Contents lists available at ScienceDirect

Materials Science & Engineering R

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

A general framework to govern machine learning oriented materials data quality

Yue Liu [a], Zhengwei Yang [a], Xinxin Zou [a], Yuxiao Lin [e], Shuchang Ma [a], Wei Zuo [a], Zheyi Zou [d], Hong Wang [f], Maxim Avdeev [g,h], Siqi Shi [b,c,*]

[a] *State Key Laboratory of Materials for Advanced Nuclear Energy & School of Computer Engineering and Science, Shanghai University, Shanghai 200444, 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] *School of Materials Science and Engineering, Xiangtan University, Xiangtan 411105, China*
[e] *School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, China*
[f] *Materials Genome Initiative Center & School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[g] *Australian Nuclear Science and Technology Organisation, Sydney 2232, Australia*
[h] *School of Chemistry, The University of Sydney, Sydney 2006, Australia*

ARTICLE INFO

Keywords:
Machine learning
Materials science
Data quality
Domain knowledge

ABSTRACT

Machine learning (ML) is increasingly applied in materials discovery and property prediction, mainly due to its advantage of low-cost and efficient data analysis process. The materials data quality can heavily influence the performance of ML models. However, most current data quality improvement approaches are purely data-driven, neglecting materials domain knowledge and data quality issues latent in the entire process of ML modelling. Here, we address the definition of high-quality data and propose a general framework for ML-oriented MATerials Data Quality Governance incorporating domain knowledge (MAT-DQG), involving nine dimensions defining WHAT materials data quality should be evaluated, lifecycle models guiding WHEN to execute data governance activities in the entire process of ML modelling, and processing models guiding HOW to detect and address issues related to materials data quality. 60 datasets from materials ML studies are assembled to demonstrate potential utility and applications of MAT-DQG, including mining complicated structure-activity relationships in metals, inorganic non-metals, polymers, and composite materials. MAT-DQG identifies and resolves issues in 17 datasets and as a result prediction accuracy improvements of up to 49 % are achieved. Our work lays a foundation for governing ML-oriented materials data and ensuring its reusability and reliability, which advances the frontiers of materials discovery and design.

With its strong capability for data analysis, machine learning (ML) plays an increasingly important role in accelerating novel materials discovery and design [1–4]. The reliability and credibility of ML models are largely affected by materials data issues (e.g., inaccuracy, noise, and insufficiency), which seem ubiquitous and inevitable due to measurement errors, equipment failures, calculation defects, and the laborious process of acquisition. Furthermore, while most materials researchers show interest in developing ML modeling for specific tasks, very few of them undertake tedious and time-consuming data work [5–7], resulting in the lack of easy access to high-quality ML-oriented data of sufficient volume [8], despite the fact that ML models accuracy strongly depends on the materials data quality and quantity [9]. Although a synergistic data quantity governance flow was proposed in the previous research by exploring the governance towards feature and sample quantities [10], the methodologies to assess and monitor the other critical factor, namely data quality, still remain to be addressed.

High-quality ML oriented materials data should possess three characteristics, i.e., it should be free from erroneous data to provide data accuracy [11], fit the purpose of ML modeling to provide data usability, and satisfy the accessibility requirements to provide compatibility with downstream tasks. Some efforts [12–16] were previously made to improve materials data quality in terms of accuracy at the data preprocessing stage [17], redundancy at the feature engineering stage [18], or unbalanced distribution at the model construction stage [19].

[*] Corresponding author at: 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.mser.2025.101050
Received 3 December 2024; Received in revised form 4 June 2025; Accepted 11 June 2025

However, several more challenges still exist for the assessment and improvement of materials data quality: (i) Comprehensive improvement of the reliability and performance of ML model analysis can be hardly achieved by improving only one specific dimension at a particular stage of ML modeling. (ii) Existing data science approaches [20–24] to ML-oriented data quality improvement primarily focus on a sample level, typically in a single data format. Although FAIR data guidelines [17], which are designed to ensure traceability, format consistency, and integrity of multi-source data, have been introduced into the field of materials science by Draxl et al. [25], practical and concrete schemes are still lacking in real-world applications. To overcome the challenge of discovering high-quality materials datasets adhering to the FAIR principles, some efforts focus on the construction of appropriate tools [8, 26–28]. For example, Schmidt et al. [29] introduced a software named Foundry-ML, which opens accessible pathway to publish and discover ML-oriented structured datasets for researchers. However, these studies ignore the fact that data quality issues exist in the entire process of ML modelling, thus comprehensive evaluation of essential dataset characteristics, e.g., data accuracy and redundancy, can further benefit ML modelling [30]. (iii) The longstanding challenges associated with purely data-driven approaches also remain in the governing process of materials data quality, such as neglecting domain knowledge, inconsistency between the simulated and real data distributions, or the inappropriate assignment of relative importance to the descriptors or samples [31–33]. Hence, a general framework, addressing different data quality issues to globally govern materials data quality throughout the entire ML process under the guidance of domain knowledge is in urgent need but still lacking [34].

To this end, we propose a general framework for MATerials Data Quality Governance incorporating domain knowledge (MAT-DQG) for ML, which ensures reusability and reliability of the materials data. This framework comprises the definition of comprehensive assessment dimensions for materials data quality, mapping between assessment dimensions and the entire lifecycle of materials data, and the organized and standard integration of each governance component. The development of the framework aims to bridge the communities of ML and materials science and can be summarized as follows:

1. The quality requirements to ML-oriented materials data at all stages of the whole life-cycle are clearly defined, such as generation, collection, processing, modeling, and application and the data quality is assessed in multiple dimensions, i.e., six inherent quality dimensions (redundancy, traceability, consistency, accuracy, time-sensitivity and completeness) and three contextual quality dimensions (insight, normalization and balance), to clarify WHAT materials data quality should be assessed.
2. A lifecycle model is further constructed to regulate all activities related to quality governance in all the nine dimensions, thus providing a standard for the practical step-by-step governance of materials data quality in ML, which determines WHEN the processing models should be performed.
3. With the incorporation of materials domain knowledge, processing models are designed, which aim to detect issues along with the process of ML modeling and provide targeted methods to improve materials data quality for all quality assessment dimensions, thus describing HOW to govern data quality. Note that considering the spectacular ability of data analysis and processing of large language models (LLMs) [35], such approach will be discussed in governance schemes aiming at accuracy and redundancy.

Regarding the application, MAT-DQG is thoroughly evaluated on 60 publicly available ML-oriented materials structured datasets as a case study. That demonstrated not only its capability of monitoring data quality through the whole process of ML modelling and improving accuracy in the assessment of materials data quality and its effectiveness of quality governance but also lead to improvement in the accuracy of the corresponding ML models. In summary, the theory and techniques discussed in this article lay a foundation for assessing and improving the quality of ML-oriented materials data, which advances the frontiers of ML towards accelerating materials discovery and design.

Framework for ML oriented Materials data quality governance incorporating domain knowledge

The overall schematics of MAT-DQG are shown in Fig. 1, which consists of data quality dimensions, lifecycle model, and processing models.

2. Dimensions of materials data quality governance

2.1. Data quality dimensions (DQDs)

Considering that data quality is a complex concept, we define nine crucial quality dimensions of materials data that affect ML modeling, as shown in Fig. 2**a.** This helps to visualize *WHAT* quality of materials data should be assessed for ML modeling. These nine dimensions can be divided into inherent and contextual types.

2.2. Inherent type

Inherent Quality Dimensions (IQDs) generally refer to objective and native data attributes that constitute the focus of researchers in different fields for various data tasks (e.g., data analysis [36]). Herein, we provide definitions of six crucial IQDs, as listed below.

2.2.1. Traceability

measures whether the acquisition and ML-based processing of materials data is traceable. Repeatability is another key factor in determining the reliability of a ML project. Since sources of materials data are diverse and have various degrees of trustworthiness and quality, it is necessary to ensure that the input data of ML can be traced. In addition, data processing is generally a semi- or non-transparent process accompanied by randomness, which should also be traced.

2.2.2. Completeness

measures whether all data used for training ML models and other critical additional information are recorded with no missing entries. Most ML models cannot be effectively trained on the datasets with missing values. The loss of data may occur during the process of data generation, entry, and transmission. For instance, Bond Valence Site Energy (BVSE) method [37] cannot calculate the energy barrier of compounds containing hydrogen due to the intrinsic limitations and thus yields incomplete datasets. Moreover, the integrity of additional information (e.g., meta-data) is closely related to the traceability and reusability of ML projects.

2.2.3. Time-sensitivity

captures the varying physical/chemical properties that materials data may exhibit across different time periods. This evaluation dimension enables researchers to gain deeper insights into the underlying patterns embedded within current materials data, thereby facilitating more rational organization and utilization of material data to implement more meaningful ML modelling.

2.2.4. Consistency

measures whether samples in the multi-source dataset are represented in the same way and whether data with dependencies are correctly correlated. It makes no sense to train ML models based on the dataset with inconsistencies. Different sources generally have different specifications for data generation, characterization, and storage, leading to inconsistency issues in multi-source dataset, such as different coding rules, measurement units, etc. In addition, some operations may produce

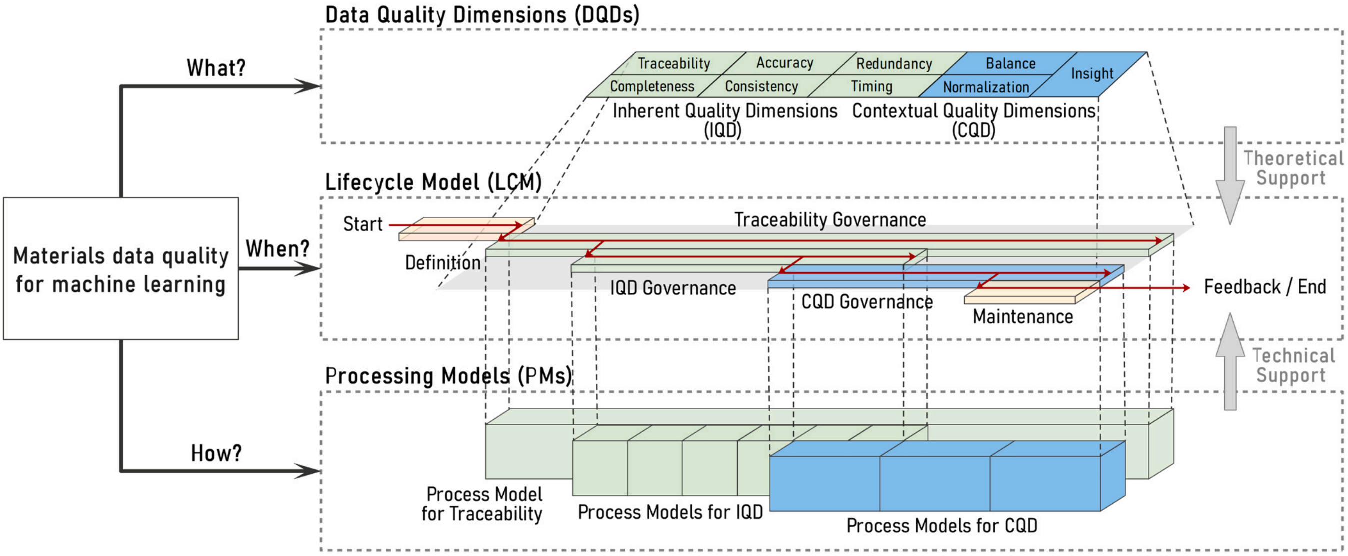


Fig. 1. Overall schematic of MAT-DQG. The data quality dimensions (DQDs) define the issues of data quality and provide life-cycle model (LCM) with input; LCM provides the routes of data quality governance for researchers according to DQDs; Following the LCM, the processing models (PMs) provide specific solutions. The framework comprehensively evaluates the data quality in terms of *WHAT*, *WHEN* and *HOW*.

inconsistencies in materials data during the process of ML application and data quality governance. For example, the structure files may have already been modified, whereas the calculated data are still based on the original structures.

2.2.5. Accuracy

measures whether data is recorded correctly and reflects realistic values. Given the “garbage in, garbage out” principle, erroneous data will lead to incorrect ML results, thus reducing the reliability of ML. Materials data often have well-known uncertainty or error, which hinders effective ML models modelling [38]. For example, E_{hull} (energy relative to convex hull) is a typical estimate of thermodynamic stability and largely relied on as a key criterion for materials design. However, its computation is highly sensitive to the choice of competing phases in phase diagrams, and the incompleteness of phase diagrams or the presence of artificially stabilized phases can lead to large errors in E_{hull} [39]. Meanwhile, for unstructured materials data, one vital process is the transformation of “unstructured” into “structured” type, which ensures that ML models can accurately capture key information latent in such data.

2.2.6. Redundancy

measures whether data contains redundant information. Redundant information will bias ML results. Unfortunately, redundant features (descriptors) are ubiquitous in the materials datasets. For instance, while it was reported that 111 features may possibly be associated with ionic transport, only 23 of them are effective descriptors to the prediction of diffusion barriers in FCC solute [40]. In addition, the same entity may be represented in different sources with certain differences, contributing to redundancy through duplicate samples. There is growing evidence that the accuracy of ML models hinges on large amount of diverse training data, namely if critical subdomains are underrepresented, predictions in those fields may falter, nevertheless [41].

2.3. Contextual type

As the framework proposed in this work focuses on ML-oriented data, we here define contextual quality dimensions (CQDs) for the data only employed for ML modelling. CQDs include the following three dimensions.

2.3.1. Balance

measures whether the population of different classes in the entire dataset is balanced. The imbalanced dataset may result in biased ML results, which often occurs when the number of samples representing one class is much larger (or smaller) than those of other classes.

