowledge-assisted materials data anomaly detection towards constructing high-performance machine learning models

Yue Liu [a], Shuchang Ma [a], Zhengwei Yang [a], Duo Wu [a], Yali Zhao [a], Maxim Avdeev [d, e], Siqi Shi [b, c, *]

[a] State Key Laboratory of Materials for Advanced Nuclear Energy & School of Computer Engineering and Science, Shanghai University, Shanghai, 20444, China
[b] State Key Laboratory of Materials for Advanced Nuclear Energy & School of Materials Science and Engineering, Shanghai University, Shanghai, 200444, China
[c] Materials Genome Institute, Shanghai University, Shanghai, 200444, China
[d] Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC NSW, 2232, Australia
[e] School of Chemistry, The University of Sydney, Sydney, 2006, Australia

ARTICLE INFO

Article history:
Received 18 December 2024
Received in revised form
19 February 2025
Accepted 19 March 2025
Available online 24 April 2025

Keywords:
Machine learning
Materials science
Data anomaly
Domain knowledge

ABSTRACT

Machine learning (ML) is widely applied to accelerate materials design and discovery due to its outperforming capability of data analysis and information extraction. However, experimental and computational errors typically lead to emerging data anomalies, harming the performance of ML models. Most currently used anomaly detection methods are purely data-driven, which has limited capability of learning complicated factors in materials data. Here, we propose a domain knowledge-assisted data anomaly detection (DKA-DAD) workflow, where materials domain knowledge is encoded as symbolic rules. Three detection models are designed for evaluating the correctness of individual descriptor value, correlation between descriptors, and similarity between samples, respectively, and one modification model is constructed for comprehensive governance. We construct 180 synthetic datasets by injecting noise into 60 structured materials datasets collected from materials ML studies, to validate its potential utility and applications. DKA-DAD achieves a 12% F1-score improvement in anomaly detection accuracy on synthetic datasets compared to purely data-driven approach and the ML models trained on materials datasets processed through DKA exhibit an average 9.6% improvement in R^2 for the property prediction. Our work provides a data anomaly detecting approach under the guidance of materials domain knowledge towards accelerating materials design and discovery based on ML.

1. Introduction

Data-driven machine learning (ML) is widely used in materials research owing to its excellent prediction accuracy, efficiency, and applicability [1–6], of which generalizability and reliability heavily relies on the quality of input data [7]. Notably, data accuracy is one of the critical dimensions in ensuring the construction of high-performance ML models. Currently, materials data are primarily acquired *via* experimental measurements, theoretical calculations, and industrial production protocols [8–11]. However, those approaches are prone to introducing uncertainties such as experimental environment differences or calculation software errors, thus decreasing the data accuracy and reducing the model performance [12]. Therefore, it is essential to detect and revise or exclude the anomalies of material samples, thereby enhancing the accuracy of ML predictions.

In recent years, the critical role of data accuracy in ML applications has garnered increased attention from the research community. Consequently, a variety of techniques, including manual inspection, basic statistical methods, and advanced classification algorithms, have been employed to evaluate and enhance data accuracy. For instance, Beal *et al.* [13] employed a distance-based anomaly detection algorithm to identify and remove material samples far from the center of the sample during the construction of artificial neural network (ANN) model, achieving R^2 of 92% in the

* Corresponding author. State Key Laboratory of Materials for Advanced Nuclear Energy & School of Materials Science and Engineering, Shanghai University, Shanghai, 200444, China.
E-mail address: sqshi@shu.edu.cn (S. Shi).

https://doi.org/10.1016/j.jmat.2025.101066

ionic conductivity prediction task. Hemmati-Sarapardeh *et al.* [14] applied the lever detection method to identify anomalous data through the William plot of the Hat matrix components and detected six outliers due to erroneous reports in the literature or errors made by researchers in experimental measurements. This correction led to an impressive accuracy of 98.86% in predicting the density of ionic liquids by Least Square Support Vector Machine (LSSVM) algorithm. Similarly, Amiri-Ramsheh *et al.* [15] detected four anomalies outside the valid range using the leverage detection method in their study on the solubility of carbon dioxide in polymers.

Most of the above studies rely on ML or statistical learning methods, evaluating anomalies within dataset solely from the perspective of data distribution. In materials science, available domain knowledge also plays a decisive role in influencing ML model outcomes. For instance, Wenzlick *et al.* [16] evaluated anomalies in alloy datasets by determining the optimal number of clusters for *K*-means based on the composition and values of the processing parameters, leveraging materials science principles and expertise in expected alloy behavior. Furthermore, by incorporating insights rooted in physics and materials science into the *K*-means clustering, they enhanced interpretability of both the clustering model and the anomaly detection model. Li *et al.* [17] introduced domain knowledge into the feature selection process to determine the feature subset when predicting the properties of soft magnetic metallic glasses. The results showed that domain knowledge-assisted feature design can greatly reduce the number of features without significantly decreasing the prediction accuracy. In our previous work, we explored the issues related to data quality in materials datasets, specifically highlighting the necessity and effectiveness of integrating domain knowledge into the entire ML process. Based on this, we integrated domain knowledge into the data-driven feature selection process, successfully identifying descriptors that not only align with domain expertise but also exhibit low internal correlation [18,19]. Therefore, plenty of evidence underscores the value of conducting anomaly detection and evaluation from multiple perspectives with the assistance of expert experience and domain knowledge, thus further improving the comprehensiveness of anomaly detection and obtaining higher-quality datasets.

Here, a Domain Knowledge-Assisted Data Anomaly Detection (DKA-DAD) workflow is proposed, which consists of four parts: Single-Descriptor Accuracy Detection model (S-DAD), Multi-Descriptor Correlation Detection Model (M-DCD), Sample Reliability Detection model (SRD) and Modification model. This workflow detects potential anomalies in datasets hierarchically from both data-driven and domain knowledge perspectives to provide high-quality datasets for ML modelling. The effectiveness of DKA-DAD is demonstrated on 180 simulated datasets and 60 real materials datasets. Experiments show not only does DKA-DAD detect and correct potential anomalies but also improve the predictive performance of materials property prediction model to varying degrees.

The remainder of the paper is structured as follows: Section 2 introduces the extraction and symbolic representation of descriptor expert knowledge and the structure of the proposed DKA-DAD framework. Section 3 illustrates the effectiveness of DKA-DAD by conducting experiments on simulated and real datasets. Finally, conclusions and outlook are presented in Section 4.

2. Anomaly detection supervised by domain knowledge

Based on our previous work [20], Domain Knowledge-Assisted Data Anomaly Detection workflow (DKA-DAD) is proposed in which the detection process is optimized and a modification model is added to detect possible abnormal data in single-dimensional, multi-dimensional and full-dimensional descriptor space based on data and knowledge. As depicted in Fig. 1, the overall framework of DKA-DAD consists of four parts: Single-Descriptor Accuracy Detection model (S-DAD), Multi-Descriptor Correlation Detection Model (M-DCD), Sample Reliability Detection model (SRD), and Modification model. Each detection model is executed through a collaboration with domain knowledge and data-driven approaches, generating the modification strategy for anomaly detection. The modification model evaluates all strategies, with domain knowledge remaining a vital component throughout the process. It ensures the unified decision making on the modification strategies generated during the detection process, determining whether data requires modification and the approach for such, ultimately leading to enhanced accuracy. At last, to verify the effectiveness of anomaly detection, the high-accuracy data is subjected to regression testing through the S-DAD, M-DCD, and SRD models. Throughout the entire data anomaly detection process, the detection-modification procedure may be executed once or multiple times, ensuring comprehensive and precise detection of all anomalies.

2.1. Anomaly detection based on the rule for value of individual descriptors

Note that different descriptors possess specific physical meanings, data types, value ranges, data sources, and quantification methods. Hence, we obtain information of specific material types from relevant literature and material experts as illustrated by an example in Table 1. Since multiple samples within the same dataset may originate from different sources, it is crucial to evaluate whether the data type, value range, and units of each descriptor in any given sample are consistent with those obtained from domain knowledge.

On the basis of various phenomena of materials data mentioned above, we here represent it as define *the rule for value of descriptors* (Hereinafter referred to as Rule 1), shown in Eqs. (1) and (2), which uses the information of descriptors, including value detection and grammar detection. The former determines whether the value range of each descriptor is consistent with the domain knowledge, while the latter determines whether the data type and unit of each descriptor is consistent with the domain knowledge.

The rule for value of descriptors: Given a triple $\langle D,T,R\rangle$, where $D=\{\overrightarrow{d_1},\overrightarrow{d_2},\ldots,\overrightarrow{d_n}\}$, n is the total number of descriptors and $\overrightarrow{d_l}(l=1,\ldots,n)$ is the vector of the l-th descriptor; $T=\{t_1,t_2,\ldots,t_n\}$ and $t_l(l=1,\ldots,n)$ is the data type of the l-th descriptor; $R=\{r_1,r_2,\ldots,r_n\}$ and $r_l=<r_l^{\min},r_l^{\max}>$ is the experience value range of the l-th descriptor; $r_l^{\min}$ and $r_l^{\max}$ are the minimum and maximum values of empirical values respectively; $U=\{u_1,u_2,\ldots,u_n\}$ and $u_l(l=1,\ldots,n)$ is the unit of the l-th descriptor. Let m be the total number of samples, given any descriptor $d_l^j(j=1,\ldots,m)$, Eq. (1) and (2) can be used to determine whether it is a potential anomalous data point:

$$\mathrm{value}\left(d_l^j\right)=\begin{cases}0, & \text{if } \left\{\mathrm{range}\left(d_l^j\right)\in\left(r_l^{\min},r_l^{\max}\right)\right\}\\ 1, & \text{otherwise}\end{cases} \tag{1}$$

$$\mathrm{grammer}\left(d_l^j\right)=\begin{cases}0, & \text{if } \left\{\mathrm{type}\left(d_l^j\right)=t_l\right\}\&\left\{\mathrm{unit}\left(d_l^j\right)=u_l\right\}\\ 1, & \text{otherwise}\end{cases} \tag{2}$$

where type$(\cdot)$, range$(\cdot)$, and unit$(\cdot)$ indicates the data type, value range and unit of any data point, respectively. 1 and 0 indicate that

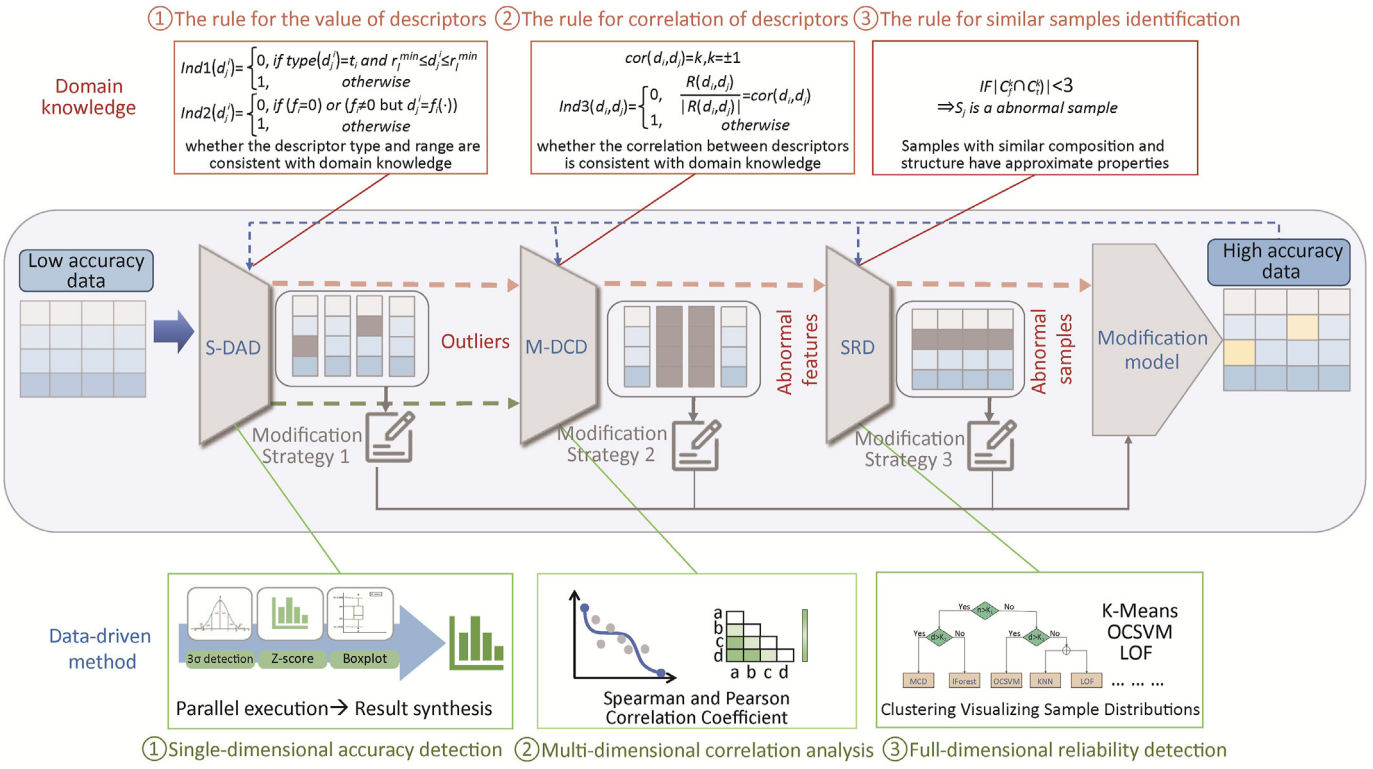

Fig. 1. Workflow of DKA-DAD. There are three models in the middle, and the anomalies are filtered out hierarchically through the three models. The upper and bottom are specific methods. The blue dashed line represents the detection-governance feedback loop.

d_l^j is or is not a potential anomaly data point. If d_l^j exceeds the empirical value range $[r_l^{\min}, r_l^{\max}]$ of the descriptor it belongs to, or the type(d_l^j) does not match the empirical value type t_l of the corresponding descriptor, or the unit(d_l^j) does not match the empirical unit u_l, it indicates that the d_l^j is anomaly and should be further analyzed.