2.3.2. Normalization

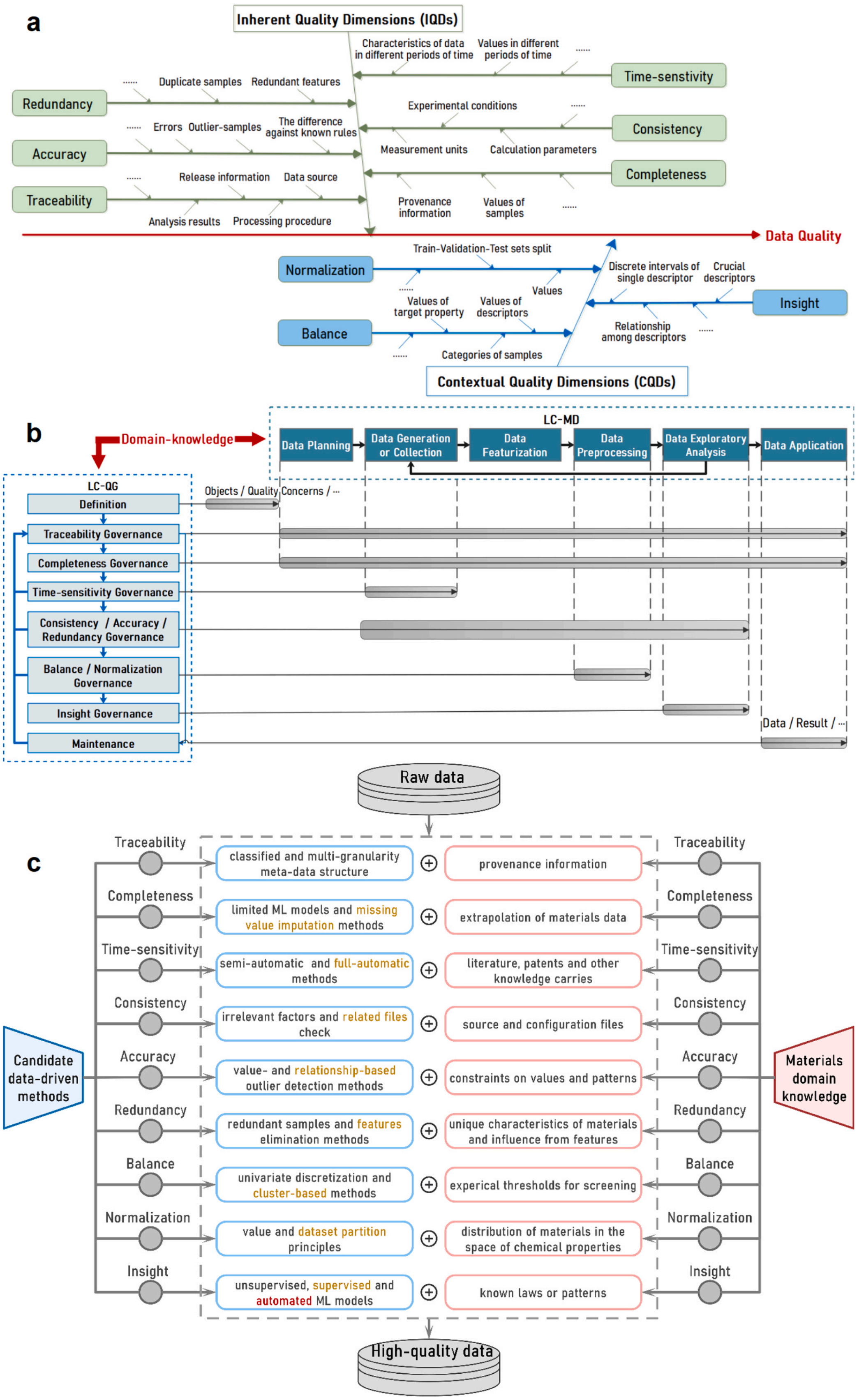
measures whether materials data has been converted into the representation or organization forms suitable for ML modeling as required. Typically, the values of different features should be normalized into the same magnitude under principles recognized by ML community. Nevertheless, materials data often consists of multi-scale features with different magnitudes, which can interfere with the feature importance evaluation. For example, non-normalized data can introduce noise, which impacts the accuracy and interpretability of ML models.

2.3.3. Insight

preliminarily quantifies the learnability of materials data before time-consuming analyses. For any ML project, a prerequisite for its success is that the learning goals must be potentially learnable from available data. But the definition of goals typically relies on domain knowledge of specific experts, making the learnability of the collected materials data questionable. Hence, it is necessary to rapidly explore potential patterns (i.e., insight) using various ML techniques—such as clustering, classification, and regression—before undertaking more time-consuming analyses.

3. Lifecycle model (LCM) of materials data quality governance

Rather than being a one-off step, the materials data quality governance should be ongoing during the entire process of ML application since data operation occurs at almost every stage of ML. Different data operations may also affect different data quality dimensions. Sometimes, the issue detection and quality improvement may even require the introduction of different methods at different stages. Herein, we establish a lifecycle model to govern materials data quality for ML, which guides *WHEN* to execute governance activities. This model consists of a lifecycle of materials data (LC-MD) during the whole process of ML application and lifecycle of data quality governance (LC-QG). These two lifecycles, together with their relationship with the nine dimensions, are shown in Fig. 2**b**.


a
Inherent Quality Dimensions (IQDs)
Characteristics of data in different periods of time
Values in different periods of time
......
Time-senstivity
......
Duplicate samples
Redundant features
Redundancy
Experimental conditions
......
Consistency
......
Errors
Outlier-samples
The difference against known rules
Accuracy
Measurement units
Calculation parameters
Completeness
......
Release information
Data source
Traceability
Analysis results
Processing procedure
Provenance information
Values of samples
......
Data Quality
Train-Validation-Test sets split
Normalization
Discrete intervals of single descriptor
Crucial descriptors
......
Values
Values of target property
Values of descriptors
Insight
Balance
Relationship among descriptors
......
......
Categories of samples
Contextual Quality Dimensions (CQDs)
b
Domain-knowledge
LC-MD
Data Planning
Data Generation or Collection
Data Featurization
Data Preprocessing
Data Exploratory Analysis
Data Application
LC-QG
Definition
Traceability Governance
Completeness Governance
Time-sensitivity Governance
Consistency / Accuracy / Redundancy Governance
Balance / Normalization Governance
Insight Governance
Maintenance
Objects / Quality Concerns / ...
Data / Result / ...
Raw data
c
Traceability
Completeness
Time-sensitivity
Consistency
Accuracy
Redundancy
Balance
Normalization
Insight
Candidate data-driven methods
classified and multi-granularity meta-data structure
limited ML models and missing value imputation methods
semi-automatic and full-automatic methods
irrelevant factors and related files check
value- and relationship-based outlier detection methods
redundant samples and features elimination methods
univariate discretization and cluster-based methods
value and dataset partition principles
unsupervised, supervised and automated ML models
provenance information
extrapolation of materials data
literature, patents and other knowledge carries
source and configuration files
constraints on values and patterns
unique characteristics of materials and influence from features
experical thresholds for screening
distribution of materials in the space of chemical properties
known laws or patterns
Materials domain knowledge
High-quality data

(caption on next page)

Fig. 2. Components of MAT-DQG. a, Data quality dimensions for ML in materials science. Inherent and contextual quality dimensions are indicated by green and blue boxes, respectively. Influence factors of different quality dimensions are represented by the words used as the starting point of the smallest arrows. **b**, The lifecycle of materials data quality governance for ML modelling, and its mapping relationships with the lifecycle of materials data of ML. The flow represented by cyan boxes is $LC - MD$ (lifecycle of materials data). The flow illustrated by blue boxes is $LC - QG$ (lifecycle of materials data quality governance). Mapping relationships between $LC - MD$ and $LC - QG$ are represented by the gray boxes. **c**, PMs for nine quality dimensions, where blue boxes mean the statistical methods for assessing and improving data quality and the red ones means materials domain knowledge-embedded approaches matching the statistical methods.

3.1. Lifecycle of materials data (LC-MD)

The general process of ML in materials science includes target identification, sample construction, model building and application stages [42,43]. According to different operations on materials data at different stages, LC-MD is proposed. As shown in Fig. 2**b**, it includes data planning, data generation or collection, data featurization, data pre-processing, data exploratory analysis, and data application. Materials domain knowledge plays different roles at each stage.

Data planning is the initial stage of LC-MD corresponding to the target definition stage of the ML general process. At this stage, researchers determine materials data acquisition by combining domain knowledge, such as materials type, target property, and data source. Subsequently, a raw materials dataset for ML is constructed after the *data generation or collection* and *data featurization* stages of LC-MD. At the former stage, the source files of materials data specified at the data planning stage are collected, such as materials structure and target property files. At the latter stage, parameters (termed features) influencing target properties specified at data planning stage are first appropriately determined according to physical meaning and then transformed into numerical values based on source files. Finally, a raw dataset that can be processed by computer for ML modeling is obtained. In general, descriptor selection is a step that involves most human intervention and has a shaping influence on ML model performance.

Data pre-processing is a critical stage to improve the quality of raw materials dataset for ML modeling. At this stage, many data quality issues are identified and dealt with by means of various data-driven cleaning approaches generally, such as missing values, outliers, and class imbalance [44]. Nevertheless, the values of descriptors or target properties always have physical restrictions in the context of materials knowledge, and it is difficult to automatically catch anomalies that break these restrictions for purely data-driven methods. Therefore, it is important to develop novel methods for materials data pre-processing by combining with materials domain knowledge [45].

Data exploratory analysis stage provides *data application* stage with guidance for final ML modeling. Due to complexity and specificity of materials issues, general-purpose ML models may be ineffective in many cases. For example, the generalization of some ML models will be poor with small-size, imbalanced, or too complicated materials data. [46,47] In this stage, the characteristics of the underlying pattern of materials data are explored with standard ML methods. These “characteristics” are not only helpful for the design of better ML models but also provide researchers a valuable opportunity to re-examine the rationality of target definition and data generation or collection under the guidance of domain knowledge.

3.2. Lifecycle of materials data quality governance

Different data operations at each stage of LC-MD induce that the data quality dimensions involved in each stage also differ. Identifying data quality dimensions at each stage of LC-MD is beneficial for targeted improvement of the reliability of data quality governance. To this end, LC-QG is defined by responding to LC-MD. LC-QG is divided into nine stages, including one definition stage, seven quality governance stages, and the final maintenance stage. Materials domain knowledge is involved at each stage of LC-QG. Finally, there exists a comprehensive mapping relationship between the various stages of LC-QG and LC-MD, enabling the synchronization of the machine learning process with the data quality governance process.

The execution order is shown in Fig. 2. Firstly, the *definition* stage should be finished before LC-MD starts. Its task is to determine data quality governance objects, dimensions, and so on. Herein, the taxonomy of governance objects is **master-** and **meta-data**. The former is used for ML modeling, and the latter is used to understand the physical meaning and quality of materials data, such as data type, coverage, source, and materials domain knowledge about target properties and features (e.g., descriptors). Once the data planning stage of LC-MD begins, the *traceability governance* is initiated and does not stop until LC-MD ends. It records information about the generation, collection, and processing of master- and meta-data. *Completeness governance* is required to evaluate and improve the integrity of dataset for ML modeling and crucial additional information (*e.g.*, provenance information), so it goes along with traceability governance to verify the integrity of provenance information. Subsequently, *data consistency*, *accuracy*, and *redundancy governance* are sequentially activated and executed. Specifically, data consistency ensures uniform value dimensions within each feature, optimizing the efficacy of downstream operations; Data accuracy verification validates the correctness of values and features, which in turn enables data redundancy governance to efficiently detect and eliminate redundant features and samples. Note that in any case these three governances must be finished before the data application stage of LC-MD, otherwise the results of ML model may be unreliable. *Time-sensitivity*, *balance*, *normalization*, and *insight* governance should also be finished in this period. Therein, time-sensitivity is finished at data generation or collection stage of LC-MD, because the time-related characteristics are always related to materials domain knowledge. Balance and normalization governance are finished at the data pre-processing stage of LC-MD. Insight governance stage is the last and started at data exploration analysis stage of LC-MD. Of particular notice is that these quality governance stages should cover all manipulations of traditional data pre-processing in theory, and the intermediate data generated from them should be fed back to the traceability governance. Moreover, they consider data quality evaluation, improvement, and validation based on materials domain knowledge.

Note that since available materials data represents a sample of the real world which constantly accumulates, the materials data quality detection becomes outdated as time passes. Therefore, maintaining a certain level of materials data quality cannot be limited to a one-shot approach. The maintenance stage of LC-QG focuses on the update of materials data and ML results after materials data is first used to tackle materials issues. Updating stored information and re-doing related governance operations should be executed at this stage. For example, if the master-data is changed, it is necessary to restart DQDs governance processes.

4. Processing models (PMs) for materials data

Although various data-driven methods to address data quality issues are available, their reasonable selection remains challenging for material scientists. Without appropriate consideration of domain knowledge, certain methods can even lead to flaws in ML models. To this end, under the guidance of LCM, we construct nine content-rich PMs for materials data in the form of two-dimensional table, aiming to provide insights on *HOW* to govern materials data quality through rational combination of data-driven methods and materials domain knowledge. As shown in Fig. 2**c**, each PM corresponds to one of the nine quality dimensions described above.