Note that the detection of units typically relates to the value range. For example, "°C" and "K" are both for the unit of temperature, while the difference in value between these two units is 272.15 (namely K = °C + 272.15). Therefore, the materials domain knowledge needs to participate in the estimation of unifying value units according to the value range, and then, data anomaly detection methods can be used to filter the sample without the target unit.

Commonly used data anomaly detection methods based on statistical analysis include 3σ detection, *Z*-Score, boxplot, Grubbs' Test, *etc.* These individual methods primarily analyze data *via* data distribution. However, materials data are often collected from multiple sources and the distribution characteristics of such data are complex and diverse. Therefore, as shown in Eq. (3), the ensemble strategy is employed to ensure that data detection is accurate and reliable. Concisely, three commonly used methods are executed simultaneously, including 3σ detection, *Z*-Score, and boxplot, to detect the value of descriptors from the perspective of data distribution, and then combine their results through majority voting.

$$\begin{aligned}\text{SingleAnomalies} = &((3\sigma \cap \text{zscore}) \text{or} (3\sigma \cap \text{boxplot}) \text{or} \\ &(\text{zscore} \cap \text{boxplot})) \cap \text{rule 1}\end{aligned} \tag{3}$$

where rule 1 indicates the anomaly detection result of the rule for value of descriptors 1. If a sample point is detected as being an anomaly by any two of the 3σ detection, *Z*-Score and boxplot methods and violates Rule 1, it is identified as a single dimensional anomaly sample.

2.2. Anomaly detection based on the rule of descriptor cross-correlation

While single-descriptor accuracy detection provides a certain level of assurance for individual descriptors, it does not guarantee accuracy between descriptors. Correlation between descriptors can influence the results of feature selection, which in turn affects the model prediction performance.

According to the domain knowledge related to descriptors, some descriptors may share similar or same physical meaning and exhibit identical influencing mechanisms on material properties, *i.e.*, there is a certain correlation between these descriptors. To this end, *the rule for correlation of descriptors* (Hereinafter referred to as Rule 2) can be defined as shown in Eqs. (4) and (5), including **descriptor qualitative correlation rules** (QLR) and **descriptor quantitative correlation rules** (QTR). QLR refers to the correlation between descriptors or between descriptors and material properties that can be obtained directly from the literature; QTR refers to the correlation between descriptors or between descriptors and material properties obtained by extracting and processing calculation formulas in literature. Both QLR and QTR are discussed in detail below.

The rule for correlation of descriptors: Given binary group $\langle D, F\rangle$, where $D = \{\overrightarrow{d_1}, \overrightarrow{d_2}, \ldots, \overrightarrow{d_n}\}$, n is the total number of descriptors and $\overrightarrow{d_l}(l = 1, \ldots, n)$ is the vector of the l-th descriptor; $C = \{c_1, c_2, \ldots, c_k\}, c_p = < d_i, d_j, cor(d_i, d_j) > (p = 1, \ldots, k)$ indicates the rule for correlation of descriptors derived from material domain knowledge, $\text{cor}(d_i, d_j)$ indicates the correlation (positive correlation or negative correlation) between descriptors d_i and d_j , and its

Table 1
The meanings, types, ranges, and sources of all descriptors in NASICON.

No.	Descriptors	Description	Type	range	Source
1	Occu_6b	Occupancy of Na in 6b site	Float	[0,1]	CIF
2	Occu_18e	Occupancy of Na in 18e site	Float	[0,1]	CIF
3	Occu_36f	Occupancy of Na in 36f site	Float	[0,1]	CIF
4	*C*_Na	Na^+ concentration	Float	$(0, +\infty)$	Formula
5	Occu_M1	Occupancy of element M1	Float	[0,1]	CIF
6	Occu_M2	Occupancy of element M2	Float	[0,1]	CIF
7	EN_M1	Electronegativity of element M1	Float	$(0, +\infty)$	Pauling electronegativity meter
8	EN_M2	Electronegativity of element M2	Float	$[0, +\infty]$	Pauling electronegativity meter
9	EN_avg_M	Average effective electronegativity of M site	Float	$(0, +\infty)$	Formula
10	Radius_M1	Ionic radius of element M1	Float	$(0, +\infty)$	Shannon radius table
11	Radius_M2	Ionic radius of element M2	Float	$[0, +\infty]$	Shannon radius table
12	Radius_avg_M	Average effective ionic radius of M site	Float	$(0, +\infty)$	Formula
13	Valence_M1	Valence of element M1	Int	$(0, +\infty)$	CIF
14	Valence_M2	Valence of element M2	Int	$[0, +\infty]$	CIF
15	Valence_avg_M	Average effective ionic valence of M site	Float	$(0, +\infty)$	Formula
16	Occu_X1	Occupancy of element X1	Float	[0,1]	CIF
17	Occu_X2	Occupancy of element X2	Float	[0,1]	CIF
18	EN_X1	Electronegativity of element X1	Float	$(0, +\infty)$	CIF
19	EN_X2	Electronegativity of element X2	Float	$(0, +\infty)$	CIF
20	EN_avg_X	Average effective electronegativity of X site	Float	$(0, +\infty)$	Formula
21	Radius_X1	Ionic radius of element X1	Float	$(0, +\infty)$	Shannon radius table
22	Radius_X2	Ionic radius of element X2	Float	$[0, +\infty]$	Shannon radius table
23	Radius_avg_X	Average effective ionic radius of X site	Float	$(0, +\infty)$	Formula
24	Valence_X1	Valence of element X1	Int	$(0, +\infty)$	CIF file
25	Valence_X2	Valence of element X2	Int	$[0, +\infty]$	CIF file
26	Valence_avg_X	Average effective ionic valence of X site	Float	$(0, +\infty)$	Formula
27	*a*	Lattice parameter	Float	$(0, +\infty)$	CIF file
28	*c*	Lattice parameter	Float	$(0, +\infty)$	CIF file
29	V_{cell}	Lattice parameter	Float	$(0, +\infty)$	Formula
30	V_MO_6	Volume of MO_6 polyhedron	Float	$(0, +\infty)$	VESTA file
31	V_XO_4	Volume of XO_4 polyhedron	Float	$(0, +\infty)$	VESTA file
32	$V\_Na_1O_6$	Volume of Na_1O_6 polyhedron	Float	$(0, +\infty)$	VESTA file
33	$V\_Na_2O_8$	Volume of Na_2O_8 polyhedron	Float	$(0, +\infty)$	VESTA file
34	$V\_Na_3O_5$	Volume of Na_3O_5 polyhedron	Float	$(0, +\infty)$	VESTA file
35	BT1	Bottleneck	Float	$(0, +\infty)$	Formula
36	BT2	Bottleneck	Float	$(0, +\infty)$	Formula
37	min_BT	The minimum of BT2 and BT1	Float	$(0, +\infty)$	VESTA file
38	RT	Radius of largest sphere probe that can freely pass through the void space packed by framework ions	Float	$(0, +\infty)$	Geometry-based Ion-transport Analysis Library CAVD
39	EP_6b	Configurational entropy of Na in 6b site	Float	$[0, +\infty]$	Formula
40	EP_18e	Configurational entropy of Na in 18e site	Float	$[0, +\infty]$	Formula
41	EP_36f	Configurational entropy of Na in 36f site	Float		Formula

Table 1 (*continued*)

No.	Descriptors	Description	Type	range	Source
				[0, + ∞]	
42	EP_Na	Configurational entropy of Na	Float	[0, + ∞]	Formula
43	EP_M	Configurational entropy of cationic in M site	Float	[0, + ∞]	Formula
44	EP_X	Configurational entropy of cationic in X site	Float	[0, + ∞]	Formula
45	*T*	Temperature	Float	(0, + ∞)	Reference

formal expression is shown in Eqs. (4) and (5).

$$\mathrm{cor}(d_i, d_j) = k, k = \pm 1 \tag{4}$$

$$\mathrm{MultiDim}(d_i, d_j) = \begin{cases} 0, & \dfrac{R(d_i,\ d_j)}{|R(d_i,\ d_j)|} = \mathrm{cor}(d_i, d_j) \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where 1 and 0 indicate that d_i, d_j is or is not a potential anomaly descriptor respectively, $R(\cdot)$ denotes the correlation coefficient between any two descriptors d_i and d_j obtained by data-driven correlation analysis. k is equal to 1 or -1 which indicates a positive or negative correlation between descriptors d_i and d_j, respectively. When k is equal to 1 (-1) but $R(\cdot)$ is less (greater) than 0, it indicates that the correlation between descriptors obtained by the data-driven correlation analysis method is not consistent with the domain knowledge, requiring the acquisition and calculation of descriptors to be redefined to correct the dataset.

- **Descriptor qualitative correlation rules (QLR)**

We here take the NASICON-type solid electrolyte materials as an example, the symbolic representation for from materials domain knowledge are as follows.

In NASICON-type solid electrolytes of space group $R\ \bar{3}$ c, both lattice constants (a, b, c) and the volume of the unit cell can characterize the unit cell size. The lattice constant a is equal to the lattice constant b, and has the following relationship with the volume of the unit cell:

$$V_{\mathrm{cell}} = \frac{\sqrt{3}}{2} a^2 c \tag{6}$$

where V_{cell} is the volume of the unit cell, a and c are the lattice constant.

Therefore, V_{cell} have an obvious positive correlation with a and c, which can be expressed as Eq. (7) and (8):

$$a \propto V_{\mathrm{cell}} \tag{7}$$

$$c \propto V_{\mathrm{cell}} \tag{8}$$

where V_{cell} is the volume of the unit cell, a and c are the lattice constant.

The larger the bottleneck, the easier the migration of ions in the ion transport channel, and the smaller the activation energy of ion migration in the compound [21,22]. As the minimum bottleneck in the ion migration channel, the conduction threshold min_BT also has a negative correlation with the activation energy, which can be expressed as Eqs. (9)–(11):

$$\mathrm{BT1} \propto -E_{\mathrm{a}} \tag{9}$$

$$\mathrm{BT2} \propto -E_{\mathrm{a}} \tag{10}$$

$$\mathrm{min\_BT} \propto -E_{\mathrm{a}} \tag{11}$$

where BT1 and BT2 represent two bottlenecks in NASICON respectively; min_BT is the minimum value between BT1 and BT2; E_{a} is the activation energy, which is usually used as a material property in NASICON-type solid electrolyte materials. According to Eqs. (9)–(11), BT1, BT2 and Min_BT are all negatively correlated with E_{a}, so three QLRs are obtained.

- **Descriptor quantitative correlation rules** (QTR)

In the NASICON solid electrolyte material data, the meaning, data type, value range and data source of the descriptors, shown in Table 1, can be obtained according to the relevant literature. The data type of the descriptor "Radius_avg_M″ is floating-point, with values ranging from $(0, +\infty)$. Meanwhile, the "Radius_avg_M″ is related to the occupancy and radius of the elements, and the correlation between them is calculated as shown in Eq. (13).

$$\mathrm{adius\_avg\_M} = (\mathrm{Occu\_M1})\mathrm{Radius\_M1} + (\mathrm{Occu\_M2})\mathrm{Radius\_M2} \tag{13}$$

where ion radius come from Shannon radius table [23], OccuM1 and OccuM2 represent the M1 and M2 element occupancy; Radius_M1 and Radius_M2 represent the M1 and M2 element radius. Note that if data type of "Radius_avg_M″ is not floating-point, the range of values is not $(0, +\infty)$, or the value of "Radius_avg_M″ do not match the values calculated in Eq. (7), then the point is a potentially abnormal data point.

The construction of high-accuracy material datasets requires the reduction, or ideally the elimination, of correlations between descriptors. Several methods are available to detect correlations between multi-dimensional data, such as Pearson Correlation Coefficient (PCC) [24] and Spearman Correlation Coefficient (SCC) [25]. PCC has advantages in measuring the correlation between descriptors and material properties. SCC is commonly used to measure the correlation between descriptors because of its insensitivity to data errors and extreme values.

In summary, through embedding into the multi-descriptor correlation detection method based on correlation rules, the abnormal values of descriptors and highly correlated descriptors can be identified accurately, which are marked and flagged as potential anomalies. Then, the experts need to interpose further analysis for whether to remove highly correlated descriptors (*i.e.*, execution of feature engineering) or not.

2.3. Anomalies sample detection based on the rule of similarity identification

Following the single-descriptor accuracy detection and multi-descriptors correlation detection, the accuracy of materials data in each descriptor is expected to increase. However, this does not guarantee the accuracy of every sample associated with multiple descriptors.

The properties of materials are influenced by various factors such as composition, structure, experimental conditions and environment. Materials with similar compositions and structure are expected to exhibit similar properties. In this section, similar sample identification strategy (Hereinafter referred to as Rule 3) is defined as shown in Eq. (13). The descriptors (*i.e.*, features) characterizing the structure and composition of materials and material properties are clustered separately using *K*-Means [26] to identify anomalous samples. When the clustering results by features and by material properties belong to the same or adjacent clusters, these samples are considered to be correct; When the clustering results by features do not belong to the clusters adjacent to the material properties, such samples are flagged as anomalous. In summary, the samples with similar composition and structure but differing material properties are considered as potential anomalous samples.

The rule for similar sample identification: Given group $\langle F, T, C_F, C_T\rangle$, F is the set of values for all samples with a particular descriptor, T is the material property data for all the samples. $C_F = \{c_f^1, c_f^2, \ldots, c_f^k\}$ and $C_T = \{c_t^1, c_t^2, \ldots, c_t^k\}$ respectively denote the clustering results on feature F and material property T, k is the number of clusters. For a given sample $S_j (j = 1, \ldots, m)$, Eq. (13) can be used to determine whether they are potential anomalous samples.

$$\mathrm{Full}(S_j) = \begin{cases} 1, & \left|c_f^k \cap c_t^k\right| < 2 \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

where $|\cdot|$ is the number of elements in the set, c_f^k and c_t^k respectively denote the set of clustering results of any sample S_j on F and T. For the set c_f^k and c_t^k to which S_j belongs, if the number of samples in the intersection $c_f^k \cap c_t^k$ is less than 2, it means that the samples that are clustered with S_j in terms of features are not clustered in terms of material properties or even far away from each other. Therefore, S_j is a potentially anomalous sample.

Existing methods for anomaly sample detection rely on data analysis, such as distance-based approaches like k-nearest neighbors (KNN) [27], partitioning-based methods like Isolation Forest (IForest) [28], density-based techniques like Local Outlier Factor (LOF) [29], support vector machines for one-class classification based on hyperplane distance (OCSVM) [30], and methods involving covariance matrix estimation such as Minimum Covariance Determinant (MCD) [31]. Considering the varied perspectives, applicability, and pros and cons of these anomaly detection methods (see Supplementary Table S1), we propose a strategy for selecting the most suitable method for comprehensive sample reliability assessment. When dealing with substantial datasets (the number of samples $\boldsymbol{n}$ greater than $\boldsymbol{K_1}$) and high dimensions (the number of feature $\boldsymbol{d}$ greater than $\boldsymbol{K_2}$), MCD is the preferred choice. In other cases, Isolation Forest is recommended. For smaller datasets ($\boldsymbol{n}$ less than or equal to $\boldsymbol{K_1}$) with high dimensions ($\boldsymbol{d}$ greater than $\boldsymbol{K_2}$), One-Class SVM (OCSVM) is recommended. Otherwise, KNN or LOF should be considered.

2.4. Data accuracy improvement by modification model

After applying the three detection models discussed above, anomalies within individual descriptors, between descriptors, and among samples are detected independently. Concisely, a modification strategy is generated for each stage based on an analysis of the detection results with the collaboration of domain knowledge. The execution process of modification model is illustrated in Fig. 2. Initially, domain experts thoroughly evaluate all modification strategies to determine whether each identified anomaly point is genuine and analyze its underlying cause, thus establishing a comprehensive modification strategy for handling anomalies, such as deletion, correction, or rechecking. Subsequently, the original dataset is modified according to this strategy, resulting in a revised dataset. Finally, ML models are constructed using both revised dataset and original dataset. The effectiveness of anomaly detection is assessed by evaluating the performance of these models.

3. Experimental section

3.1. Datasets

3.1.1. Materials datasets

All datasets employed in the experiments are derived from published materials research literature spanning from 2003 to 2023. These datasets encompass a diverse range of material types, including inorganic nonmetallic materials, metallic materials, polymer materials, composite materials, and others. Detailed statistical information regarding the specific datasets is shown in Fig. 3. Fig. 3a and b displays the number of samples and features for each of the 60 datasets, respectively. Note that most datasets contain fewer than 750 samples and fewer than 50 features. Fig. 3c summarizes the distribution of datasets across different material types, with inorganic non-metallic materials being the most prevalent category. Fig. 3d illustrates the prediction performance of the datasets extracted from the source literature for various material property prediction tasks. Notably, two datasets employ custom evaluation metrics, making them incomparable with others, and three datasets do not provide information about model performance. Further details are available in the Supplementary Information (SI).

3.1.2. Simulated datasets

To validate the effectiveness of individual detection models and the benefits of embedding domain knowledge, additional 180 simulated datasets are generated based on 60 real datasets. Fig. 4 illustrates the process of generating three simulated datasets from one real dataset.

Anomalies are injected into the real dataset based on the principles of single-dimensional, multi-dimensional, and full-dimensional anomaly detection, and then appended to form simulated datasets.

Individual descriptor anomalies: Due to experimental equipment errors or computational software defects, materials datasets may contain certain noise during collection, which leads to inadequate model performance when used directly for machine learning modeling. In this dimension, m columns are randomly selected for uniform sampling $\mathrm{Unif}(\min(x^m) + p, \max(x^m) + p)$ to generate uniform noise.

Multiple descriptor anomalies: Materials datasets may include uncorrelated descriptors, the existence of which is superfluous and affects the predictive accuracy of machine learning models. In this dimension, we randomly select k descriptors for uniform sampling $\mathrm{Unif}(\min(x^k), \max(x^k))$ to generate uniform noise descriptors. The number of abnormal descriptors added does not exceed 50% of the total input features. In this study, we set k as 2. That is, two columns (descriptors) of each dataset are selected randomly and then execute noise addition.

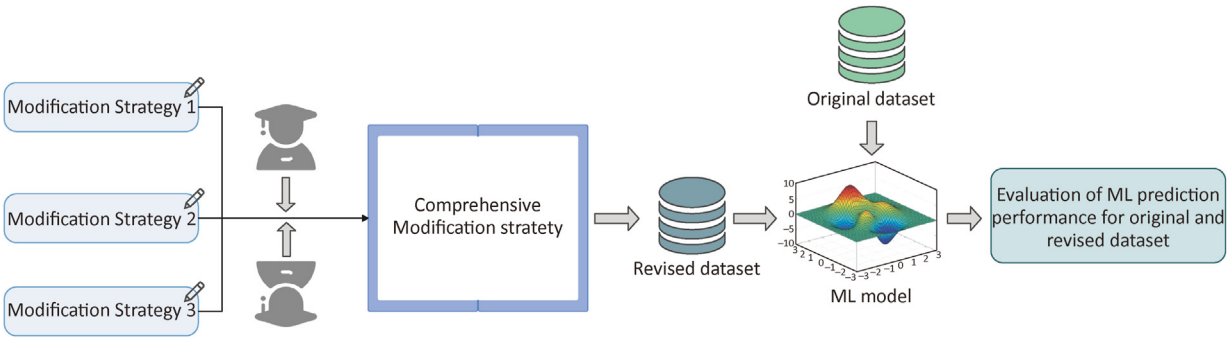

Fig. 2. Processes of data accuracy improvement.

Sample anomalies: Due to recording errors or other factors, materials datasets may contain duplicate or abnormal samples, which will affect the subsequent modeling process of machine learning. In this stage, anomaly samples are generated based on a uniform distribution $\mathrm{Unif}\left(\alpha*\min\left(x^f\right), \alpha*\max\left(x^f\right)\right)$ for each descriptor f. Specifically, for the feature and materials property, α is set to 1 or 10, respectively, aiming to simulate the distinction between the feature and the materials property.

3.2. Experimental setups

3.2.1. Parameter setup of DKA-DAD

In S – DAD, the *Z*-score method based on normal distribution assumptions and boxplot were adopted for anomaly detection. In the boxplot, the data interval IQR of the upper and lower edges is set to 1.5, which means that the data fractions exceeding 5% and 95% of the data interval are identified as anomalies. In M – DCD, PCC and SCC are used to evaluate the correlation between descriptors. In SRD, the sample size and dimension thresholds K_1 and K_2 are set to 250 and 25, respectively. The number of *K*-means clusters is set to 8.

3.2.2. Candidate ML models

Due to the unsupervised nature of the above anomaly detection algorithm, we cannot set ground truth to evaluate whether outliers should be removed. Thus, in this paper, six ML models that are commonly used in the materials field are selected as candidate models, including Least Absolute Shrinkage and Selection Operator (LASSO) [32], Gaussian Process Regression (GPR) [33], Ridge Regression (Ridge) [34], Support Vector Regression (SVR) [35], K-Nearest Neighbor Regression (KNN) and Random Forest (RF) [36]. ML models are built on the original and revised datasets, respectively. The model performance is evaluated by 10-fold cross-validation.

3.2.3. Evaluation metrics

Anomaly detection is inherently an imbalanced data classification problem, where the amount of anomalous data is significantly lower than that of normal data. Consequently, relying solely on accuracy to evaluate the effectiveness of anomaly detection lacks meaningful reference. Therefore, Recall, Precision, and F1-score are selected as evaluate metric to evaluate the performance of anomaly detection in this paper. More details can be seen in Section S2.2 of SI.

Additionally, to evaluate the performance of ML models, we select the Root Mean Square Error (RMSE), the Mean Absolute Percent Error (MAPE), and *R*-square (R^2) as evaluation metrics. All the algorithms are implemented by the Python-based scikit-learn toolkit.

3.3. Validation on simulated datasets

To validate the effectiveness and necessity of DKA-DAD and materials domain knowledge embedding into each detection model, sufficient verification experiments based on the different simulated datasets are designed and executed.

3.3.1. Single-descriptor accuracy detection

According to Eq. (2), materials domain knowledge such as descriptor value range and data type can be obtained. To compare the disparities between purely data-driven data-accuracy detection method and domain knowledge-assisted data accuracy detection method, four purely data-driven methods are selected, *i.e.*, Boxplot (M1), *Z*-score (M2), 3δ detection (M3), Boxplot + *Z*-score+3δ detection (M4), and one domain knowledge-assisted method (M4 combined with Rule 1, S-DAD). The experiments are conducted on single-dimensional simulated datasets. The detection results of 60 datasets on five methods are shown in Fig. 5 and Fig. 6. Fig. 5 demonstrates the 60 simulated dataset results of S – DAD. As shown in Fig. 5, the line graph gradually stabilizes from M1 to M4, indicating that the combination of multiple methods comprehensively detects anomalies within the dataset compared to individual methods. Compared to M4, S-DAD towards smoother, and the prediction accuracy of the F1-score exceeds 90%. This suggests that the incorporation of domain knowledge can rectify errors or undetected anomalies in purely data-driven analysis, thereby improving the accuracy of anomaly detection by identifying real anomalies and reducing false anomalies.