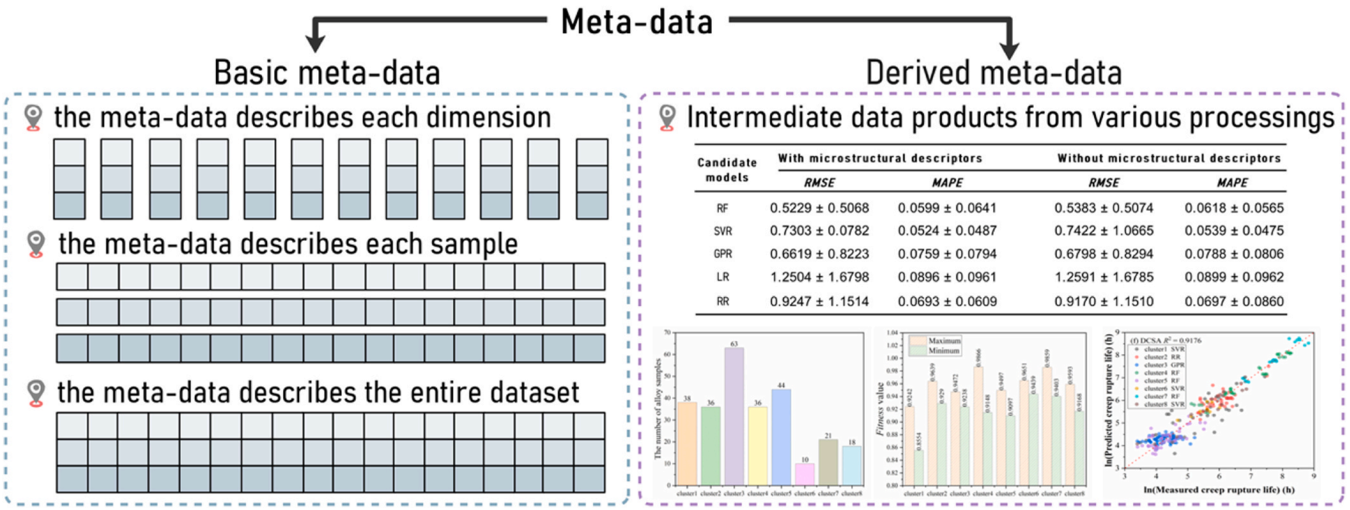
4.1. Traceability assurance

focus on characteristics of master-data itself (i.e., Basic meta-data) and its insight performance evaluated by ML models (i.e., Derived meta-data). For evaluating the credibility of ML results, it is important to clarify the content, sources, and acquisition way of the original master-data constructed at the early stage of LC-MD. Using the original master-data, lots of derived data will be generated subsequently, which may be the revised master-data from quality governance, or the hyperparameters and results of processing models. These derived data capture the processing stages of LC-MD and showcase the added value of master-data. Hence, to make it easy for researchers to verify each stage of LC-MD, it is necessary to make ML-oriented data acquisition and processing transparent to ensure the traceability of both original master-data and derived-data. The process of data acquisition and processing can be traced by establishing a classified and multi-granular meta-data structure and applying a provenance tracing mechanism. Note that the existing platform [48], ioChem-BD Platform, provides an insight of the setting of meta-data, based on the Dublin Core meta-data schema [49]. It captures not only the most basic bibliographic information about any digital asset but also records the descriptive information of quantum chemistry documents. However, such standards have failed to adequately account for the requirements of data services oriented towards ML. To this end, we divided the meta-data set in PM into basic meta-data and derived meta-data (i.e. the derived data), as shown in Fig. 3. The former refers to meta-data that describes the basic information of master-data, of which details can be seen in Section S1.1 in SI, while the latter refers to meta-data generated during the processing of master-data, which documents the analysis and application processes involved. Accordingly, the process of data acquisition and processing can be traced.

Moreover, for traceability of data processing, the scientific workflow (SWF) is a flexible tool for capturing scientific data and performing complex analysis on it [50]. There are many existing scientific data management systems for particular fields [50,51], and SWF provenance mechanism has increasingly become a core function of these systems [52]. Although the provenance mechanism of materials data has not been studied, some cross-domain SWF models may provide a promising solution, such as the cross-domain Content-rich and Fine-grained SWF Provenance Model (CF-PROV) propose by us [53]. It provides normative transformations and documentation declarations for multi-field SWFs, as well as can consolidate the traceability governance of master-data acquisition and processing. On the one hand, the derivation of master-data during various processing can be clearly depicted by the topology of the provenance graph of CF-PROV. On the other hand, the basic and derived meta-data (e.g., coding rule, model hyper-parameters, domain knowledge, visual tables and figures) can be embedded into the provenance mechanism of CF-PROV as extended attributes of the entity and activity.

4.2. Completeness protection

focuses on the completeness of master-data itself to ensure it can facilitate ML models to reveal real-world phenomenon accurately and its reusability and traceability (i.e., the completeness of meta-data). To this context, we believe that the completeness of both master- and meta-data need to be valued, which reflect value-level governance and the governance of its traceability. For meta-data, completeness can be maintained through a classified and multi-granular meta-data system. Therefore, the completeness protection of meta-data can be achieved by pre-defining the organizational structure and specific content of meta-data, and by real-time monitoring the generation and recording of meta-data during the processes of data acquisition, analysis, and processing. For master-data, the completeness of both features and values is equally important. Specifically, to ensure feature completeness, an effective approach is to leverage natural language processing (NLP) technologies to automatically collect features (e.g., descriptors) from literature [54,55]. Such approaches help mitigate human subjectivity and bias while enabling the collection of a relatively comprehensive set of highly relevant features, aligned as closely as possible with domain requirements; Following feature determination, any missing values or outliers (e.g., non-numeric values in numerical fields) are detected and corrected. For instance, Lin et al. [56] assembled an ML-oriented dataset after determining the descriptor combination but encountered issues with missing values that hindered accurate ML modeling. To address this, they employed a multiple imputation method [57] to impute the missing data. It is important to note that this data cleaning step occurs after the features of the master-data have been finalized. Moreover, limited ML models have been expanded for pattern recognition based on incomplete data such as artificial neural network, decision tree, support vector machine, XGBoost, and some fuzzy methods [58]. However, these models may exhibit poor prediction performance with missing values in key descriptors. Accordingly, missing imputation is another way to process the missing values of master-data. There are many



Candidate models	With microstructural descriptors		Without microstructural descriptors	
	RMSE	*MAPE*	*RMSE*	*MAPE*
RF	0.5229 ± 0.5068	0.0599 ± 0.0641	0.5383 ± 0.5074	0.0618 ± 0.0565
SVR	0.7303 ± 0.0782	0.0524 ± 0.0487	0.7422 ± 1.0665	0.0539 ± 0.0475
GPR	0.6619 ± 0.8223	0.0759 ± 0.0794	0.6798 ± 0.8294	0.0788 ± 0.0806
LR	1.2504 ± 1.6798	0.0896 ± 0.0961	1.2591 ± 1.6785	0.0899 ± 0.0962
RR	0.9247 ± 1.1514	0.0693 ± 0.0609	0.9170 ± 1.1510	0.0697 ± 0.0860

Fig. 3. Diagram of the composition of meta-data. Basic meta-data is employed to record basic information of master-data (i.e., feature, sample and entire dataset). Derived meta-data is employed to record the information generated from in a master- data processing project that records the analysis and application of the master-data.

general-purpose and purely data-driven methods, such as KNN, fuzzy c-means, rough K-Means, self-organizing maps, and probability density function-based methods, all with the caveat that simulated values in materials data may not exist in the real world. Therefore, the key is to test the rationality of generated data with materials domain knowledge, such as its empirical range and relationship with other data.

4.3. Time-sensitivity characterization

focuses on the featurization of time-sensitivity latent in master-data. On the one hand, *d*ue to the difference in materials domain knowledge, synthesis techniques, and representation methods, materials data generated in different periods may exhibit distinct characteristics. For example, nickel-based single-crystal superalloys have evolved through four generations. Although these alloys fall within the same class of functional materials, each generation exhibits marked variations in multiple critical aspects, e.g., chemical composition, phase constitution, heat treatment protocols, creep deformation behavior, and strengthening mechanisms [59]. On the other hand, depicting a timeline for each sample of master-data indicating its characteristics published at different periods can provide an intuitive way of capturing data time-sensitivity (e.g., time-series data). For example, Li et al. [60] found that to ensure the best state of battery health diagnostic accuracy, the data used for modeling should be recorded when the voltage reaches middle value in a charge cycle. Similar conclusion is observed in molecular dynamics [61], namely recording the error of on-the-fly simulation in real time can facilitate addressing the issues of data insufficiency and generalization in complex phase transition simulations. Moreover, the development history of materials reflects the exploration and discovery process of their property actuating mechanism, which can improve ML performances by providing prior guidance information for model selection and construction. Similarly, Meredig et al. [62] and Ling et al. [63] advocated to use potential patterns of previous materials data as the test set to evaluate the generalization ability of ML models in novel material discovery. It is worth noting that time-sensitivity cannot be used as an absolute indicator to evaluate the quality of materials data, rather than reflect the time-related characteristics of data quality. Therefore, time-sensitivity governance aims to find time-related characteristics as early as possible, so that ML tasks can be designed more rationally for material problems.

In the past, time-sensitivity characterization was usually executed manually and relied heavily on domain knowledge of materials experts. But now, statistical or ML technology promises to make this work semi-automatic, even full-automatic. *Semi-automatic way* refers to assistance to materials experts in identifying time-related characteristics that are not easily observed directly by using data-driven methods. For example, Bandt et al. [64] proposed permutation entropy to measure the complexity of time series in a simple, robust, and computationally efficient way. Based on this, Pessa et al. [65] proposed a simple and open-source Python module (named *ordpy*) that implements permutation entropy and several of the principal methods to analyze time series data. Through this way, the implicit rules of materials data are extracted in advance by means of clustering or association rule mining methods, then materials domain knowledge is used to verify whether there is a mapping between the potential pattern of materials data and time. *Full-automatic method* automatically establishes relationship between the potential pattern of materials data and time using data-driven methods. NLP and text mining techniques are promising. They can automatically process unstructured text, as a proxy for the accumulated domain knowledge, and learn underlying patterns. For example, Tshitoyan et al. [66] successfully developed a method to identify promising thermoelectric materials by introducing time factors and employing NLP techniques. Nie et al. [67] developed a knowledge graph named MatKG through NLP techniques and successfully constructed a graph for $LiFePO_4$ used for lithium-ion batteries.

4.4. Consistency check

focuses on the detection of uniform representation approach of master-data. Effective ML modeling requires that the “irrelevant factors”, which influence target properties but not explored, are as consistent as possible. Therefore, the consistency detection of master-data checks whether all samples are uniformly represented in the same way. Samples whose “irrelevant factors” are not expected values need to be regenerated or retrieved or even removed from master-data. For example, if an ML project is only to study the relationship between component- and structure-related descriptors and target properties, then all samples should be captured under the same conditions including experimental environment, measuring instrument parameters, coding rules, and so on. Beyond consistency of the samples, the consistency detection of the information among files will ensure that data is correctly identified and utilized in future.

4.5. Accuracy improvement

focuses on the outlier from value-based (i.e., data value of single feature and data values of associate features) and relationship-based (i.e., relationships among features and feature and target property) perspectives. Noise from the process of data acquisition results in outliers, which heavily increases the aleatoric uncertainty. Hence, uncertainty quantification should be valued and can be employed to assist the detection of outliers. Existing statistical techniques generally detect outliers from structured data as data points deviating from the global data distribution. However, the materials data quantity is typically smaller than that in other fields, which may lead to the inaccuracy of outlier detection by traditional data-driven techniques. Thus, the incorporation of materials domain knowledge about the relationship among data becomes necessary. Hence, a classified and layered accuracy improvement system is shown as Fig. 4, which can help researchers to quantify the aleatoric uncertainty, meanwhile, design and organize various outlier detection methods based on values and relationships.

For the accuracy governance of structured data values, single- and multi-dimensional outlier detection methods are proposed in ML community, which identify “outlier-values” in single-dimensional data space and “outlier-samples” in multi-dimensional data space. In single-dimensional data space, data is considered abnormal if certain descriptors exhibit unreasonable data types and value ranges. Single-dimensional outlier detection measures whether a given single-dimensional datum is reasonable. Similarly, the accuracy governance of structured data relationships is a process dominated by materials domain knowledge and assisted by data-driven methods, because meaningful relationships only exist in the context of domain knowledge, such as the importance of descriptors, the dependency among descriptors, and the law between descriptors and target property. Data-driven correlation analyses such as Pearson Correlation Coefficient (PCC) [68], Spearman Correlation Coefficient (SCC) [69], and Maximal Information Coefficient (MIC) [70] provide good assistance to capture these relationships. Moreover, LLM is gradually introduced into the improvement of structured data accuracy due to its capability of one-shot or few-shot context [71,72]. For example, Narayan et al. [73] casted five data cleaning and integration tasks as prompting tasks and evaluated the performance of LLMs on these tasks, of which results showed that LLMs generalized and achieved state-of-the-art performance on data cleaning and integration tasks, even though they are not trained for these data tasks (i.e., zero-shot context). However, for unstructured materials data, representation learning plays key roles in improving its accuracy, which transforms “unstructured” into “structured” for ML modeling. In this process, values and relationships can be mapped to reasonable space.

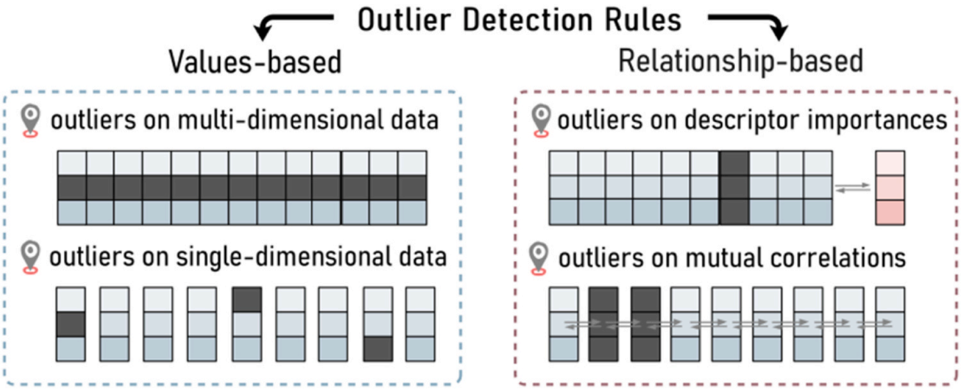

Outlier Detection Rules
Values-based
Relationship-based
outliers on multi-dimensional data
outliers on single-dimensional data
outliers on descriptor importances
outliers on mutual correlations

Fig. 4. Cla**ssified and layered structure for designing and organizing all outlier detection rules.** The dark grey square represents the outlier-values or samples. The pink square refers to the target property of interest.

4.6. Redundancy elimination

focuses on the redundancy of features and samples of master-data. Note that good representations are expressive, meaning that a reasonably sized learned representation can capture a large number of possible input configurations [74], namely high correlation among features (i.e., redundant features) can have a negative impact on ML modeling. Hence, eliminating redundant features is the goal of feature engineering for structured data [75]. Though there are many mature feature selection (FS) and transformation methods widely used in materials science, most of them are purely data-driven and ignore the prior knowledge about descriptor relative importance. Hence, it is a key issue to develop FS or feature transformation methods with the incorporation of materials domain knowledge. For example, we proposed a multi-layer FS method incorporating domain expert knowledge, where the importance of descriptors is introduced into the feature selection process via domain knowledge to ensure key features to be accurately selected [76]. It not only alleviates the risk of removing key features in the context of materials domain knowledge but also preserves the predictive precision of ML models in terms of data itself. Moreover, we transferred the materials domain knowledge for the relationships between descriptors into Non-Co-Occurrence Rules (NCOR) and embedded NCOR into the process of FS [77] and proposed a feature selection method (NCOR-FS) to reduce correlations among features by embedding domain knowledge. The correlation between various factors affecting the ion transport performance of solid electrolyte is transformed into NCOR and embedded into the objective function of the feature selection. Moreover, it is effective to design high-performance descriptors or representation approaches, to conquer the issues of redundant features, of which details can be seen in Section S1.3 of SI.

For sample redundancy, it refers to the repeated occurrence of eigenvectors representing the same entity in master-data. The naive rule for recognizing and eliminating redundant samples is that if eigenvectors of several samples are identical, then only one of them is retained. In this case, the simple complete matching can be used. However, in most cases, one entity is represented by inequivalent eigenvectors in different sources. Therefore, redundant sample recognition is knowledge-intensive and domain-specific. Moreover, both *semantic-* and *data-level similarity* should be considered to recognize redundant samples when defining redundancy entity identification rules based on domain knowledge of materials. In MAT-DQG, similarity measurement at semantic-level can be done by evaluating the similarity of some sample-grained meta-data and related materials domain knowledge. For example, if external conditions are consistent, the samples with chemical formula $Na_6Zr_{12}P_{18}O_{72}$, $NaZr_2P_3O_4$, and $NaZr_2(PO_4)_3$ are corresponding to the same compound. For similarity measurement at data level, the recognition for redundant samples can be conducted via similarity calculation among samples on key descriptors (i.e., it is easy to distinguish between entities). Herein, data-driven similarity measurement methods are useful, such as sorted neighborhood [78], or fuzzy duplicate elimination [79]. More rigorously, different recognition rules have different confidence degrees. To this end, the rules can be represented by “IF-THEN” logic with confidence coefficient [80], which can be seen in Section S1.3 in SI.

It is important to note that different material target properties correspond to distinct prediction tasks, and varying feature representation methods (e.g., different descriptor combinations) or target properties can directly impact ML models used for revealing structure-activity relationships. To this end, we primarily focus on sample redundancy within individual datasets, ensuring both traceability and reusability of each dataset. For example, Evans et al. [81] constructed datasets with two different descriptor combinations (31 and 33 descriptors) and two properties (i.e., Bulk moduli and Shear moduli), which can be regarded as four different datasets for two tasks, i.e., prediction for Bulk moduli and Shear moduli, instead of two datasets for one task. This is because ML models can construct certain relationships between one descriptor combination (or features) and one target property only.

4.7. Imbalance discovery and remediation

focuses on modifying imbalance distribution latent in master-data. Clarifying the category of samples is a prerequisite to measure whether master-data has an imbalanced distribution on sample space, but sample category may be implicit. Ideally, the materials domain knowledge is expected to be used to classify samples into reliable categories. For example, ionic conductivity $10^{-4}Scm^{-1}$ can be used to screen superionic and non-superionic structures of solid lithium-ion conductor materials [3]. The band gap energy range of $0.9 \sim 1.7$eV can be used to distinguish promising and non-promising solar cell materials [82]. However, since the property actuating mechanisms are typically complex, it is difficult to determine sample category for materials experts in most cases. At this point, data-driven discretization techniques can be used to generate prior categories for materials experts. *Univariate discretization methods* [83] can classify samples by dividing target property values into several discrete intervals. Clustering-based methods [84] are popular *multivariate discretization methods*, they classify samples by measuring similarities among samples in multi-dimensional feature space. To reduce or even eliminate the imbalance distribution of samples in a fixed feature space, the optimal way is to generate or collect more data. For example, Niblett et al. [85] obtained 200 configurations of a liquid from molecular dynamics simulations with OPLS-AA force field covering states at different temperatures and pressures to ensure the diversity of training dataset and incorporated the near-equilibrium configurations of 200 isolated ethylene carbonate molecules and 400 isolated ethyl methyl carbonate molecules to enhance the ML models’ sensitivity to intramolecular interactions, based on the anticipation of potential deficiencies in the models. Moreover, class imbalance is a hot topic with long research history in ML community [86]. A considerable amount of data-driven approaches without adding new samples have been developed to reduce negative impact of data imbalance on modeling in different domains [87], which is promising to be applied in materials science.

4.8. Data normalization

focuses on master-data value and dataset partition. For value normalization, ML results are affected by the measurement units of descriptors. In general, descriptors with smaller units have a larger range of values, and they tend to have a larger influence on target properties. Value normalization attempts to eliminate this bias for ensuring that ML model objectively reflects the influence of descriptors on target properties, and it also facilitates ML algorithms to converge [88]. Value normalization methods require that values are scaled according to transformation function with invariant relativity among values, such as min-max normalization scales or zero-centered normalization. For example, Schütt et al. [89,90] observed that the neural network can be more stable when normalizing the filter response by the number of atoms within the cutoff range. Geiger et al. [91] constrained spherical harmonics to a maximum value of 1, and initialized the weights as some normalization constant multiplying a learned parameter initialized randomly with a normalized Gaussian, to ensure the GNN accurately capture the equivariance of materials. Based on this, Batatia et al. [92] employed this normalization strategy for the implementation of the one-particle basis. Meanwhile, to construct the ML model with high performance, master-data should be divided into training, validation, and test datasets. Training dataset is used to train ML models for learning potential patterns of master-data. Model selection and hyper-parameters determination are performed on validation dataset. Test dataset is used to assess the generalization of selected and optimized ML models. There are two pitfalls to be avoided, i.e., *information leakage* and *information missing*. To solve the former, it should be ensured that no same or extremely similar samples appear in training and validation (test) datasets simultaneously. Otherwise, the generalization of ML models would be exaggerated. The latter is solved by preventing dissimilar or mutually exclusive samples that always appear separately in training and validation (testing) datasets. Otherwise, ML results would be biased because of insufficient information diversity of training dataset.

4.9. Insight exploration

focuses on independent-dimension (i.e., single feature and single relationship between one feature and other features or target properties) and joint-dimension (i.e., multiple influence of features on target properties) insights. For the first one of *independent-dimension insight*, there are two suggestions: It is beneficial to understand data distribution in terms of single-dimensional values through typical statistical measures (e.g., medians, means, quantiles), and introducing some visual diagrams (e.g., box plot, violin plot, frequency distribution histogram). Novelty analysis is the topic related to outlier detection and allows to share related techniques, which aims at detecting previously unobserved novel patterns from data [93]. Univariate novelty analysis is used to find distinctive values of crucial descriptors, and it is of importance to explore the reason why these values emerge based on materials domain knowledge. For the second one of *independent-dimension insight*, there are two suggestions: Analyzing correlation between single descriptor and one target property by two-dimensional scatter plots, which is also a common method used by materials experts [94]. However, the scatter plots are hardly adequate when the relationship is complex. Correlation analysis techniques can explore more complex relationships, such as PCC, SCC, MIC, linear regression, and polynomial fitting. Redundant or key descriptors can be found based on these techniques, then used to guide feature engineering or the construction and interpretation of ML model as prior knowledge. For the *joint-dimension insight*, exploring it is useful to understand the characteristics of potential patterns in materials data by using some easy-to-operate ML models, for finding the optimal ML model more rapidly. Three suggestions should be considered collectively: *unsupervised methods* [95], (*semi-*) *supervised methods*, and *automated ML* [96,97]; see Section S1.4 in SI for details.

5. Case study for structured materials data

5.1. Dataset

The model performance can vary substantially with different tasks. To this end, we collected 60 various materials datasets from literature to evaluate the effectiveness of MAT-DQG, with details provided in Section S2.1. Fig. 5**a** shows the category distribution of collected datasets, with a detailed description of leaf nodes. It can be seen that most researchers only focus on the dimensions of "accuracy", "insight", and "redundancy" (Fig. 5**b**). To further present the details of data quality governance, 7 representative materials datasets exhibiting different properties and data characteristics are selected (details can be found in Table 1 and Section S3 of SI). Finally, details of the data governance process of NASICON-type solid electrolyte material are presented here as the case study. Migration energy barrier reflects ion transport properties of solid electrolytes. Hence, this study leverages high-throughput screening platform for solid electrolytes (SPSE) [37,98] and its calculation program of Bond Valence Site Energy (BVSE) to obtain the values of energy barriers which is regarded as the target property.

5.2. Evaluation metrics

To comprehensively evaluate the ML models, this study employed Root Mean Square Error (RMSE) and R^2 to measure the difference between predicted and true value, as shown in Eqs. (7) and (8). Thereinto, *RMSE* is sensitive to maximum or minimum error in a set of predicted values. R^2 can reflect the model fitting degree of predicted values. To simplify the results, R^2 is employed to evaluate the performance of ML models, the results of other metrics can be seen in Section S3 in SI.

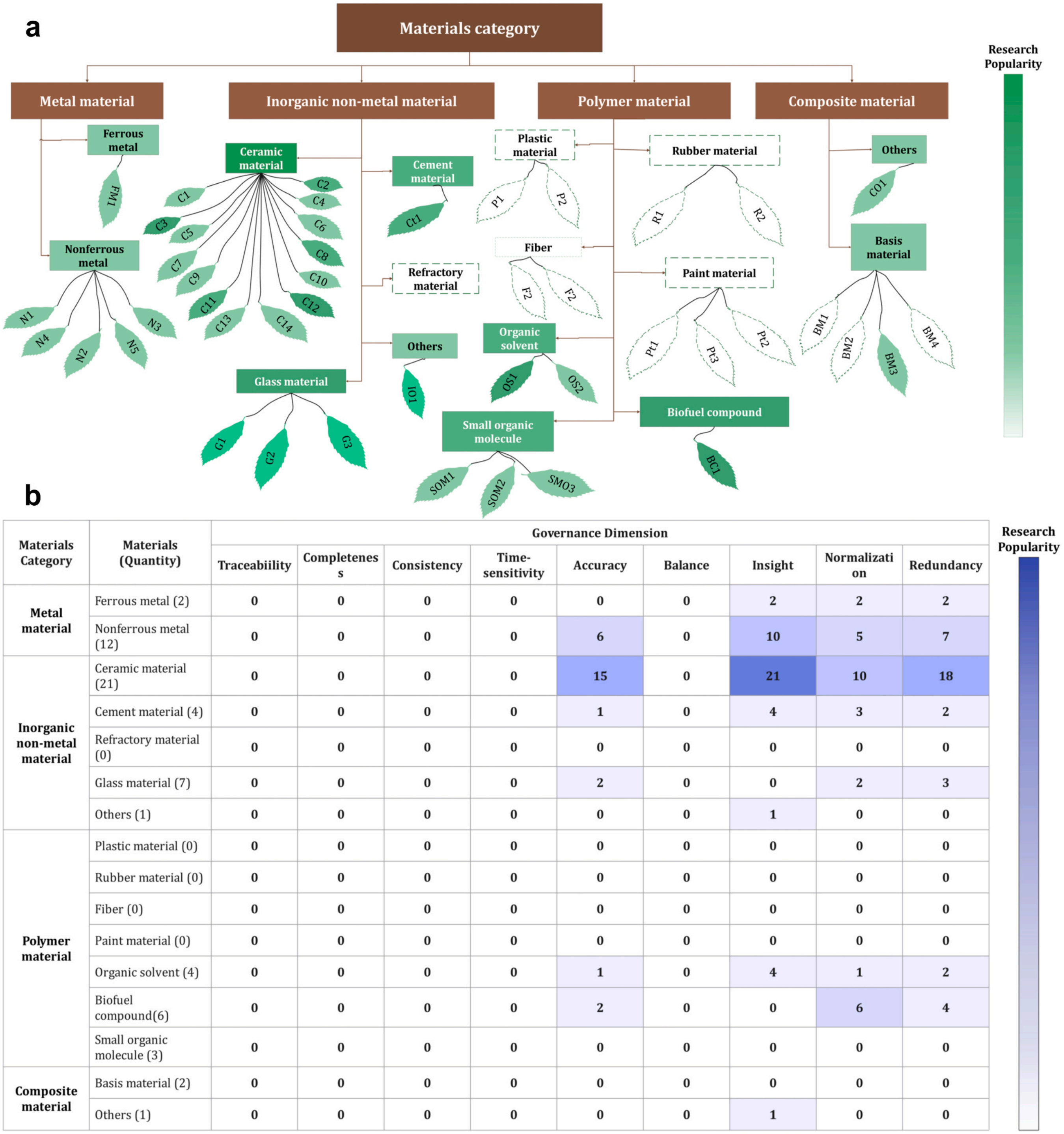
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i' - y_i\right)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i' - y_i\right)^2}{\sum_{i=1}^{n}\left(\overline{y} - y_i\right)^2} \tag{8}$$

where y_i means true value of sample i; y_i' means the predicted value of sample i; $\overline{y}$ is the average of y.

5.3. Model selection

The performance of different ML models differently depend of the materials data quality [97]. Hence, we here select 7 common ML models, of which prediction performances are employed to evaluate the quality of governed dataset, i.e., Multiple Linear Regression (MLR) [104], Ridge Regression (RR) [105], Least Absolute Shrinkage And Selection Operator (LASSO) [106], Support Vector Regression (SVR) [107], K-Nearest Neighbour (KNN), Gaussian Process Regression (GPR) [108], and Random Forest (RF) [109]. Thereinto, the kernel functions of SVR and GPR are set as radial basis function. All these ML models are implemented in Python in the Scikit-learn toolkit [110] and Bayesian optimization algorithm (BOA) [111] is employed for hyperparameter optimization of ML models above, of which details of parameter setup can be seen in Table 2. Meanwhile, each materials dataset is divided into training set and testing set with the ratio of 8:2 randomly, and the models are trained on training set by 10-fold cross-validation, and then



Materials Category	Materials (Quantity)	Governance Dimension								
		Traceabiility	Completeness	Consistency	Time-sensitivity	Accuracy	Balance	Insight	Normalization	Redundancy
Metal material	Ferrous metal (2)	0	0	0	0	0	0	2	2	2
	Nonferrous metal (12)	0	0	0	0	6	0	10	5	7
Inorganic non-metal material	Ceramic material (21)	0	0	0	0	15	0	21	10	18
	Cement material (4)	0	0	0	0	1	0	4	3	2
	Refractory material (0)	0	0	0	0	0	0	0	0	0
	Glass material (7)	0	0	0	0	2	0	0	2	3
	Others (1)	0	0	0	0	0	0	1	0	0
Polymer material	Plastic material (0)	0	0	0	0	0	0	0	0	0
	Rubber material (0)	0	0	0	0	0	0	0	0	0
	Fiber (0)	0	0	0	0	0	0	0	0	0
	Paint material (0)	0	0	0	0	0	0	0	0	0
	Organic solvent (4)	0	0	0	0	1	0	4	1	2
	Biofuel compound(6)	0	0	0	0	2	0	0	6	4
	Small organic molecule (3)	0	0	0	0	0	0	0	0	0
Composite material	Basis material (2)	0	0	0	0	0	0	0	0	0
	Others (1)	0	0	0	0	0	0	1	0	0

Fig. 5. Statistical information of 60 datasets. a, the category distribution of 60 materials dataset. **b,** distribution of data quality governance in original research.

testing set is employed to validate the stability and robustness of ML models.