Overall, from M1 to S-DAD, the F1-score exhibited an upward trend on each dataset, indicating the robustness of S – DAD in detecting anomalies across datasets of varying sizes. Fig. 6 presents the average detection results for 60 simulated datasets. M1 exhibits the highest Recall, followed by S – DAD, suggesting a slightly higher detection rate of anomalies compared to S – DAD. Table 2 represents the statistical information of different prediction metrics, which can be seen that S – DAD achieve significant differences compared with most of metrics compared with other methods. Note that M1 performs poorly in terms of Precision and F1-score. Additionally, in most cases, the model accuracy of M4 surpasses that of M1, M2, and M3, although it is lower than S – DAD. This suggests that the combination of M1, M2 and M3 compensates for the limitations of individual methods and enhances detection accuracy. The standard deviation provides insight into dataset stability. The standard deviation of S – DAD on three metrics is the smallest, indicating its stable performance across different datasets and its reliable anomaly detection capability. This reinforces the importance of incorporating domain knowledge to correct the inaccuracies prediction results, underscoring its value in strengthening data-driven methods for outlier detection. Consequently, S –

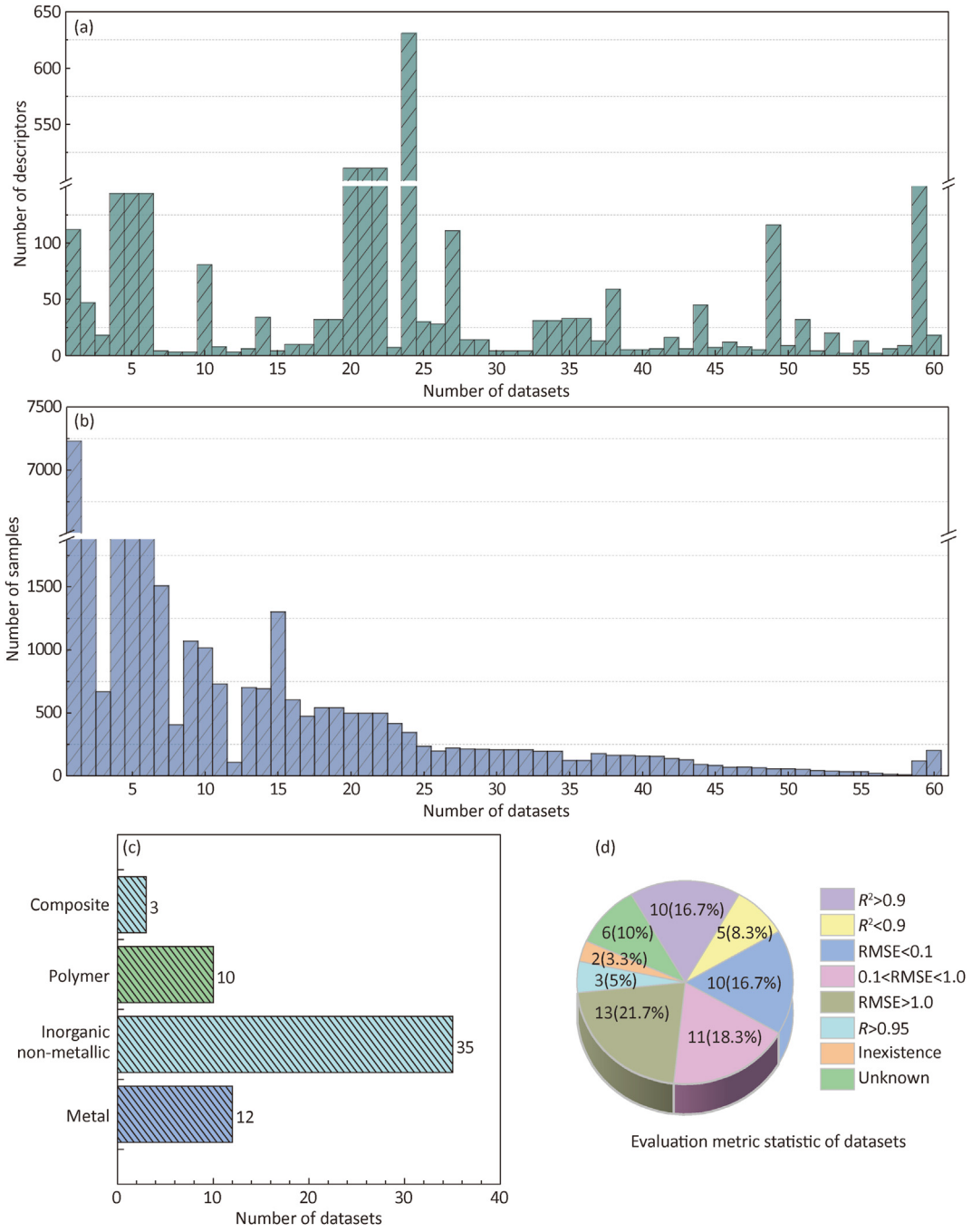

Fig. 3. Statistics on 60 datasets. (a) and (b) represent the distribution of the number of samples and features respectively; (c) represents the number of datasets contained for each material type (d) represents the prediction performance of every material property prediction model (all performance data from source literature).

DAD exhibits superior performance compared to the other four methods.

3.3.2. Multi-descriptor correlation detection model

In the multi-descriptor correlation detection, the success rate of anomaly descriptor detection, denoted as "***suc***", is defined in Eq. (14). The detection results for all datasets are summarized in Table 3. It is evident that most unrelated descriptors in the datasets can be successfully identified. However, on the four datasets (D12, D56, D57, D58), M-DCD fails to accurately identify the noise descriptors (***suc*** $= 0$). Further analysis reveals that these 4 datasets have characteristics of an extremely small or large number of descriptors. When the number of features is too small, noise descriptors may be erroneously classified as normal descriptors. Conversely, when the number of features is excessively large, the presence of sparse descriptors can interfere with the results of correlation analysis. Thus, it is crucial to apply feature quantity governance or sample quantity governance prior to employing machine learning or deep learning approaches [12].

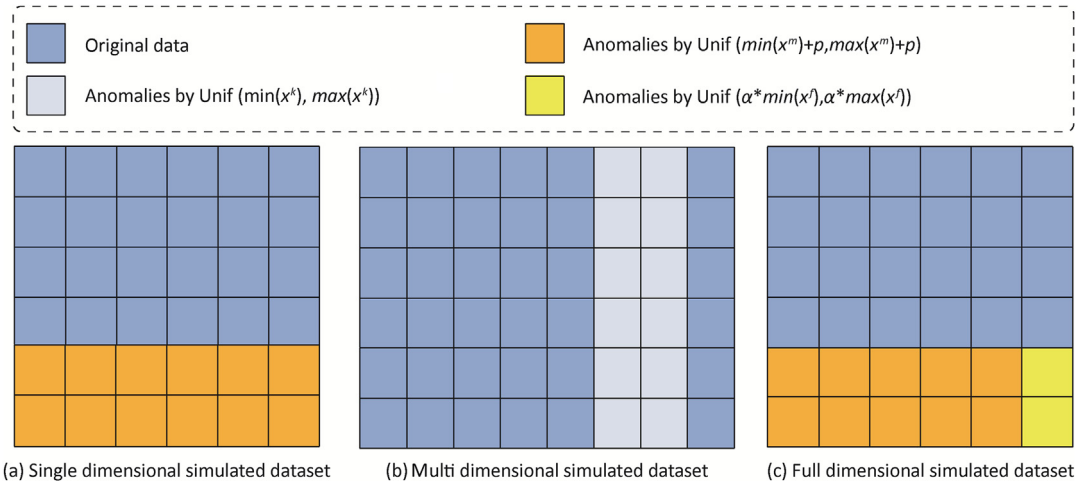

Fig. 4. Procedure for abnormal injection. (a) abnormal injection for single-dimensional type, (b) abnormal injection for multi-dimensional type, (c) abnormal injection for full-dimensional type.

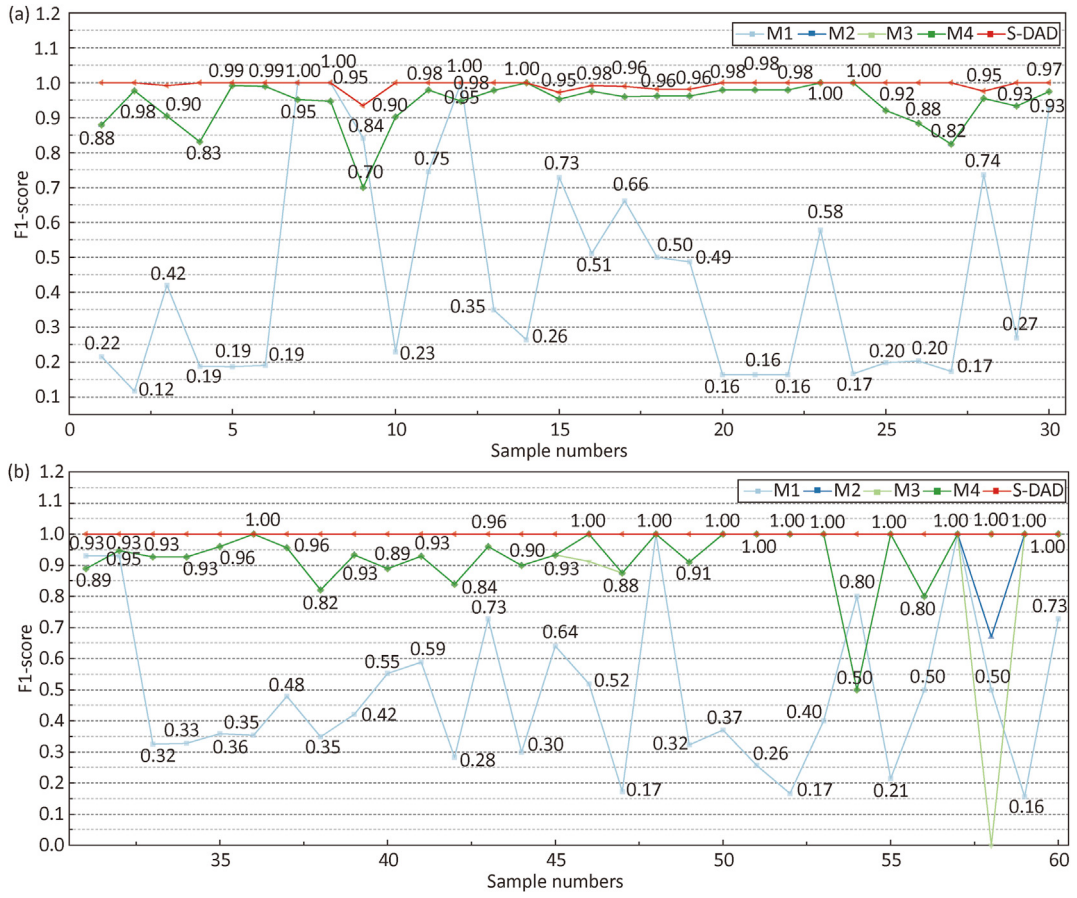

Fig. 5. F1-score of simulated datasets detected and revised by various single-descriptor accuracy detection methods. (a) Dataset 1–30, (b) Dataset 31–60. "Light blue" represents M1 (Boxplot), "Blue" represents M2 (Z-score), "Light green" represents M3 (3δ detection), "Green" represents M4 (the method of combination of Boxplot, Z-score and 3δ detection), and "Red" represents the methods of the combination of M4 and Rule 1(S-DAD).

$$\boldsymbol{suc} = \frac{\text{the number of uncorrelated descriptors detected by PCC or SCC}}{\text{the total number of uncorrelated descriptors}} \tag{14}$$

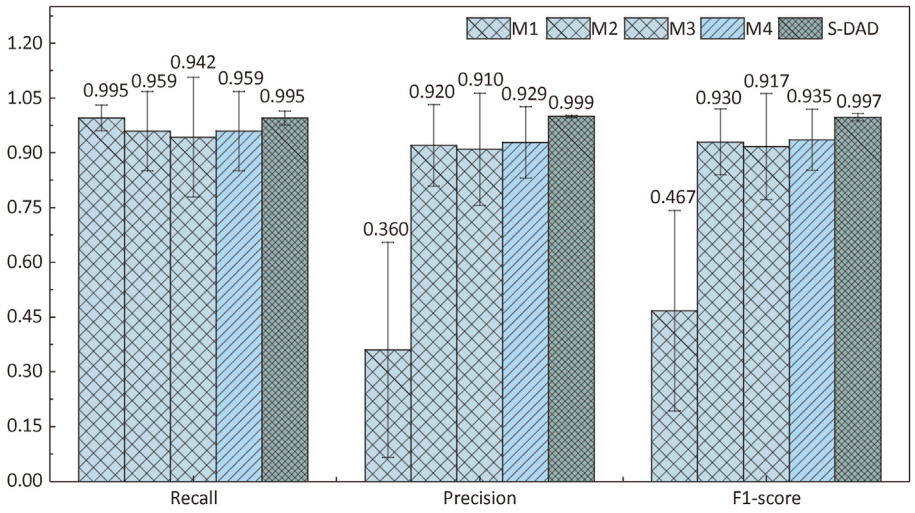

Fig. 6. Evaluation indicators of 5 methods on 60 simulated datasets. M1 represents Boxplot, M2 represents *Z*-score, M3 represents 3δ detection, M4 represents the method of combination of Boxplot, *Z*-score and 3δ detection, and S-DAD represents the methods of the combination of M4 and Rule 1.

Table 2
Statistical Information for different metrics between S-DAD and other data-driven methods. p-value$_{Rec}$, p-value$_{Pred}$, p-value$_{F1}$ represents the p value of Recall, Precision and F1-score of this method compared with S-DAD, respectively. "**CI**" represents confidence interval.

Method	p-value$_{Rec}$	p-value$_{Prec}$	p-value$_{F1}$	95% CI$_{Rec}$	95% CI$_{Pred}$	95% CI$_{F1}$
M1	0.2491	$4.816e^{-24}$	$1.868e^{-21}$	[−0.0734, 0.0294]	[0.0000, 0.9132]	[0.0000, 0.8402]
M2	$1.365e^{-3}$	$1.915e^{-8}$	$9.296e^{-8}$	[−0.0086, 0.3222]	[0.0000, 0.2660]	[0.0012, 0.2182]
M3	$1.104e^{-2}$	$1.638e^{-5}$	$5.902e^{-5}$	[−0.0086, 0.3222]	[0.0000, 0.2660]	[0.0012, 0.2182]
M4	$1.109e^{-2}$	$4.223e^{-10}$	$6.777e^{-10}$	[−0.0086, 0.1810]	[0.0000, 0.2228]	[0.0000, 0.1542]

Table 3
Success rate of uncorrelated descriptors detection.

No.	PCC	SCC	No.	PCC	SCC	No.	PCC	SCC
D1	1	1	D21	1	1	D41	0.5	1
D2	0.5	1	D22	1	1	D42	1	1
D3	1	1	D23	1	0.5	D43	0.5	1
D4	1	1	D24	1	1	D44	0.5	1
D5	0.5	1	D25	1	1	D45	1	1
D6	1	1	D26	1	1	D46	1	1
D7	1	1	D27	1	1	D47	1	1
D8	1	0.5	D28	1	0.5	D48	1	1
D9	1	1	D29	1	1	D49	1	1
D10	1	1	D30	1	1	D50	1	1
D11	1	1	D31	1	1	D51	0.5	1
D12	0	0	D32	1	0.5	D52	1	1
D13	1	1	D33	1	1	D53	1	1
D14	1	0.5	D34	1	0.5	D54	1	1
D15	1	1	D35	1	1	D55	1	1
D16	0.5	1	D36	1	1	D56	1	0
D17	1	0.5	D37	1	1	D57	0	0
D18	1	1	D38	1	0.5	D58	1	0
D19	1	1	D39	1	1	D59	1	0.5
D20	1	1	D40	1	0.5	D60	1	1

3.3.3. Sample reliability improvement detection model

Similar to Section 3.3.1, clustering is applied on both the features and decision attributes according to Rule 3 and Eq. (13). To compare the purely data-driven method and the domain knowledge-assisted data-driven method, two experiments are conducted: the pure data-driven abnormal sample detection method (P1), and the clustering analysis incorporating domain knowledge intersected with data-driven anomaly detection method (SRD).

The anomaly detection results of the 60 datasets are illustrated in Fig. 7 and Fig. 8. Fig. 7 shows that SRD achieves a significantly higher F1-score than P1 in 93% (56 out of 60) of the datasets, of which statistical information of accuracy metrics between P1 and SRD is that the p-values of Recall, Precision and F1-score are $5.869e^{-13}$, $2.187e^{-11}$ and $1.045e^{-7}$, and their 95% CIs equal [−0.2565, 0], [0, 0.9269] and [−0.0424, 0.7474]. This indicates that integrating domain knowledge can further compensate for the limitations of data-driven methods, thus identifying more anomalous points. Fig. 8 shows the average detection results of 60 simulated datasets. It can be observed that P1 exhibits a much higher Recall than SRD; however, its Precision and F1-score are relatively low. In addition, the standard deviation of P1 is more volatile compared to SRD. Based on the evaluation metrics and the average performance across all 60 datasets, SRD exhibits the best performance, with relatively stable performance across all evaluation metrics.

In conclusion, based on the verification experiment conducted on 180 synthetic datasets using the proposed detection models with three different dimensions, it is evident that the combination of multiple methods surpasses the performance of individual methods, and the integration of domain knowledge can indeed enhance the effectiveness of purely data-driven methods. Specifically, in the processes of single-descriptor and sample anomaly detection, domain knowledge can effectively assist the data-driven detection methods in identifying genuine abnormal samples, correcting misidentified samples, and produce higher quality data for further ML modelling.

3.4. Application of material datasets

In this section, DKA-DAD is employed on 60 material datasets. Fig. 9 shows the average prediction accuracy (R^2) of 6 ML models in the original and revised dataset with the statistical information that

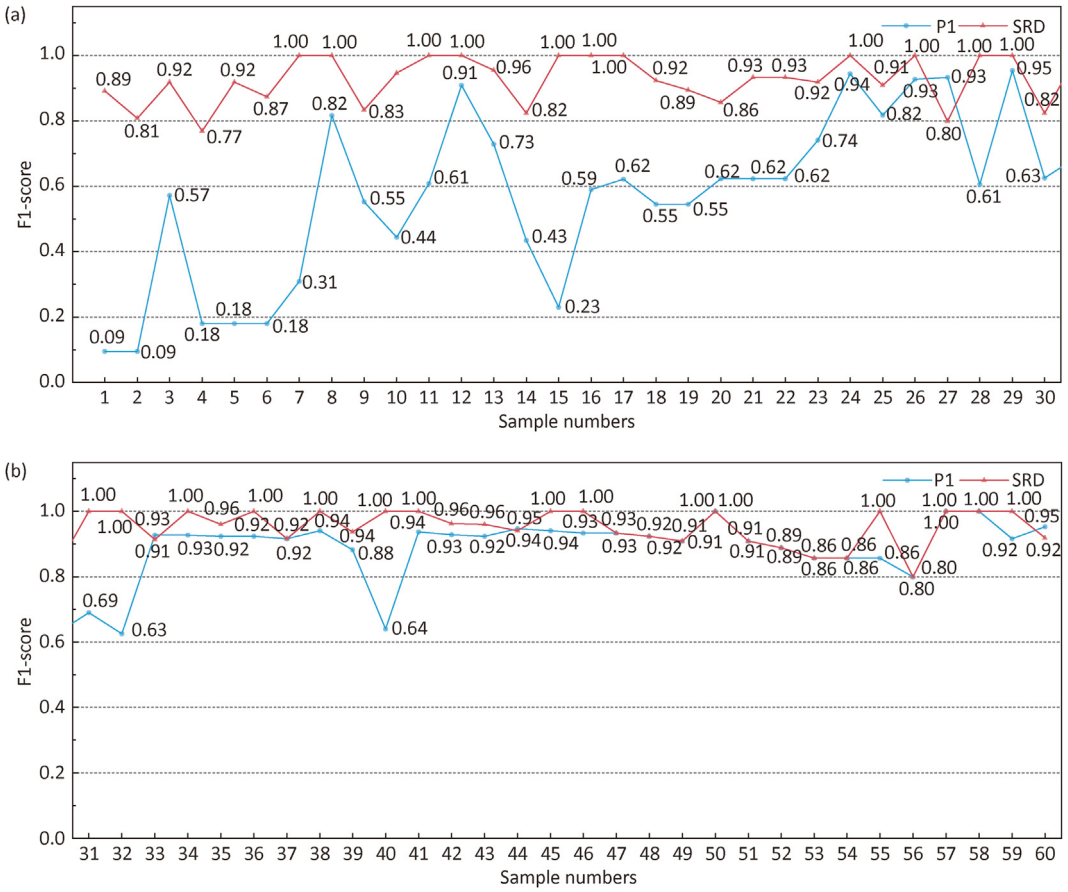

Fig. 7. F1-score of SRD on full-dimensional simulated datasets. (a) Dataset 1–30, (b) Dataset 31–60. "Sample Number" represents the number of datasets. Blue represents the datasets revised by the pure data-driven abnormal sample detection method (P1). Red represents the clustering analysis incorporating domain knowledge intersected with the data-driven anomaly detection method (SRD).

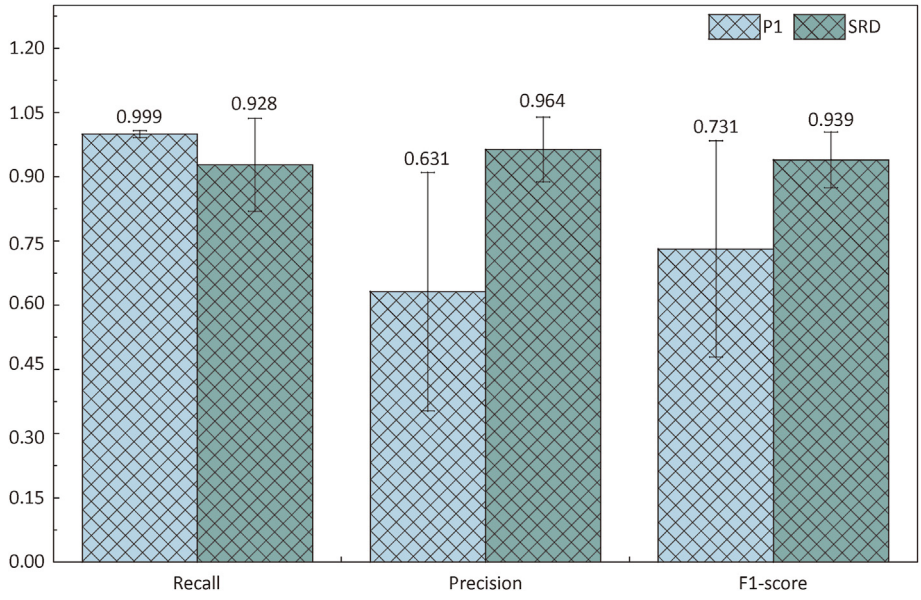

Fig. 8. Comparison of various evaluation indicators between P1 and SRD.

its p-value and 95% CI are $5.534e^{-4}$ and [−0.2373, 0.0446], respectively. As shown, experimental results indicate that 12 of these datasets exhibited potential anomalies and require correction, and the prediction accuracy of ML models on the revised datasets are improved to varying degrees compared to the original datasets, with an average increase of 9.6% across the 12 datasets. In

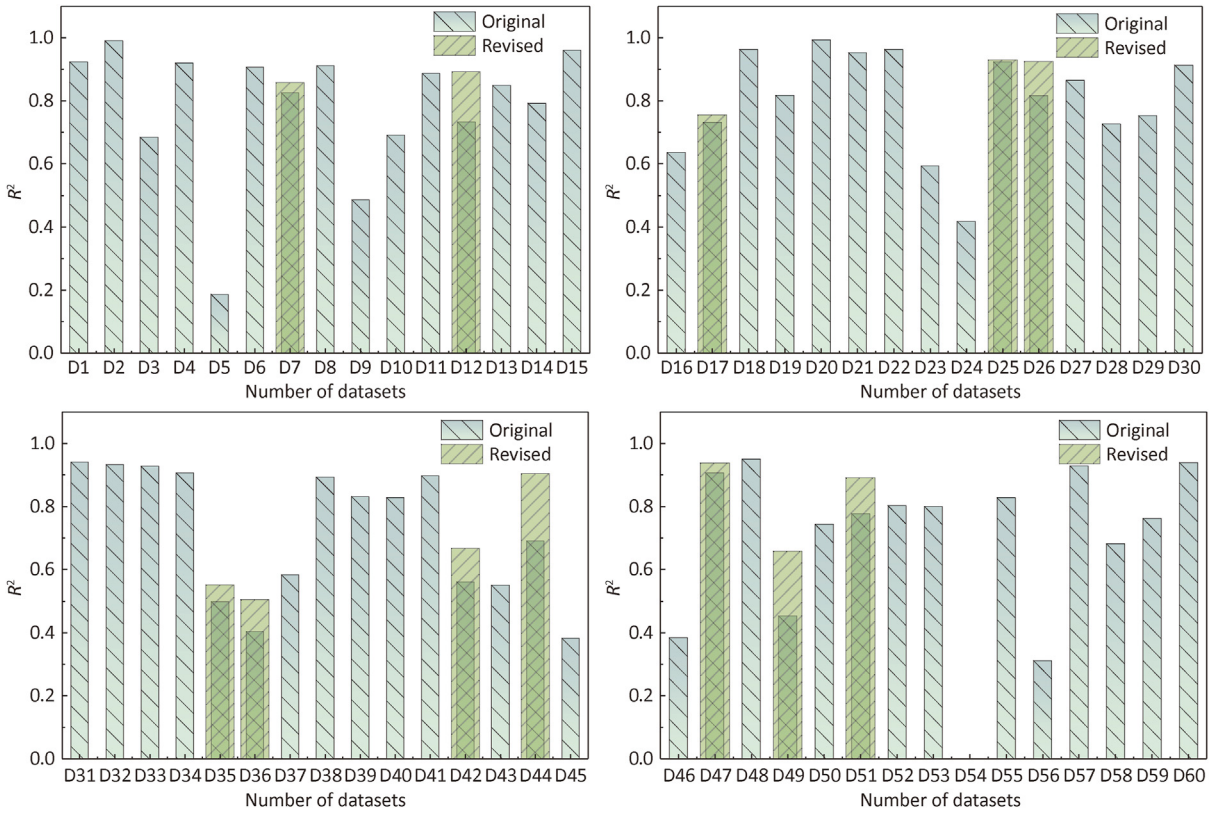

Fig. 9. Anomaly detection results (R^2) of 60 materials datasets.

contrast, the remaining datasets are assessed and found to be free from discrepancies of accuracy, requiring no further detection. Here, the NASICON-type solid electrolyte prediction dataset with complex inter-feature interaction mechanisms (D44) is used as an example to illustrate the detection and governance process, while the process for the remaining datasets is described in Section 3.2 of the SI.

3.5. Case study to NASICON-type solid electrolyte

(1) Single-descriptor accuracy detection.

In this section, single-descriptor data is assessed sequentially based on Rule 1, as well as the descriptor value ranges and data types presented in Table 1. As depicted in Table 4, anomalies are identified in two samples on the Valence_avg_M and Valence_M1 descriptors.