5.4. MAT-DQG for 60 materials datasets

Here, MAT-DQG is employed to evaluate and improve the quality of 60 materials datasets. The execution order of each evaluation dimension is guided by MAT-DQG's LCM, and the process iterates continuously until no quality issues are detected. For the step of accuracy improvement, three outlier detection methods are selected, i.e., box diagram method, Isolated Forest (IF) and Local Outlier Factor (LOF). For box diagram method, the interquartile spacing IQR is set as 1.5 and data points outside the range of 5–95 % are detected as outliers. For IF, the proportion of outliers (contamination) is set as 0.1, which runs 10 times randomly. For LOF, the nearest neighbour is set at 20. All other parameters in the three methods are kept as the default. All other parameters in the three methods are kept as the default. For redundancy elimination, the experimental setup of Non-Co-Occurrence Rules Feature Selection (NCOR-FS) is the same as the original study [77]. The detection methods of each dimension, except Insight, for 60 datasets are shown in Table S6. Fig. 6 shows the overview of the governance process of MAT-DQG on all datasets. It is worth noting that most of the datasets

Table 1
Details of seven representative datasets.

Material Types	Materials (No.)	Target property	Acquirement Manner	# of Samples	# of Descriptors	Ratio	Descriptor Type	Ref.
Nanocomposite solid polymer electrolyte	Nanocomposite solid polymer electrolyte (**MD1**)	Ionic Conductivity	Experiment	160	5	< 0.25	1–2–4	[77, 99]
Inorganic non-metal material	Zeolite (**MD15**)	Bulk Moduli	DFT calculations	121	33	0.25 ~ 0.5	1–2–4	[81]
	Cubic Li-argyrodites (**MD30**)	Energy Barrier	BVSE-based calculations	51	32	0.5 ~ 1	1–2–3–4	[100]
Metal material	Ni-based SX superalloys (**MD8**)	Lattice Misfit	Literature	136	16	< 0.25	1–3–4	[101]
	High-entropy alloys (**MD33**)	Solid solution strengthening	Experiments	162	59	0.25 ~ 0.5	1–2–4	[102]
Ionic liquid binary mixtures	Ionic liquid binary mixtures (**MD18**)	Density	Literature	405	3	< 0.25	2–3–4	[103]
Inorganic non-metal material	NASICON-type solid-state electrolyte (**MD29**)	Activation energy	Experiment	85	45	0.5 ~ 1	1–2–3–4	[77, 99]

* "#" represents number. "Ratio" means the ratio of samples and descriptors. Descriptor Type: (1-Structure; 2-Properties; 3-Process; 4-Performance). "SX" means single crystal.

Table 2
The setup of BOA for ML models.

Model	Hyperparameters	Optimization Range
RR	α	(0.01, 10)
LASSO	α	(0.0005, 1.0)
SVR	γ	(10^{-5}, 1)
	C	(10^{-4}, 500)
KNN	Number of neighborhoods	(2, 30)
	p	(1, 8)
GPR	γ	(10^{-4}, 1)
	C	(10^{-5}, 10^{5})
RF	Number of estimators	(10, 300)
	Minimum samples split	(2, 15)
	Maximum features	(0.01, 0.999)
	Maximum depth	(3, 20)

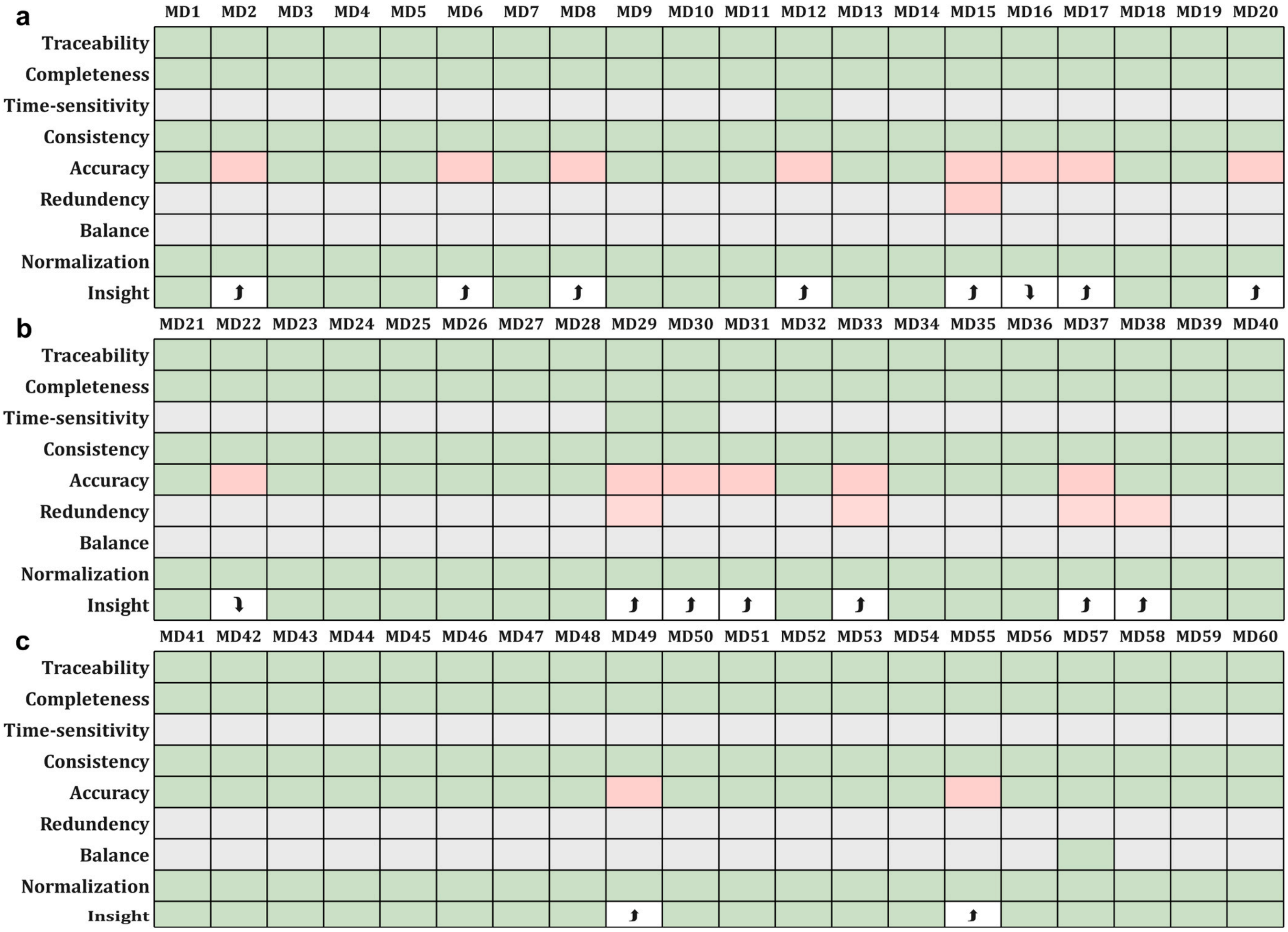
are collected from peer-reviewed published works, thus their quality is comparatively high and there are no issues in the dimensions of traceability and consistency. As different data distribution may affect the prediction performance of ML models, thus the models of insight detection are shown in Table S7. Fig. 7 illustrates the insights of 60 datasets with and without execution of MAT-DQG, where 17 datasets are identified as problematic and the insights (namely the best performance of ML model) of 16 datasets, except MD22 and MD16, gain improvement after MAT-DQG. This may be because, although these two revised datasets maintain compliance with materials domain knowledge, its expanded data distribution surpasses the modelling capacity of conventional shallow ML algorithms. Meanwhile, we observe that MD29 gains impressive improvement of insight. Actually, this dataset is constructed from scratch by us, i.e., collecting relevant CIFs, preliminarily defining the descriptor combination and calculating them for obtaining usable data. Therefore, the insight of this raw dataset is less desirable. Then, after MAT-DQG, there are three CIFs with abnormal temperature and 26 redundant features are detected. Through governance, MD29 possesses the tidy descriptor combination and accurate descriptor values, which enable these shallow ML models to accurately learn the general patterns latent in the revised dataset. More details of governance of MD29 can be seen in the **Example of NASICON-type solid electrolyte** section**.** With MAT-DQG, quality of the datasets could be increased, and performance of the ML models further improved, compared with original works.

5.5. MAT-DQG for 7 Representative materials datasets

To illustrate the governance details, we here select 7 diverse and representative materials datasets from 60 materials datasets, of which results of issue detection and governance process are presented in Table 3. Since the datasets, except MD29, are collected from the literature, their traceability, consistency and completeness can be guaranteed. To probe time-sensitivity, MD8 is from 4 different literature sources published between 1998 and 2015, with a large time span and covering four generations of nickel-based single crystal superalloy. Therefore, MD8 can be defined as a problem of time sensitivity and is necessary to reasonably divide the dataset according to the characteristics of periods. In contrast, other datasets exhibit no significant characteristics variation over the analyzed period, making it appropriate to exclude time factors from subsequent analysis. Through three-dimensional data accuracy detection, abnormal samples are detected from MD8, MD15, MD29, MD30, then we perform removement for these samples, to ensure low disturbance in the datasets. Moreover, two outlier points are detected in MD30, and we correct their values according to Ref. [100]. Then, redundant descriptors are further detected and removed from MD15, MD29 and MD33, which shows that there are latent data quality issues even in the published datasets. Table 4 illustrates the average accuracy of data quality governance in different datasets. From the prediction performance of ML models, most of the ML models gain significant improvement in predictive accuracy. However, several ML models in some datasets fail to learn the latent patterns after data quality governance, especially MD33. This phenomenon results from the fact that the distributions of the datasets have greatly changed due to the operations of data point modification and sample exclusion. ML data-driven that purely data-driven ML models greatly depend on the data, thus the incorporation of domain knowledge is necessary for the capture of important features in ML models. No issues are detected for the MD18 and MD1 and the prediction results based on them. It is obvious that MAT-DQG can accurately detect and modify the abnormalities latent in different materials datasets, which motivates the construction of high-precision ML models. The details of MAT-DQG for these 7 datasets can be seen in Section S3.1~S3.7 in SI.

5.6. Example of NASICON-type solid electrolyte

NASICON-type solid electrolytes have been widely studied in the field of electrochemical energy storage and conversion due to their good thermal stability, chemical stability, and simple and rapid synthesis process. As one of the key indicators of ion transport performance of solid-state electrolyte materials, accurate and efficient prediction of energy barrier can accelerate the discovery process of novel solid state electrolyte materials. Here, we take the prediction of NASICON-type solid electrolytes energy barrier as an example to present the governance details of MAT-DQG, namely under the guidance of LCM, this project has implemented the governance of nine DQDs of NASICON dataset based on PM. The best model (RR with R^2 of 0.961) trained on



a

	MD1	MD2	MD3	MD4	MD5	MD6	MD7	MD8	MD9	MD10	MD11	MD12	MD13	MD14	MD15	MD16	MD17	MD18	MD19	MD20
Traceability	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Completeness	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Time-sensitivity	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Consistency	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Accuracy	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟥
Redundency	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	⬜	⬜	⬜	⬜	⬜
Balance	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Normalization	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Insight	🟩	↑	🟩	🟩	🟩	↑	🟩	↑	🟩	🟩	🟩	↑	🟩	🟩	↑	↓	↑	🟩	🟩	↑

b

	MD21	MD22	MD23	MD24	MD25	MD26	MD27	MD28	MD29	MD30	MD31	MD32	MD33	MD34	MD35	MD36	MD37	MD38	MD39	MD40
Traceability	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Completeness	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Time-sensitivity	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Consistency	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Accuracy	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟩
Redundency	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	⬜	⬜	⬜	🟥	⬜	⬜	⬜	🟥	🟥	⬜	⬜
Balance	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Normalization	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Insight	🟩	↓	🟩	🟩	🟩	🟩	🟩	🟩	↑	↑	↑	🟩	↑	🟩	🟩	🟩	↑	↑	🟩	🟩

c

	MD41	MD42	MD43	MD44	MD45	MD46	MD47	MD48	MD49	MD50	MD51	MD52	MD53	MD54	MD55	MD56	MD57	MD58	MD59	MD60
Traceability	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Completeness	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Time-sensitivity	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Consistency	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Accuracy	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩
Redundency	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Balance	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	⬜	⬜	⬜
Normalization	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Insight	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	↑	🟩	🟩	🟩	🟩	🟩	↑	🟩	🟩	🟩	🟩	🟩

Fig. 6. Overview of MAT-DQG for governing 60 materials datasets. **a**, MD1∼MD20; **b**, MD21∼MD40; **c**, MD41∼MD60. The red rectangle indicates that the dataset contains anomalies in this dimension. The green rectangle signifies that the dataset is free from anomalies in this dimension. The gray rectangle denotes that the dataset is not included in the evaluation for this dimension. The arrow symbolizes an enhancement in the dataset's insights following the application of MAT-DQG.

the revised data has 49 % higher predictive performance than the best model (RF with R^2 of 0.643) based on the raw data. Details and results of each process for quality governance are recorded as derived meta-data. It is worth noting that NASICON dataset is free of inconsistency, imbalance, or time-sensitivity issues, indicating that not each dataset in practice will have issues in every quality dimension. Nevertheless, researchers should still be aware of the nine quality dimensions and provide evidence for the issue-free ones. The quality of NASICON materials data is assessed and / or improved from all nine dimensions according to the corresponding processing model, of which details are as follows.