Subsequently, 3σ detection, Z-score, and boxplot are executed in parallel. Based on Eq. (4), the comprehensive detection results of the three methods are depicted in Fig. 10. From the scatter plot, it is evident that anomalies are detected across 18 descriptors, including Radius_X1, E_a, and Occu_X1. Additionally, 9 samples including No.5, No.6, No.7, No.38 to No.43 are detected to have anomalies in multiple dimensions. Therefore, in single dimension, we can conclude **modification strategy 1:** verify all anomaly samples in Fig. 10, especially focusing on samples No. 5, 6, 7, 38, 39, 40, 41, 44, 43.

(2) Multi-descriptor correlation detection.

The relationship between dimensions is introduced in detail in Section 2.2. Through the PCC and SCC, the correlation between dimensions is obtained. The results are shown in Fig. 11. It can be found that descriptor a and c are positively correlated with V. However, the correlation between descriptors in_BT, RT, BT1, BT2, V_Na_1O_6 and E_a is very low or close to zero. This is completely inconsistent with the conclusion obtained from Eqs. (5)–(12). After analyzing the reasons, we find that the correlation related to descriptor E_a contradicts domain knowledge. Therefore, we obtain **modification strategy 2:** there may be errors in descriptor "E_a".

(3) Sample reliability detection.

According to Rule 3, the clustering results of features and target property are shown in Fig. 12. The colored dots in each subplot indicate the clustering results of features, while all points labeled with numbers represent the clustering results of the target property (E_a). Different colors represent different categories of feature clustering, and the 8 categories of target property clustering can be seen in numerical form in Fig. 12a–h. If the target property of other samples within the same feature clustering are also clustered in the same category, these samples are considered to be non-anomalous. For example, in Fig. 12b–c, e and g, both features and target property are clustered into the same category, verifying that samples with similar structures and compositions often tend to similar material properties. When there is a significant discrepancy between the feature clustering results and activation energy clustering results, and the number of points sharing the same color is less than 2, these samples are identified as abnormal samples and are submitted to experts for verification. According to Fig. 12, No.29, No.45, No.56, No.68, and No.89 are identified as anomalies. Fig. 13 displays the clustering results containing only anomalous

Table 4
Outliers in single-dimensional data accuracy improvement model.

No.	Chemical Formula	Valence_avg_M	Valence_M1
17	$Na_{16.74}Cr_{12}P_{18}O_{72}$	3.105	3.105
72	$Na_3Nb_{12}P_{18}O_{72}$	4.25	4.25

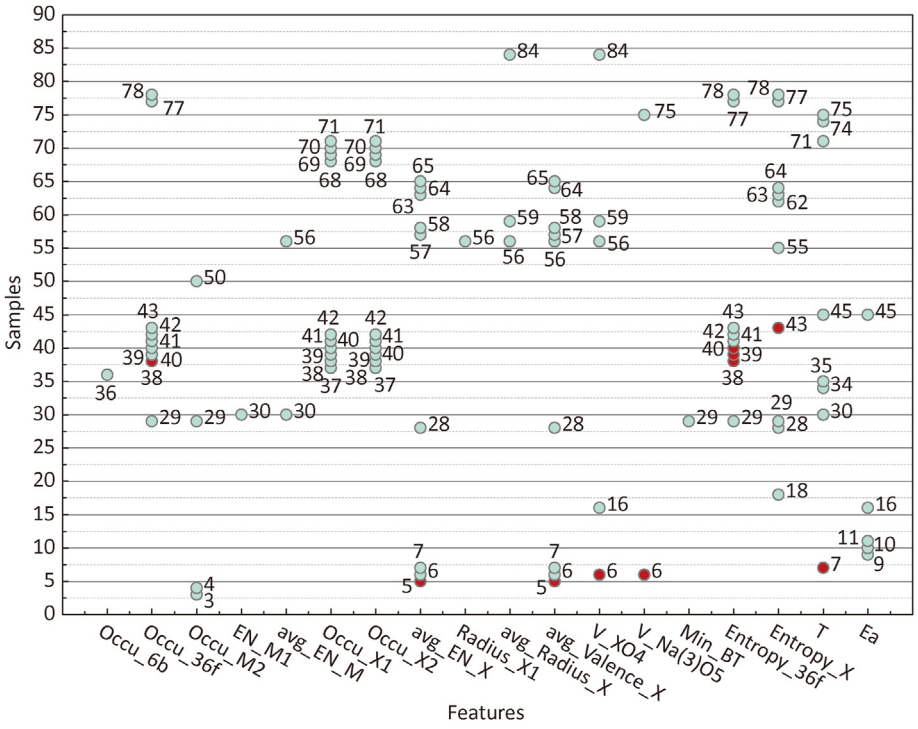

Fig. 10. Scatter of the synthesis results of the three methods, showing the anomalies in each descriptor. The horizontal axis is the descriptors (in Table S8), and the vertical axis is the sample number.

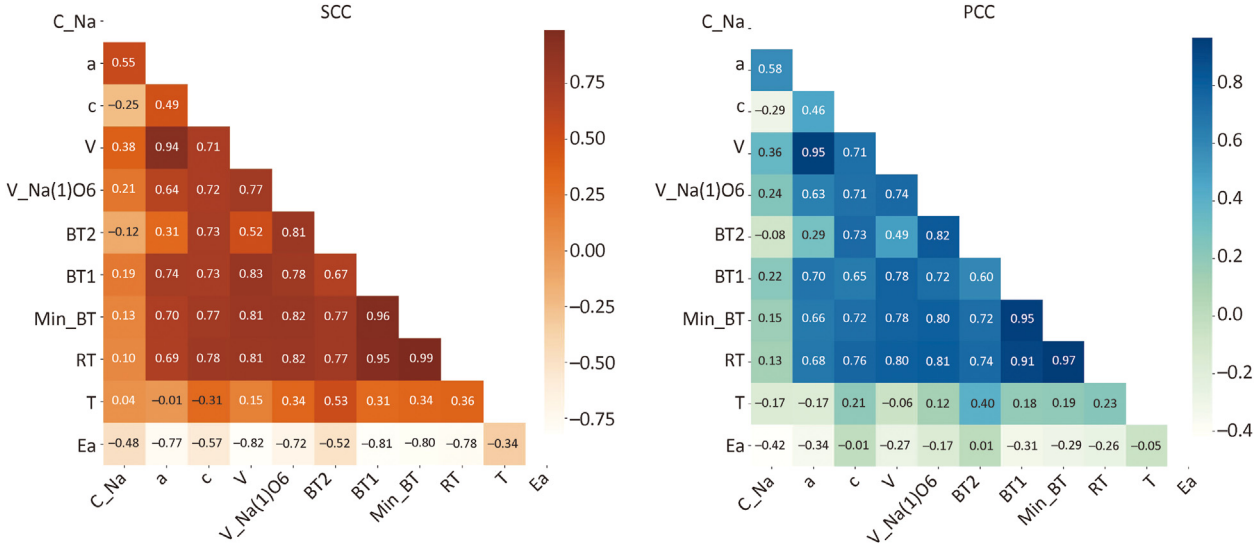

Fig. 11. Result of PCC and SCC. The color bands represent the mapping of values to colors; The darker the color, the higher the correlation, and vice versa.

samples. To clearly demonstrate the anomaly detection results based on Rule 3, only the anomalies that do not conform to the rules are marked in each category.

In order to further explore the effect of abnormal samples on clustering results, clustering of the target property is depicted in Fig. 14. It is evident that the activation energy of samples No.9, No.10, No.11, No.16, and No.45 are significantly higher than that of other samples, indicating potential anomalies. Subsequently, based on Table S1, OCSVM is selected for data-driven anomaly detection again, a total of 9 samples, including No.6, No.29, No.30, No.47, No.56, No.59, No.75, No.84, No.89, are detected as anomalous.

Finally, integrating the *K*-means clustering results with the results of data-driven anomaly detection, we get **modification strategy 3:** No.29, No.45, No.56, No.68 and No.89 may be anomaly; the decision attributes of samples No.9, No.10, No.11, No.16, No.45 may be anomaly.

(4) Comprehensive Modification.

Due to the challenging nature of collecting materials data and its uneven distribution, it is difficult to make a judgment only based on the detection results of an individual dimension. Therefore, a comprehensive analysis of the three modification strategies is deemed a reasonable approach. In S-DAD, anomalies are identified

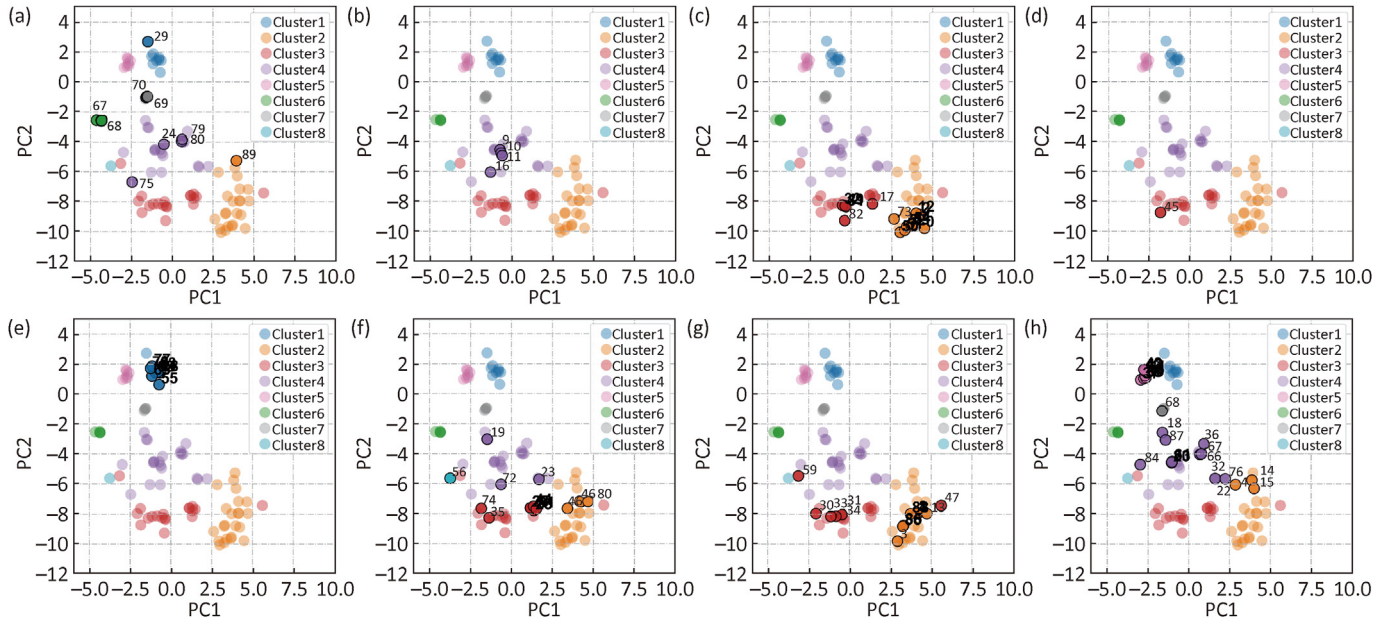

Fig. 12. Clustering results of feature and activation energy (all samples); (a–g) respectively represent the clustering overlap of features and activation energy for each of the 8 clusters. The horizontal and vertical coordinates represent the two dimensions after dimension reduction by t-SNE (t-distributed Stochastic Neighbor Embedding).

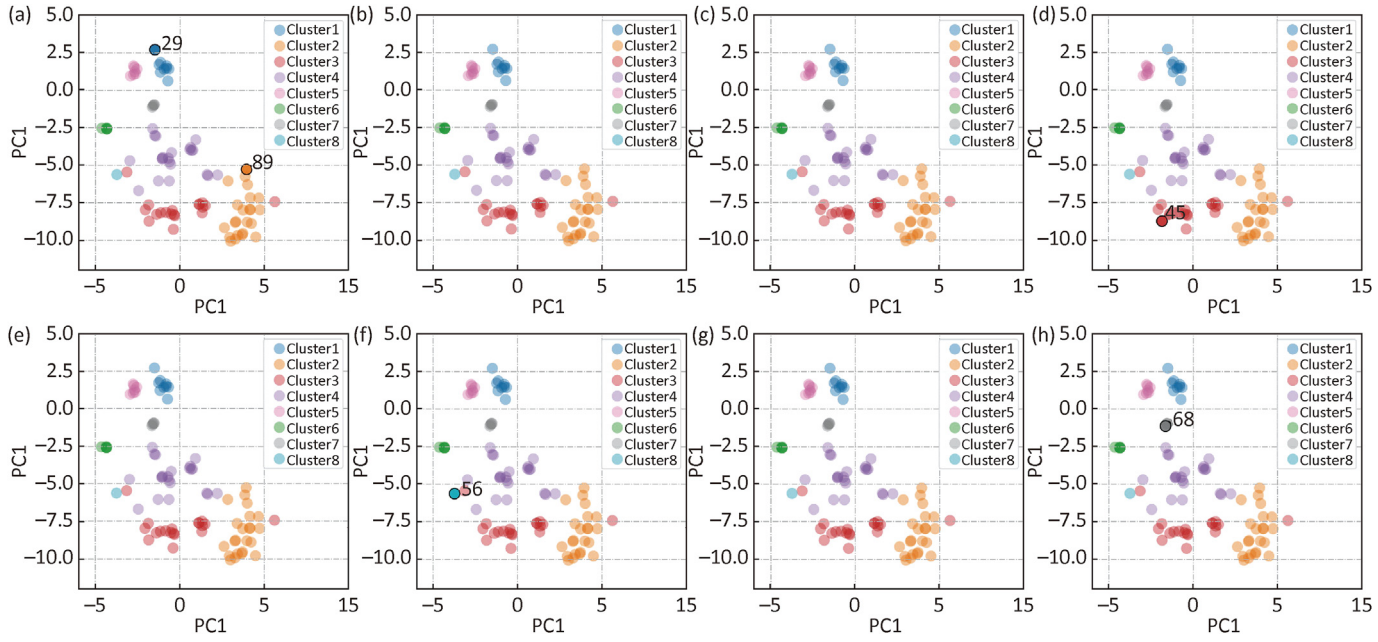

Fig. 13. Clustering results of feature and activation energy (Only mark anomaly samples).

in samples No.17 and No.72 due to non-integer values in descriptors V_avg_M and V_M1. Further analysis of their chemical structures reveals that these two samples exhibit no anomalies. In these two samples, the reason why fractional values exist is that there are different valence states of the atom at the M1 position in the materials. Therefore, two samples are retained. Although anomalies are detected in 18 descriptors according to Fig. 10, only a few anomalies are truly abnormal upon verification of material domain knowledge. Others are misidentified anomalies resulting from small sample sizes but high diversity. For instance, in descriptor Occu_X1, where most values are 1, the remaining values are identified as anomalies (Fig. 15a). However, these anomalies are attributed to the uneven distribution of the data. Similarly, in descriptors Occu_6b and Occu_36f for Na, only 1.12% of the compounds had zero occupancy rate at the Na(1) site, and only 23.6% of the compounds had non-zero occupancy rate at the Na(3) site. Thus, the smaller values in Occu_6b and the larger values in Occu_36f are not anomaly data (as shown in Fig. 15c and d). In descriptor Ea, 5 samples are identified as abnormal due to their large values, and anomalies in this descriptor are also detected in M-DCD and SRD. Due to the complexity of the real scenario, these potential anomalies should be further verified by experts. Upon further verification, apart from the E_a, other descriptors prove accurate. For the five anomalies, the Ea calculations are incorrect due

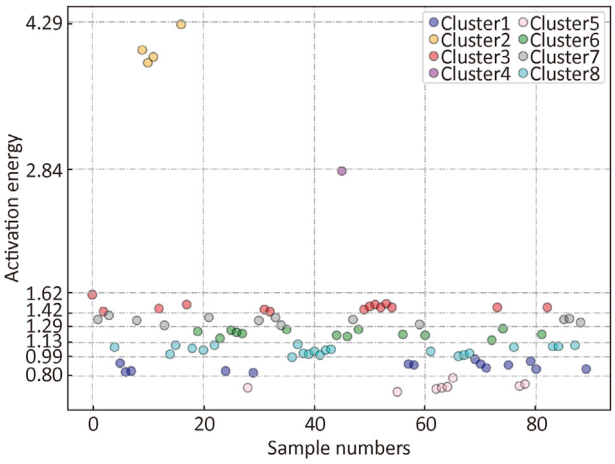

Fig. 14. Clustering results of activation energy ($K = 8$).

to atomic doping. Further investigation reveals that the calculation error stems from mixing migrating ions (Na/Cs and Na/K). Cs and K are employed as skeleton ions in the calculation program. When we change the mixed occupation of Na/Cs, and Na/K to that Na completely occupies the site, the calculated data returns to normal. However, for the sake of rigor, these five samples are removed from the dataset. Subsequently, Pearson and Spearman correlation analysis are conducted again, as shown in Fig. 16. Compared with Fig. 11, in_BT, RT, BT1, BT2, V_Na_1O_6 and Ea exhibit strong correlations, consistent with domain knowledge.

Then, further analysis of these marked abnormal samples (No.5, No.6, No.7, No.29, No.38, No.39, No.40, No.41, No.42, No.43, No.45, No.56, No.68, No.89). Following expert inspection, it is discovered that the lattice constants in the CIF files of No.5, No.6, No.7 are wrong, leading to calculation errors in unit cell volume, polyhedron volume, bottleneck, and activation energy data. Further examination finds that it is caused by the input error of the original files. Hence, these three samples are corrected, and the dataset is updated. (Table 5). With the assistance of domain knowledge, the correctness of each individual descriptor, the correlation between descriptors, and the reliability between samples are comprehensively evaluated and integrated governance. As a result, 3 samples are modified and 5 samples are deleted, resulting in a revised dataset containing 85 samples and 45 features.

(6) Prediction of activation energy and discussion of descriptors.

To confirm the effectiveness of DKA – DAD, we employed six prediction models (LASSO, GPR, Ridge, SVR, KNN, and RF) to predict E_a on the NASICON dataset. The average RMSE, MAPE and R^2 of 10 iterations of each model on both the raw dataset and revised dataset are presented in Fig. 17. Lower values of RMSE and MAPE indicate better predictive performance of the model. Compared with the original dataset, the prediction accuracy of each model on the revised dataset is significantly improved. Note that KNN causes a decrease in performance trained on the revised dataset and similar phenomenon occurs in Fig. S29. This is because due to the limited scale of the two datasets, while the revised dataset

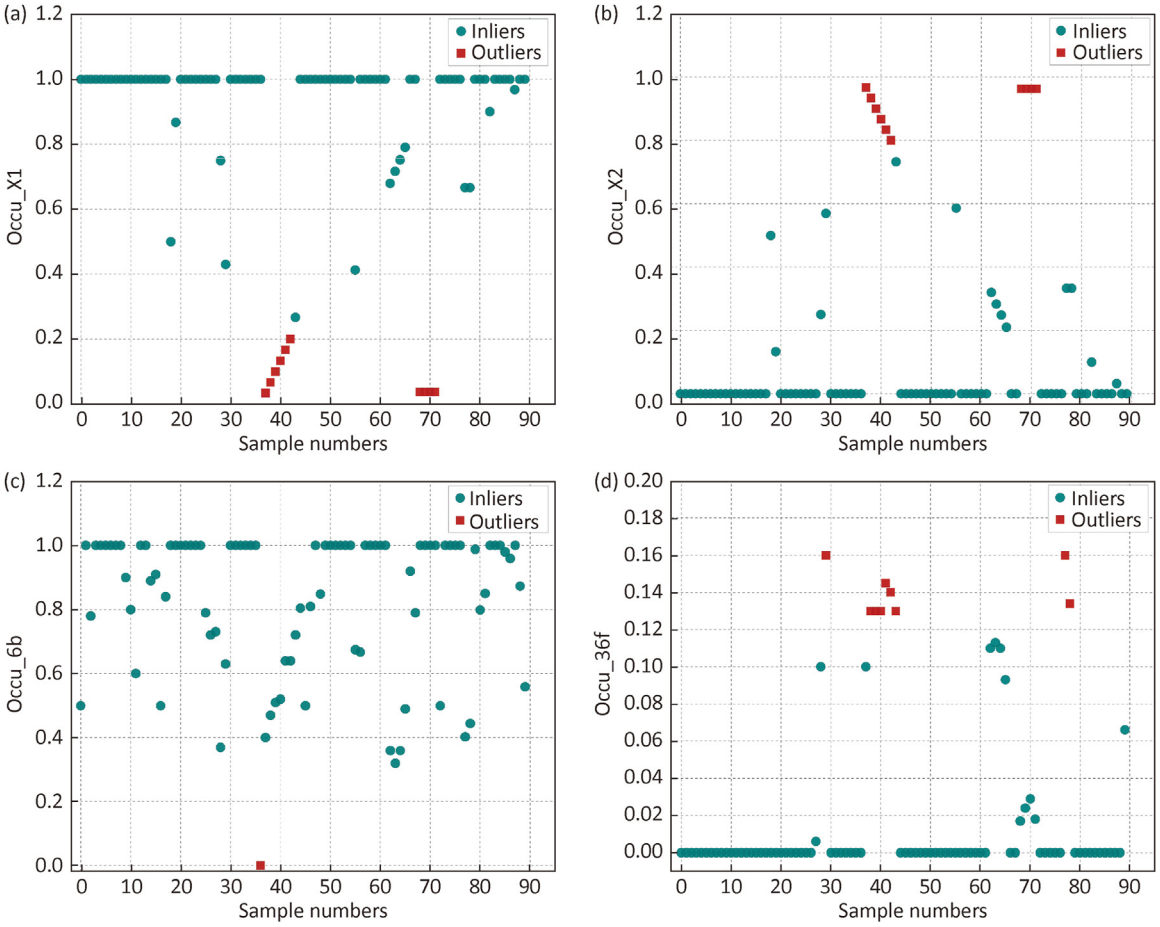

Fig. 15. Anomalous points in (a) Occu_X1, (b) Occu_X2, (c) Occu_6b, (d) Occu_36f.

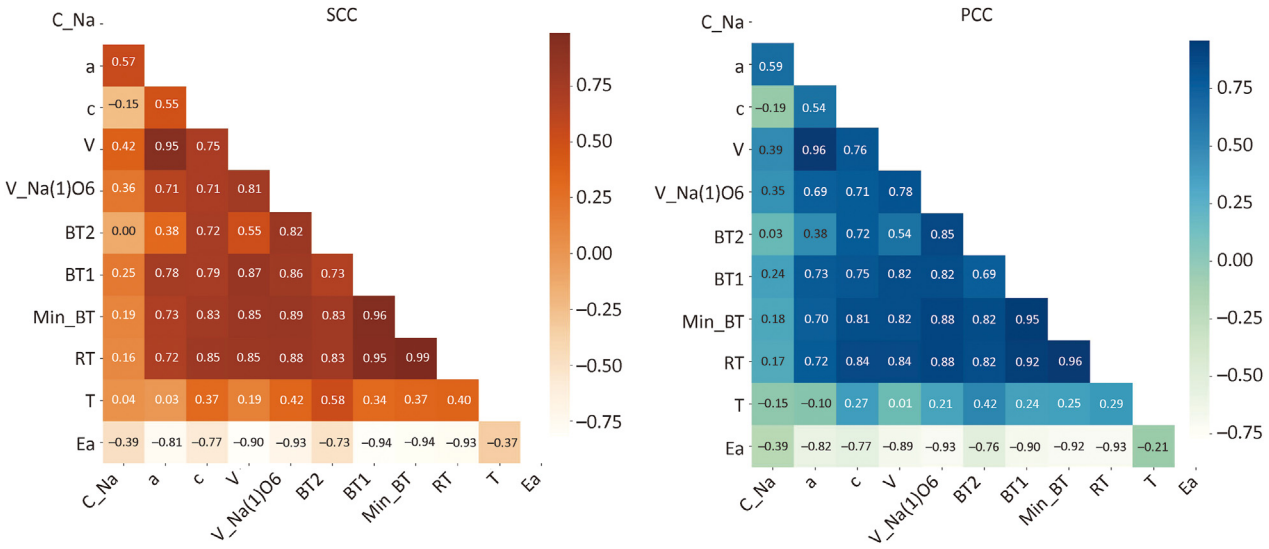

Fig. 16. Result of PCC and SCC (on revised dataset). The color bands represent the mapping of values to colors; The darker the color, the higher the correlation, and vice versa.

Table 5
Abnormal data detection and correction.