5.7. Traceability Assurance and Completeness Check

Under the guidance of materials domain knowledge relevant to migration energy barriers, we construct a raw dataset for ML modeling consists of 45 descriptors and 90 energy barrier data with no missing values, which are based on Crystallographic Information Files (CIFs) of 90 NASICON compounds under $R\overline{3}c$ symmetry with general formula $Na_xM_2(XO_4)_3$ $(M = Zr^{4+}, Sc^{3+}, Ti^{4+}, \cdots; X, P^{5+}, Si^{4+}, Ge^{4+}, Mo^{6+}, \cdots)$ from the Inorganic Crystal Structure Database (ICSD). It is worth noting that experimental measurement of the energy barrier is costly and time-consuming, which limits its application in screening superionic conductors with excellent ion transport property [98]. Hence, the energy barrier of 90 NASICON compounds is calculated by the Bond Valence Site Energy (BVSE) method, as implemented in the SPSE platform. The information about raw dataset generation is provided in Table S2 and Table S3.

5.8. Time-sensitivity Capture

To the best of our knowledge, there is no evidence that NASICON compounds have significantly different characteristics over the covered period, making it reasonable to exclude time factors from subsequent data analysis.

5.9. Consistency Detection

In the raw dataset, all component- and structure-based descriptors as well as the barrier energy for all compounds are calculated using identical calculation procedures and parameters, and the experimental temperature is taken from CIFs. Ideally, all external variables in structure measurement experiments should be considered. As far as we know, the influences of variables other than the main ones are not recorded or published by any experimentalist. Furthermore, it is almost impossible to control all of them to be the same, especially given the existence of uncontrollable factors, such as measurement device precision. Therefore, we believe that the raw dataset has acceptable consistency, similar to other related work based on computational or experimental data.

5.10. Accuracy Improvement

Correctness of CIFs is a fundamental factor affecting data accuracy,

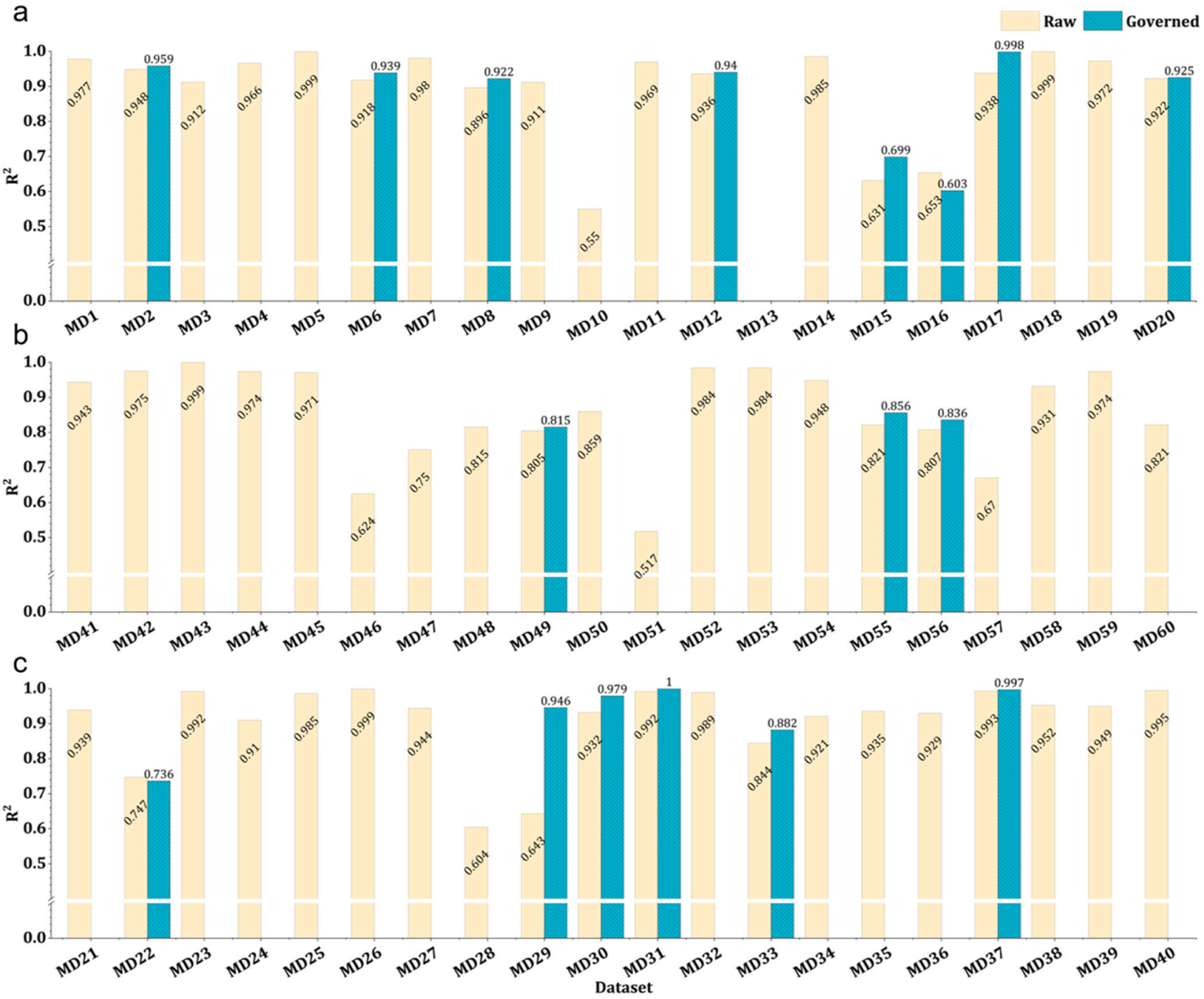


Fig. 7. The insight of 60 datasets before and after data quality governance via MAT-DQG (average R^2 of the best ML models with 10 cross-validation). **a**, MD1~MD20; **b**, MD21~MD40; **c**, MD41~MD60.

Table 3
The results of data quality governance for different datasets.

Dataset Dimensions	MD1	MD8	MD15	MD18	MD29	MD30	MD33
Traceability	√	√	√	√	√	√	√
Completeness	√	√	√	√	√	√	√
Time-sensitivity	-	√	-	-	-	-	-
Consistency	√	√	√	√	√	√	√
Accuracy	√	Deleting 4 samples	Deleting 3 samples	√	Deleting 5 samples	Modifying 2 points, deleting 1 sample	√
Redundancy	-	-	Deleting 23 redundant features	-	Deleting 26 redundant features	-	Deleting 35 redundant features
Balance	-	-	-	-	-	-	-
Normalization	√	√	√	√	√	√	√
Insight	High quality	Model performance improved	Model performance improved	High quality	Model performance improved	Model performance improved	Model performance improved

* “-” represents this dataset fails to meet the definition of this dimension. “√” represents this dataset has been analyzed through data quality governance and there is no issue in this dimension.

so it is evaluated firstly. According to the materials knowledge that lattice parameters in crystalline compounds should not remain the same with varying temperatures, three abnormal samples are identified. As shown in Table 5, their structures are measured at different temperatures (25℃, 300℃, and 620℃, respectively), but the lattice parameters “a” and “c” in CIFs are reported to be the same (9.186 and 22.181 Å,

Table 4
The average R^2 of the best ML models with 10 cross-validation on different datasets during data quality governance.

Dataset	MD1			MD8			MD15		
Metrics Models	R^2_{RAW}	R^2_{ACC}	R^2_{REDUN}	R^2_{RAW}	R^2_{ACC}	R^2_{REDUN}	R^2_{RAW}	R^2_{ACC}	R^2_{REDUN}
MLR	0.551	√	√	0.836	0.836	√	0.510	0.558	**0.639**
RR	0.547			0.799	**0.814**		0.401	0.662	**0.699**
LASSO	0.976			0.839	0.838		\	0.308	**0.595**
SVR	0.546			0.915	0.915		0.547	0.551	**0.621**
KNN	0.977			0.786	**0.871**		0.308	0.474	**0.685**
GPR	0.952			0.895	0.892		0.631	0.660	**0.673**
RF	0.967			0.896	**0.922**		0.602	0.655	0.644
Avg	0.788			0.852	**0.870**		0.500	0.552	**0.651**
Dataset	**MD18**			**MD29**			**MD30**		
Metrics Models	R^2_{RAW}	R^2_{ACC}	R^2_{REDUN}	R^2_{RAW}	R^2_{ACC}	R^2_{REDUN}	R^2_{RAW}	R^2_{ACC}	R^2_{REDUN}
MLR	0.865	√	√	\	\	\	\		\
RR	0.816			0.313	**0.960**	**0.961**	0.873	0.851	0.857
LASSO	0.864			0.118	**0.951**	0.938	0.845	0.824	**0.873**
SVR	0.864			0.108	**0.925**	**0.960**	\	\	
KNN	0.931			0.108	**0.913**	**0.955**	0.872	0.841	**0.979**
GPR	0.938			0.073	**0.959**	0.948	0.876	0.876	**0.893**
RF	0.999			0.643	**0.966**	0.956	0.932	**0.957**	**0.979**
Avg	0.897			0.227	**0.946**	**0.952**	0.880	0.870	**0.916**
Dataset	MD33								
Metrics Models	R^2_{RAW}	R^2_{ACC}	R^2_{REDUN}						
MLR	0.775	√	**0.856**						
RR	0.888		0.867						
LASSO	0.816		**0.864**						
SVR	0.849		**0.881**						
KNN	0.844		**0.882**						
GPR	0.861		**0.868**						
RF	0.896		0.893						
Avg	0.847		**0.873**						

"\" represents this model gains very poor prediction performance. "√" represents this dataset has been analyzed through data quality governance and there is no issue in this dimension. R^2_{Raw} represents the coefficient of determination (R^2) of raw datasets. R^2_{ACC} represents the coefficient of determination (R^2) of datasets after accuracy governance. R^2_{REDUN} represents the coefficient of determination (R^2) of datasets after redundancy governance. "Avg" means the average of R^2 of all ML models. The bold font means the improvement of ML models after governance of this dimension.

Table 5
Outlier-values and its revised values.

ICSD Number	Formula	Temperature	a	c	Revised a	Revised c
15545	$Na_{24}Zr_{12}Si_{18}O_{72}$	25℃	9.186	22.181	9.198 [112]	22.210 [112]
15546	$Na_{24}Zr_{12}Si_{18}O_{72}$	300℃	9.186	22.181	9.199 [112]	22.470 [112]
15547	$Na_{24}Zr_{12}Si_{18}O_{72}$	620℃	9.186	22.181	9.199 [112]	22.706 [112]

respectively). Therefore, the values of 'a', c', and other structure-based descriptors are modified by comparing structural information in CIFs and source literature.

Box plot is used to detect outliers for each descriptor and barrier energy. As shown in Fig. 8**a**, there are 17 descriptors and one target property (barrier energy) with outlier-values (green dots outside the box). Only a few outlier-values are truly abnormal according to materials domain knowledge, while others are anomalies due to the small size but high diversity of the raw dataset. For example, on Na sites, Na (1) sites are mostly occupied with lowest energy, and Na(3) sites with high energy do not tend to be occupied during the migration process in most of NASICON compounds [113]. Unsurprisingly, in the collected compounds from ICSD database, only 1.12 % of samples have zero occupancy on Na (1) sites and 23.6 % of samples have non-zero occupancy on Na(3) sites. As a result, smaller values on $Occu\_Na(1)$ and larger values on $Occu\_Na(3)$ are identified as outlier-values through box plot in Fig. 8**b** and Fig. 8**c**. Outlier values on $Occu\_Na(3)$ affect the outlier detection on $Entropy\_Na(3)$ in Fig. 8**d**, because $Entropy\_Na(3)$ is calculated according to $Occu\_Na(3)$. Therefore, these values are normal because they are known to behave as expected, and others are explained in SI.

As shown in Table 6, there are five outliers in terms of the energy barrier. We find that these outlier values corresponded to the compounds with species mixing on Na sites (Cs^+ and K^+). These compounds violate the predefined general chemical formula. Furthermore, according to the materials domain knowledge that Cs^+ and K^+ are identified as skeleton ions by BVSE procedure, so their calculated values of energy barrier greatly deviate. As a result, these samples are removed from the raw dataset.

Finally, Local Outlier Factor (LOF) and Isolation Forest (IF) are used to detect outlier-samples in 46-dimensional space. LOF is a density-based method, which identifies outlier samples whose density is significantly different from that of surrounding samples and is suitable for detecting local outlier samples. IF is a partition-based method, detecting outliers based on how far a data point is from the rest of the sample, and detects global outliers. Herein, to avoid false detection, IF is