No.	ICSD	Formula	a	c	V_{cell}	Revised a	Revised c	Revised V_{cell}
5	15545	$Na_{24}Zr_{12}Si_{18}O_{72}$	9.186	22.181	1621.04	**9.198**	**22.210**	**1627.29**
6	15546	$Na_{24}Zr_{12}Si_{18}O_{72}$	9.186	22.181	1621.04	**9.199**	**22.470**	**1646.70**
7	15547	$Na_{24}Zr_{12}Si_{18}O_{72}$	9.186	22.181	1621.04	**9.199**	**22.706**	**1663.99**

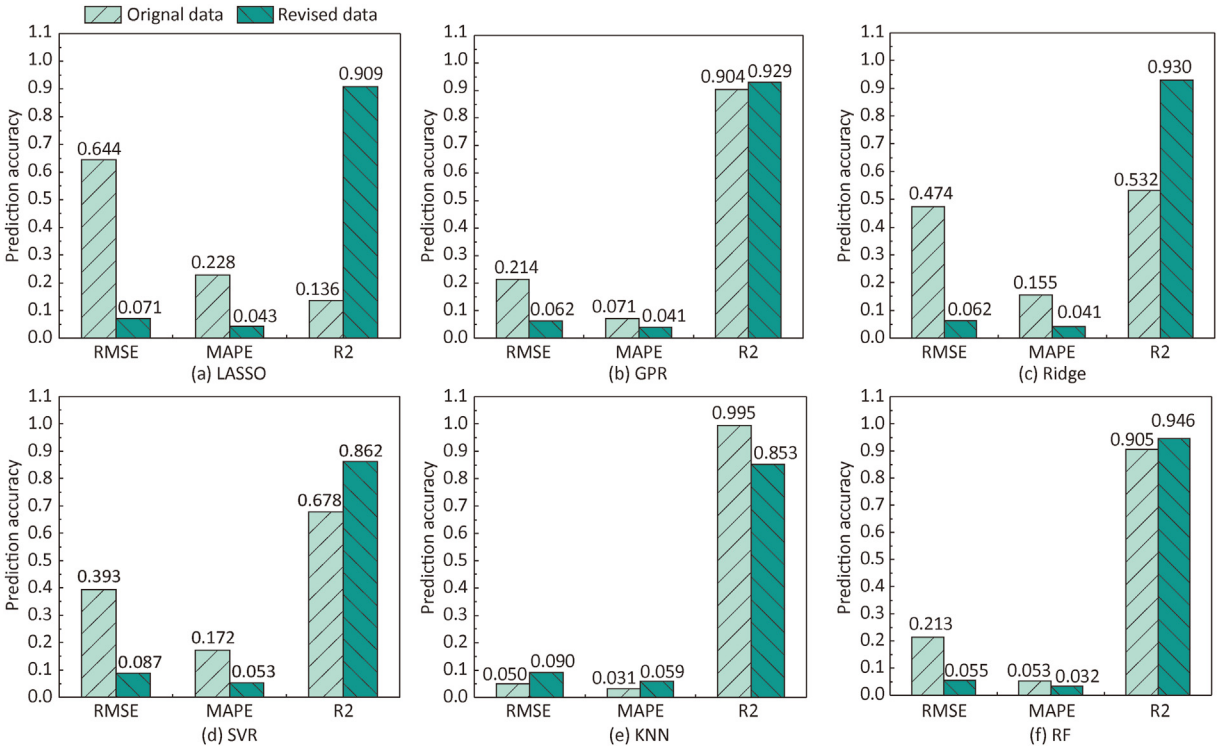

Fig. 17. The prediction accuracy of 6 ML models on raw and revised datasets.

maintains compliance with materials domain knowledge, its expanded data distribution surpasses the modeling capacity of conventional shallow ML algorithms. Notably, RF exhibits robust predictive performance in both original and revised datasets. This observation aligns with the established superiority of tree-based architectures in structured data analysis, as systematically

demonstrated by Borisov *et al.* [37], where such models demonstrate enhanced adaptability to complicated data hierarchies.

In summary, the revised dataset based on DKA – DAD proves to be reasonable and effective. It shows that the use of DKA – DAD is able to detect the dataset comprehensively to improve its accuracy, further demonstrating the feasibility and efficiency of the method. Moreover, based on this revised dataset, we explored the influence of feature engineering further and proposed a feature selection method embedded with materials domain knowledge [18]. This method effectively filters redundant descriptors, *e.g.*, BT1 and BT2 are removed due to their high correlation to Min_BT. Fourteen key descriptors are maintained from 45 descriptors for ML modelling, which significantly improves 8.7% prediction accuracy of the optimal ML model compared with the original dataset, of which RMSE decreases from 0.046 to 0.042.

4. Conclusions and outlook

In this study, we propose a data anomaly detection workflow (DKA – DAD) in which three detection models (S – DAD, M – DCD, and SRD) and a Modification model are integrated to identify potential anomalies in the data and improve the accuracy of the dataset. We innovatively combine domain knowledge and data-driven methods into the analysis of the accuracy of single descriptors, inter-descriptors and inter-samples to obtain high-quality dataset for ML modelling. The effectiveness of S – DAD, M – DCD and SRD, as well as the benefits of embedding domain knowledge, is demonstrated on the validation experiments of 180 synthetic datasets. Subsequently, DKA – DAD is applied to evaluate the accuracy of 60 structure-activity relationship research datasets, 12 datasets with potential anomalies are identified and accuracy governance carried out. Compared to the original datasets, the average R^2 of ML model on the 12 revised datasets improved by 9.6%, which illustrates that DKA – DAD can accurately identify potential anomaly samples and performs reasonable corrections, thus obtaining high-quality samples for ML modes and providing effective tool support for materials experts to obtain high-quality datasets.

Note that the connotation of knowledge in the field of materials still has strong limitations. Therefore, methods have varying requirements for the representation forms of domain knowledge in materials science and lack unified symbolic standards for materials domain knowledge. To enable broader and more systematic development of materials domain knowledge embedded into the detection processes, future efforts should actively explore richer material domain knowledge content and establish unified representation formats. This will provide standardized methodologies for more materials researchers to construct domain knowledge-embedded machine learning models.

At the same time, the growing adoption of ML techniques in materials science exhibits new challenges. For instance, while our current method applies to structured data, developing analysis methods for other descriptor types, such as graph-based descriptors, remains a future consideration. Moreover, while descriptors were manually processed in our work, our vision is to automate descriptor extraction through natural language processing technology. Finally, there is a need for more generalized approaches to enhance the collection and quality monitoring of materials data for ML applications.

CRediT authorship contribution statement

Yue Liu: Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Shuchang Ma:** Writing – original draft, Validation, Software, Investigation. **Zhengwei Yang:** Writing – original draft, Software, Methodology, Data curation. **Duo Wu:** Writing – original draft, Resources, Formal analysis, Data curation. **Yali Zhao:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation. **Maxim Avdeev:** Writing – review & editing. **Siqi Shi:** Writing – review & editing, Project administration, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 92270124, 92472207 and 52073169) and National Key Research and Development Program of China (No. 2021YFB3802101).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmat.2025.101066.

References

[1] Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature 2015;521(7553):452–9.
[2] Liu Y, Guo BR, Zou XX, Li YJ, Shi SQ. Machine learning assisted materials design and discovery for rechargeable batteries. Energy Storage Mater 2020;31:434–50.
[3] Choubisa H, Todorović P, Pina JM, Parmar DH, Li Z, Voznyy O, et al. Interpretable discovery of semiconductors with machine learning. NPJ Comput Mater 2023;9(1):117.
[4] Meng H, Wei P, Tang Z, Yu H. Data-driven discovery of formation ability descriptors for high-entropy rare-earth monosilicates. J Materiomics 2024;10(3):738–47.
[5] Liu Y, Yang ZW, Yu ZY, Liu ZT, Liu DH, Lin HL, et al. Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. J Materiomics 2023;9(4):798–816.
[6] Takamoto S, Okanohara D, Li QJ, Li J. Towards universal neural network interatomic potential. J Materiomics 2023;9(3):447–54.
[7] Xu S, Chen Z, Qin M, Cai B, Li W, Zhu R, et al. Developing new electrocatalysts for oxygen evolution reaction *via* high throughput experiments and artificial intelligence. npj Comput Mater 2024;10(1):194.
[8] Katcho NA, Carrete J, Reynaud M, Rousse G, Casas-Cabanas M, Mingo N, et al. An investigation of the structural properties of Li and Na fast ion conductors using high-throughput bond-valence calculations and machine learning. J Appl Crystallogr 2019;52(1):148–57.
[9] Xu YJ, Zong Y, Hippalgaonkar K. Machine learning-assisted cross-domain prediction of ionic conductivity in sodium and lithium-based superionic conductors using facile descriptors. J Phys Commun 2020;4(5):055015.
[10] Liu YW, Yu HL, Meng H, Chu YH. Atomic-level insights into the initial oxidation mechanism of high-entropy diborides by first-principles calculations. J Materiomics 2024;10(2):423–30.
[11] Fang C, Wang H, Shi SQ. Quantifying structural distortion manipulation for desired perovskite phase: Part I. Paradigm demonstration in tungsten oxides. J Materiomics 2024;10(2):293–303.
[12] Liu Y, Yang ZW, Zou XX, Ma SC, Liu DH, Avdeev M, et al. Data quantity governance for machine learning in materials science. Natl Sci Rev 2023;10(7):nwad125.
[13] Beal MS, Hayden BE, Le Gall T, Lee CE, Lu X, Mirsaneh M, et al. High throughput methodology for synthesis, screening, and optimization of solid state lithium ion electrolytes. ACS Comb Sci 2011;13(4):375–81.
[14] Hemmati-Sarapardeh A, Tashakkori M, Hosseinzadeh M, Mozafari A, Hajirezaie S. On the evaluation of density of ionic liquid binary mixtures: modeling and data assessment. J Mol Liq 2016;222:745–51.
[15] Amiri-Ramsheh B, Nait Amar M, Shateri M, Hemmati-Sarapardeh A. On the evaluation of the carbon dioxide solubility in polymers using gene expression programming. Sci Rep 2023;13(1):12505.
[16] Wenzlick M, Mamun O, Devanathan R, Rose K, Hawk J. Assessment of outliers in alloy datasets using unsupervised techniques. J Occup Med 2022;74(7):2846–59.
[17] Li X, Shan GC, Zhao HB, Shek CH. Domain knowledge aided machine learning method for properties prediction of soft magnetic metallic glasses. Trans

Nonferrous Metals Soc China 2023;33(1):209–19.
[18] Liu Y, Zou XX, Ma SC, Avdeev M, Shi SQ. Feature selection method reducing correlations among features by embedding domain knowledge. Acta Mater 2022;238:118195.
[19] Liu Y, Wu JM, Avdeev M, Shi SQ. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties. Advanced Theory and Simulations 2020;3(2):1900215.
[20] Shi SQ, Sun SY, Ma SC, Zou XX, Qian Q, Liu Y. Detection method on data accuracy incorporating materials domain knowledge. J Inorg Mater 2022;37(12):1311.
[21] Qui DT, Capponi JJ, Gondrand M, Saïb M, Joubert JC, Shannon RD. Thermal expansion of the framework in Nasicon-type structure and its relation to Na^+ mobility. Solid State Ionics 1981;3–4:219–22.
[22] Losilla ER, Aranda M, Bruque S, París MA, Sanz J, West AR. Understanding Na Mobility in NASICON materials: a rietveld, 23Na and 31P MAS NMR, and impedance study. Chem Mater 1998;10(2):665–73.
[23] Shannon R. Revised effective ionic radii and systematic study of inter atomic distances in halides and chalcogenides. Acta Crystallogr A 1976;32:751–67.
[24] Sedgwick P. Pearson's correlation coefficient. BMJ Br Med J (Clin Res Ed) 2012;345:e4483.
[25] Zhou Y, Li S. BP neural network modeling with sensitivity analysis on monotonicity based Spearman coefficient. Chemometr Intell Lab Syst 2020;200:103977.
[26] Tzortzis G, Likas A. The MinMax k-Means clustering algorithm. Pattern Recogn 2014;47(7):2505–16.
[27] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data. Dallas, Texas, USA: Association for Computing Machinery; 2000. p. 427–38.
[28] Liu FT, Ting KM, Zhou ZH, Isolation-Based Anomaly Detection. ACM trans. Knowl. Discov. Data 2012;6(1). Article 3.
[29] Breunig MM, Kriegel HP, Ng RT, Sander J. OPTICS-OF: identifying local outliers. In: Zytkow JM, Rauch J, editors. Principles of data mining and knowledge discovery; 1999. p. 262–70.
[30] Ma J, Perkins S. Time-series novelty detection using one-class support vector machines. In: Proceedings of the international joint conference on neural networks; 2003.
[31] Hubert M, Vandervieren E. Vandervieren. An adjusted boxplot for skewed distributions. Computational Statistics & Data Analysis 2008;52(12):5186–201.
[32] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006;101(476):1418–29.
[33] Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. Chem Rev 2021;121(16):10073–141.
[34] McDonald GC. Ridge regression. WIREs Computational Statistics 2009;1(1):93–100.
[35] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
[36] Breiman L. Using iterated bagging to debias regressions. Mach Learn 2001;45:261–77.
[37] Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep neural networks and tabular data: a survey. IEEE Transact Neural Networks Learn Syst 2024;35(6):7499–519.

Yue Liu obtained her B.S. and M.S. in computer science from Jiangxi Normal University in 1997 and 2000. She finished her Ph.D. in control theory and control engineering from Shanghai University (SHU) in 2005. She has been working with the School of Computer Engineering and Science of SHU since July 2000. During that time, she has been a curriculum R&D manager at the Sybase-SHU IT Institute of Sybase Inc. from July 2003 to July 2004 and a visiting scholar at the University of Melbourne from Sep. 2012 to Sep. 2013. At present, she is a professor of SHU. At present, she is a professor of SHU. Her current research interests focus on research of data mining, machine learning, and their applications in materials science.

Siqi Shi obtained his B.S. and M.S. from Jiangxi Normal University in 1998 and in 2001, respectively. He finished his Ph.D. from Institute of Physics, Chinese Academy of Sciences in 2004. During this period, focused on the electrolyte, electrode materials and relevant interfaces for lithium-ion batteries, he carried out the first-principles calculation and design on the ionic transport physics, cooperative electron/ion transport control problem earliest in China. After that, he joined the National Institute of Advanced Industrial Science and Technology of Japan and Brown University of USA as a senior research associate until joining Shanghai University as a professor in early 2013. His current research interests focus on the fundamentals and multiscale calculation of electrochemical energy storage materials and materials design and performance optimization using machine learning.