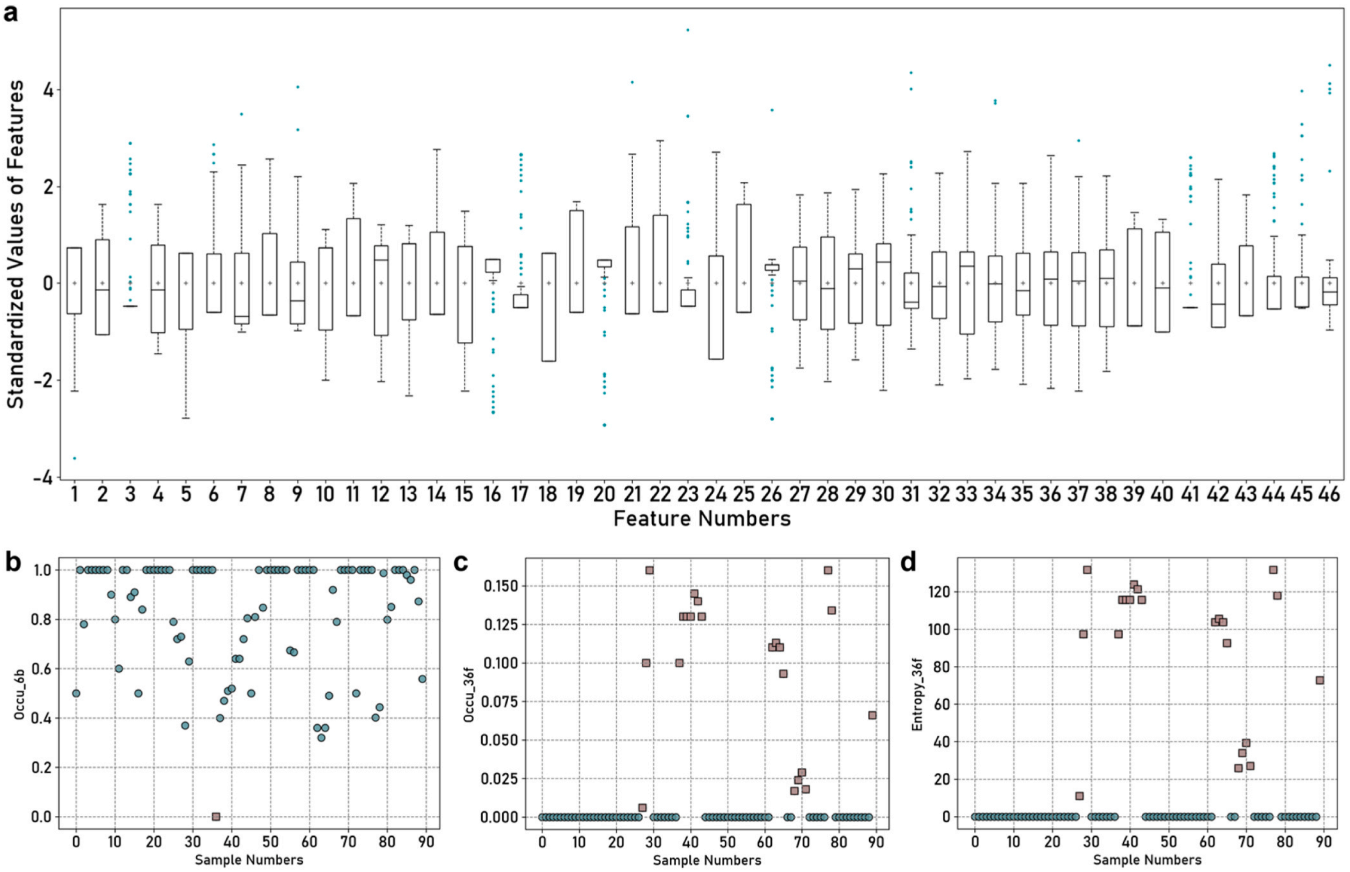


Fig. 8. Box plot for all descriptors and target property. The horizontal axis shows the number of descriptors (in Table S3), and the final feature represents target property, i.e., energy barrier. Vertical axis shows normalized values through zero-centered normalization. “+ ” and long “-” in the box represent mean and median, respectively. Short “-” outside the box indicate the lower or upper limits of normal. Green dots outside the box indicate outlier-values; **b**, **c**, and **d** are outlier-values distributions in Occu_Na(1), Occu_Na(3), and Entropy_Na(3), respectively. “Circles” and “Squares” represent non-outlier-values and outlier-values, respectively.

Table 6
The outlier-values on the energy barrier.

ICSD Number	Formula	Energy Barrier (eV)
167728	$Cs_{0.6}Na_{5.4}Zr_{12}P_{18}O_{72}$	4.0137
167729	$Cs_{1.2}Na_{4.8}Zr_{12}P_{18}O_{72}$	3.8867
167730	$Cs_{2.4}Na_{3.6}Zr_{12}P_{18}O_{72}$	3.9453
174461	$K_3Na_3Hf_{12}P_{18}O_{72}$	4.2676
250394	$K_3Na_3Ti_{12}P_{18}O_{72}$	2.8223

executed randomly 10 times, and the occurrence of each sample as outlier-samples are counted, then the top 10 most frequent samples are regarded as final outlier-samples. In total, there are 14 outlier samples, but no errors found by examining literature sources. This is likely due to the small size of data with high diversity. The final dataset with 85 samples for ML is determined to be free of inaccuracy issues.

5.11. Redundancy Elimination

Herein, a feature selection embedded with materials domain knowledge, named Non-Co-Occurrence Rules Feature Selection (NCOR-FS), is employed to accurately detect redundant features. Firstly, data-driven correlation analysis techniques and descriptor associations in material domain knowledge are used to obtain the non-co-occurrence relationship between descriptors and symbolize them as NCOR. Then, the non-co-occurrence rule violation calculation function of any descriptor set is established, and the prediction error evaluation function of the machine learning model is combined as the objective function of the feature selection method based on the optimization algorithm to evaluate the suitability of the descriptor subset. Considering the uncertainty of the impact of NCOR on the prediction performance of machine learning models, a two-stage evolutionary process is employed to optimize the process of feature selection methods. As a result, 26 redundant features are deleted. Table 7 illustrates the average R^2 (with standard deviation) of the various ML models on 10-fold cross validation before and after redundancy elimination, respectively. Besides MLR, the prediction performances of all ML models are improved to different extents. This confirms that despite the flexibility in ML algorithms, careful maintenance of data is still necessary. The analysis of screening thresholds is shown in Figure S21.

“\” represents this model gains very poor prediction performance.

5.12. Imbalance Discovery

K-Means is employed to divide 85 samples into $K(K = 8, 6, 4, 2)$

Table 7
Performance of ML models before and after redundancy elimination (mean ± standard-deviation).

Model	R^2 before redundancy elimination	R^2 after redundancy elimination
MLR	\	\
RR	0.959 ± 0.002	**0.960 ± 0.002**
SVR	0.915 ± 0.003	**0.933 ± 0.001**
GPR	0.959 ± 0.002	**0.960 ± 0.002**
LASSO	0.954 ± 0.007	**0.955 ± 0.005**
KNN	0.895 ± 0.011	**0.948 ± 0.003**
RF	0.955 ± 0.004	**0.956 ± 0.004**

classes based on all data, complete descriptors data, and energy barrier, respectively. Two-dimensional *t*-distribution stochastic neighbor embedding visualizations of class distribution are displayed in Figure S22. The sample numbers of all classes are approximately the same, validating the balance of the dataset.

5.13. Data Normalization and Insight Exploration

To assess the impact of outlier values on ML model performance, four types of datasets are used to train ML models: raw dataset; the dataset in which three samples in Table 5 are revised but five samples in Table 6 are not removed (semi-revised data1); the dataset in which five samples in Table 6 are removed but three samples in Table 5 are not revised (semi-revised data2); and completely revised data. MLR shows very poor performance on all the datasets, therefore its performance is omitted. R^2 of the other 6 models on test sets from four datasets are shown in Fig. 9. R^2 of six models based on revised data, semi-revised data1, and semi-revised data2 is always higher than that of models based on raw dataset. It reflects that the predictive ability of models can be improved by revising three samples in Table 5 and/or removing five samples in Table 6. And R^2 of the six models based on semi-revised data2 is always higher than that of models based on semi-revised data1. This indicates that five outlier-values in Table 6 have a greater impact on the predictive performance of ML models than three outlier-values in Table 5. One possible explanation is that the original values in Table 4 are already close to the corresponding revised values, while the removed outliers in Table 6 share a large difference with other values. This also makes R^2 of models based on semi-revised data2 and revised data comparable. In addition, the model can be “right for the wrong reasons” when the true signal is correlated with a false one in the data. Therefore, the R^2 values of some models based on semi-revised data2 are higher than those of the models based on revised data. As can be seen from Fig. 9, based on the revised data, the RR and GPR model achieved the highest R^2 of 0.959, and it reflects that the revised dataset is learnable.

5.14. Traceability assurance of data processing

A panorama of the finished ML project is shown in Fig. 10. Following the proposed lifecycle of data quality governance LC-QG, this project has implemented the governance of the nine DQDs of NASICON data as the lifecycle of materials data LC-MD goes on. Details and results of each process for quality governance were recorded as derived meta-data, and presented in the body of manuscript, supplemental information, and experimental codes.

6. Discussion and conclusion

Given the critical role of data quality in the entire process of materials property prediction (from data collection to model application), we believe data quality should be evaluated and treated effectively at all ML modelling stages. To this end, a rigorous execution sequence of DQD is established (i.e., LCM), to guarantee reliability, traceability, and correctness of information in materials data. Subsequently, the implementation of PM provides methodological support for LCM. Following this, we then evaluate 60 structured materials datasets in a case study, detecting issues through sequential analysis. Although all datasets are nominally AI-ready, 17 exhibit data quality issues, focusing primarily on issues of accuracy and redundancy. After governance, the revised datasets show an average 5.6 % improvement in insights. Notably, MD29 (a self-constructed dataset developed from scratch), achieves about 30 % insight improvement. Another interesting result is that data accuracy and redundancy play a key role in improving the ML model prediction accuracy, which is induced by the direct modification of data values and features. However, these modifications cannot always play a role for improving the prediction performance of ML models, e.g., MD15 and MD33, because the modified data distribution may be unsuitable for the fitting mechanisms of such ML models. Note that the insights of those datasets with the identical descriptor combination but different properties possess significant variations, as shown in Fig. 11, whose optimal ML models (Table S7) and corresponding performance exhibit fluctuations. This is because ML models establish relationships—either implicitly or explicitly—between descriptor combinations and target properties. Consequently, changes in the input of information (i.e., features) or the task objective (i.e., target property) can alter data characteristics, such as its distribution, and potentially affect the latent structure-activity relationships. Hence, governing data quality marks the outset of the ML pipeline, determining the ceiling for model performance. However, a model’s analytical capability ultimately hinges on both its learning algorithm and the underlying assumptions about data distribution. Domain knowledge can play a key role in guiding the entire ML learning process [33]. Importantly, in traditional ML workflows, domain knowledge is often primarily applied during data preprocessing and feature engineering, becoming deeply integrated within the learning process itself. As a result, at the moment it cannot be employed as an independent module or through separated representations and further work on formalizing integration of the domain knowledge into ML modeling is required.

Critically, the MAT-DQG framework is adaptable to any data format. Unlike structured data, unstructured data governance fundamentally shifts focus from descriptor selection and value correction (structured

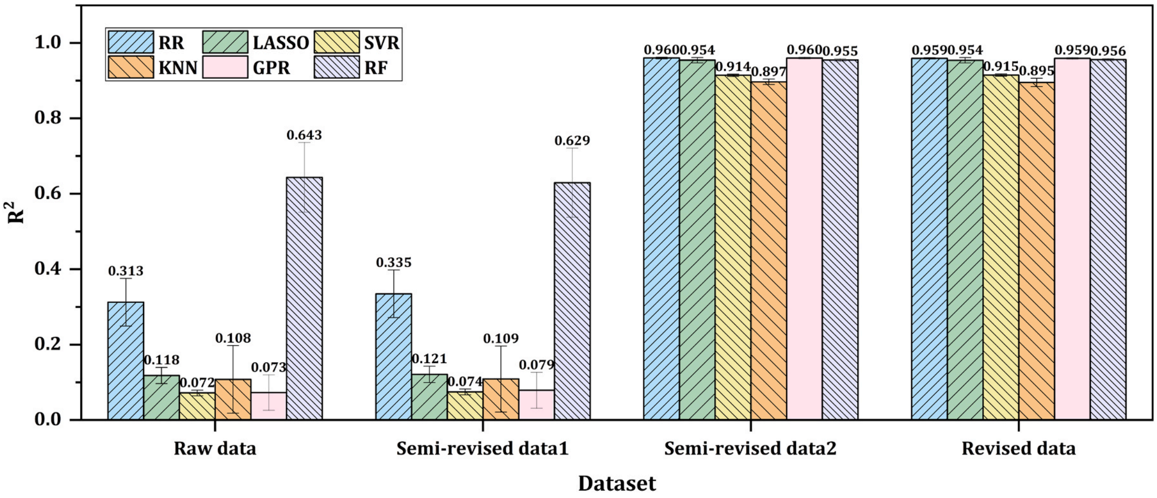


Fig. 9. R^2 of six models on test sets from the raw dataset, semi-revised data1, and semi-revised data2, and revised data (mean ± standard-deviation).

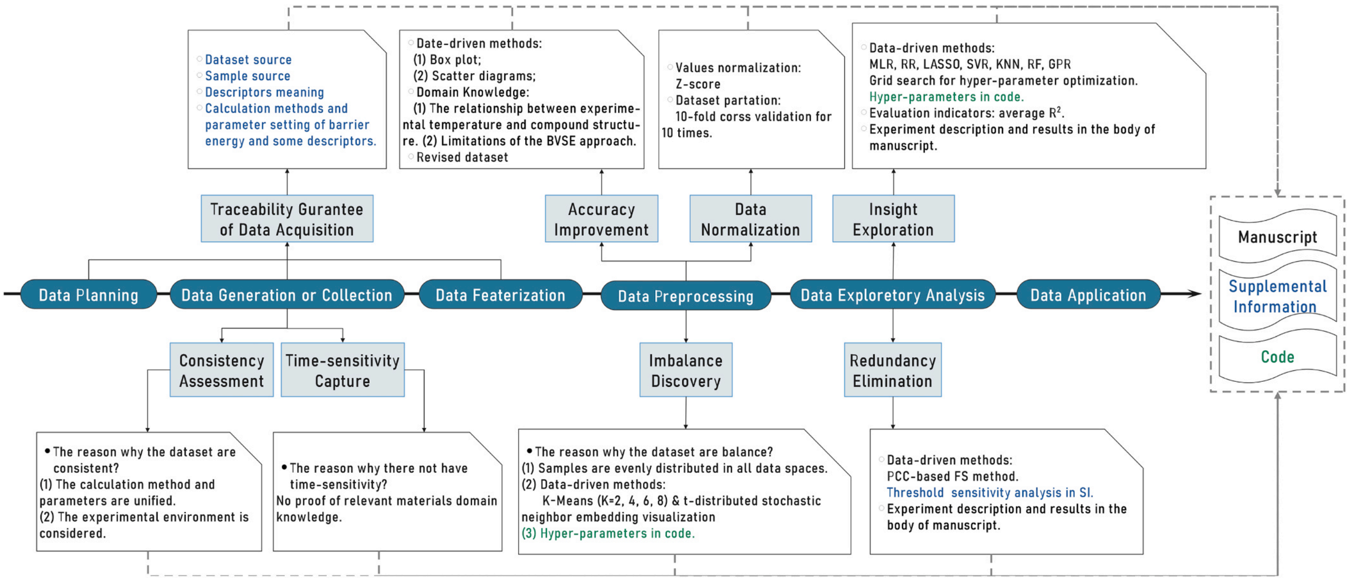

Dataset source
Sample source
Descriptors meaning
Calculation methods and parameter setting of barrier energy and some descriptors.
Date-driven methods:
(1) Box plot;
(2) Scatter diagrams;
Domain Knowledge:
(1) The relationship between experimental temperature and compound structure. (2) Limitations of the BVSE approach.
Revised dataset
Values normalization:
Z-score
Dataset partation:
10-fold corss validation for 10 times.
Data-driven methods:
MLR, RR, LASSO, SVR, KNN, RF, GPR
Grid search for hyper-parameter optimization.
Hyper-parameters in code.
Evaluation indicators: average R^2.
Experiment description and results in the body of manuscript.
Traceability Gurantee of Data Acquisition
Accuracy Improvement
Data Normalization
Insight Exploration
Manuscript
Supplemental Information
Code
Data Planning
Data Generation or Collection
Data Featerization
Data Preprocessing
Data Exploretory Analysis
Data Application
Consistency Assessment
Time-sensitivity Capture
Imbalance Discovery
Redundancy Elimination
The reason why the dataset are consistent?
(1) The calculation method and parameters are unified.
(2) The experimental environment is considered.
The reason why there not have time-sensitivity?
No proof of relevant materials domain knowledge.
The reason why the dataset are balance?
(1) Samples are evenly distributed in all data spaces.
(2) Data-driven methods:
K-Means (K=2, 4, 6, 8) & t-distributed stochastic neighbor embedding visualization
(3) Hyper-parameters in code.
Data-driven methods:
PCC-based FS method.
Threshold sensitivity analysis in SI.
Experiment description and results in the body of manuscript.

Fig. 10. The schematic of ML project. Cyan components represent the lifecycle of NASICON-type solid electrolyte materials data being used. Blue components stand for governance activities of all DQDs. Words in green and blue indicate that more detail is displayed in codes and SI.

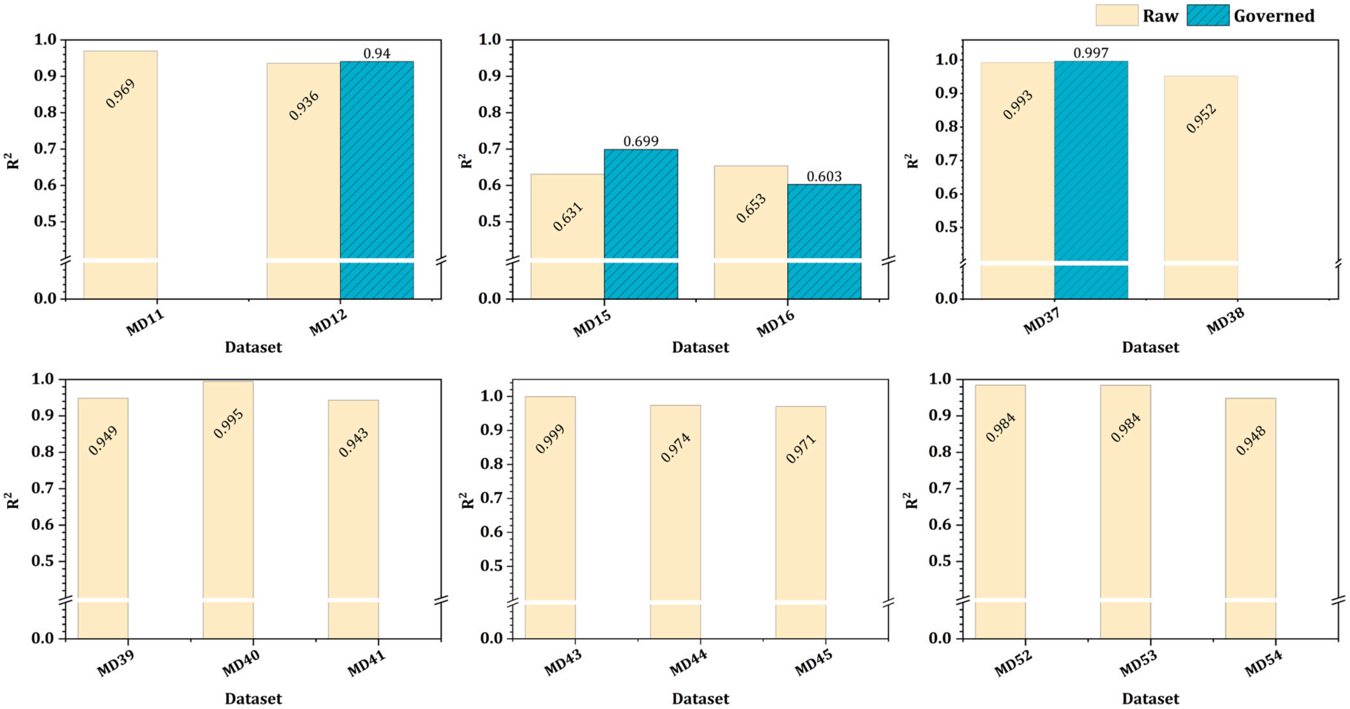


Fig. 11. Insight of datasets with the identical descriptor combination but different target properties.

data) toward representation patterns that control the quantity of valuable information [114]. To illustrate this, we use crystal representation as an example, highlighting that the choice of representation directly impacts both its accuracy and redundancy. Consequently, representation selection should be informed by the target ML models. For instance, while image-based representation captures only partial crystal information, it allows access to a wide range of image-oriented deep learning models, such as CNNs and their variants [115], as well as ViTs and their variants [116]. Conversely, graph-based representation—a mainstream approach—effectively accommodates non-Euclidean transformations. However, GNNs, the primary models for this representation, still suffer from over-smoothing and over-squashing due to inherent structural and learning limitations. Recently, efforts have emerged to integrate insights from GNNs and LLMs, aiming to leverage their complementary strengths to enhance both prediction accuracy and model interpretability [117, 118]. However, this strategy establishes only a superficial linkage between graph and text representations. Consequently, the integrated LLM embeddings fail to contextually adapt to structural patterns learned by GNNs. By contrast, string-based representations retain essential crystal information while excluding spatial relationships. Unlike graph-type data, this format is directly compatible with language models, circumventing information loss from intermediate encoding steps. Moreover, we developed a text-data governance pipeline under MAT-DQG guidance to construct high-quality datasets for materials science text mining

[9]. This approach reduces the substantial overhead in building large-scale supervised textual datasets while enhancing prediction accuracy in downstream text-mining models.

In summary, under the guidance of materials domain knowledge, we propose a general framework which comprehensively evaluates, monitors, and improves the materials data quality. Applying this framework resulted in significant improvement of the ML model performance in the case study with 60 materials structured datasets, suggesting its surprisingly versatile nature and great potential to improve the data quality for highly diverse materials. With the guidance of MAT-DQG, materials scientists can conduct reliable, reproducible, and interpretable data analysis in an orderly manner, and construct high-quality learning samples and high-accuracy ML models. Furthermore, the processing schemes proposed in this paper can be easily extended to large material systems. Thus, this work not only paves the way towards high-quality data foundation for ML modeling but also pushes forward ML-assisted research and development of novel materials. By combining the existing data analysis tools and extending the nine quality governance models to semi-structured and unstructured materials data, we hope to develop a series of automated tools to perform quality governance soon to effectively accelerate the research and development of novel materials.

CRediT authorship contribution statement

Yue Liu: Writing – original draft, Writing – review & editing, Conceptualization, Methodology, Supervision, Projection administration, Investigation. **Zhengwei Yang**: Writing – original draft, Writing – review & editing, Methodology, Software, Data curation, Visualization. **Xinxin Zou**: Writing – original draft, Methodology, Software, Visualization. **Yuxiao Lin**: Writing – review & editing. **Shuchang Ma**: Software, Visualization, Investigation. **Wei Zuo**: Writing – review & editing. **Zheyi Zou**: Validation, Investigation. **Hong Wang**: Supervision. **Maxim Avdeev**: Writing – review & editing. **Siqi Shi**: Writing – review & editing, Supervision, Projection administration.

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.

Acknowledgment

This work was supported in part by National Key Research and Development Program of China (No. 2021YFB3802101), National Natural Science Foundation of China (Nos. 92270124, 92472207 and 52102313), Hunan Provincial Natural Science Foundation of China (No. 2023JJ40635), and Shandong Province Natural Science Foundation (No. ZR2022ZD11). We appreciated the High Performance Computing Center of Shanghai University and Shanghai Engineering Research Center of Intelligent Computing System for providing the computing resources and technical support.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.mser.2025.101050.

Data availability

Data will be made available on request.

References

[1] W. Ye, C. Chen, Z. Wang, I.H. Chu, S.P. Ong, Deep neural networks for accurate predictions of crystal stability, Nat. Commun. 9 (2018) 3800.

[2] C.J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, G. Ceder, A critical examination of compound stability predictions from machine-learned formation energies, npj Comput. Mater. 6 (2020) 97.

[3] A.D. Sendek, Q. Yang, E.D. Cubuk, K.-A.N. Duerloo, Y. Cui, E.J. Reed, Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials, Energy Environ. Sci. 10 (2017) 306–320.

[4] N. Artrith, K.T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, Best practices in machine learning for chemistry, Nat. Chem. 13 (2021) 505–508.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, ACM Comput. Surv. 54 (2021) 1–35.

[6] A. Halevy, P. Norvig, F. Pereira, The Unreasonable Effectiveness of Data, IEEE Intell. Syst. 24 (2009) 8–12.

[7] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, L.M. Aroyo, Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems Article 39, Association for Computing Machinery, Yokohama, Japan, 2021.

[8] B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, I. Foster, A data ecosystem to support machine learning in materials science, MRS Commun. 9 (2019) 1125–1133.

[9] Y. Liu, D.-H. Liu, X.-Y. Ge, Z.-W. Yang, S.-C. Ma, Z.-Y. Zou, S.-Q. Shi, A high-quality dataset construction method for text mining in materials science, Acta Phys. Sin. 72 (2023) 070701.

[10] Y. Liu, Z. Yang, X. Zou, S. Ma, D. Liu, M. Avdeev, S. Shi, Data quantity governance for machine learning in materials science, nwad125, Natl. Sci. Rev. 10 (2023). nwad125.

[11] D. Firmani, M. Mecella, M. Scannapieco, C. Batini, On the meaningfulness of "big data quality", Data Sci. Eng. 1 (2016) 6–20.

[12] A.Y.T. Wang, R.J. Murdock, S.K. Kauwe, A.O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T.D. Sparks, Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, Chem. Mater. 32 (2020) 4954–4965.

[13] V. Kuznetsova, Á. Coogan, D. Botov, Y. Gromova, E.V. Ushakova, Y.K. Gun'ko, Expanding the Horizons of Machine Learning in Nanomaterials to Chiral Nanostructures, Adv. Mater. 36 (2024) 2308912.

[14] Y. Zhang, M. Safdar, J. Xie, J. Li, M. Sage, Y.F. Zhao, A systematic review on data of additive manufacturing for machine learning applications: the data quality, type, preprocessing, and management, J. Intell. Manuf. 34 (2023) 3305–3340.

[15] K.T. Butler, K. Choudhary, G. Csanyi, A.M. Ganose, S.V. Kalinin, D. Morgan, Setting standards for data driven materials science, npj Comput. Mater. 10 (2024) 231.

[16] Buriak, J.M., Akinwande, D., Artzi, N., Brinker, C.J., Burrows, C., Chan, W.C.W., Chen, C., Chen, X., Chhowalla, M., Chi, L., Chueh, W., Crudden, C.M., Di Carlo, D., Glotzer, S.C., Hersam, M.C., Ho, D., Hu, T.Y., Huang, J., Javey, A., Kamat, P. V., Kim, I.-D., Kotov, N.A., Lee, T.R., Lee, Y.H., Li, Y., Liz-Marzán, L.M., Mulvaney, P., Narang, P., Nordlander, P., Oklu, R., Parak, W.J., Rogach, A.L., Salanne, M., Samorì, P., Schaak, R.E., Schanze, K.S., Sekitani, T., Skrabalak, S., Sood, A.K., Voets, I.K., Wang, S., Wang, S., Wee, A.T.S. & Ye, J. Best Practices for Using AI When Writing Scientific Manuscripts. *ACS Nano* **17**, 4091-4093 (2023).

[17] M. Wenzlick, O. Mamun, R. Devanathan, K. Rose, J. Hawk, Assessment of Outliers in Alloy Datasets Using Unsupervised Techniques, JOM 74 (2022) 2846–2859.

[18] B. Li, W. Zhang, F.Z. Xuan, Machine-learning prediction of selective laser melting additively manufactured part density by feature-dimension-ascended Bayesian network model for process optimisation, Int. J. Adv. Manuf. Technol. 121 (2022) 4023–4038.

[19] W. Li, R. Jacobs, D. Morgan, Predicting the thermodynamic stability of perovskite oxides using machine learning models, Comput. Mater. Sci. 150 (2018) 454–463.

[20] H. Chen, J. Chen, J. Ding, Data Evaluation and Enhancement for Quality Improvement of Machine Learning, IEEE Trans. Reliab. 70 (2021) 831–847.

[21] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, E. Damer, Data quality of platforms and panels for online behavioral research, Behav. Res. Methods 54 (2022) 1643–1662.

[22] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for Data Quality Assessment and Improvement, ACM Comput. Surv. 41 (2009) 1–52.

[23] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch, Eff. Data Qual. Mach. Learn. Perform. (2022) arXiv:2207.14529 (.

[24] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, V. Munigala, Overview and importance of data quality for machine learning tasks. in Proc of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, ACM, 2020.

[25] C. Draxl, M. Scheffler, NOMAD: The FAIR concept for big data-driven materials science, MRS Bull. 43 (2018) 676–682.

[26] L. Sbailò, Á. Fekete, L.M. Ghiringhelli, M. Scheffler, The NOMAD Artificial-Intelligence Toolkit: turning materials-science data into knowledge and understanding, npj Comput. Mater. 8 (2022) 250.

[27] M. Scheidgen, L. Himanen, A.N. Ladines, D. Sikter, M. Nakhaee, Á. Fekete, T. Chang, A. Golparvar, J.A. Márquez, S. Brockhauser, NOMAD, A distributed web-based platform for managing materials science research data, J. Open Source Softw. 8 (2023) 5388.

[28] R. Amaro, J. Åqvist, I. Bahar, F. Battistini, A. Bellaiche, D. Beltran, P.C. Biggin, M. Bonomi, G.R. Bowman, R. Bryce, need Implement FAIR Princ. Biomol. Simul. arXiv Prepr. arXiv 240716584 (2024).

[29] K. Schmidt, A. Scourtas, L. Ward, S. Wangen, M. Schwarting, I. Darling, E. Truelove, A. Ambadkar, R. Bose, Z. Katok, Foundry-ML-Software and Services to Simplify Access to Machine Learning Datasets in Materials Science, J. Open Source Softw. 9 (2024) 5467.

[30] P.-O. Côté, A. Nikanjam, N. Ahmed, D. Humeniuk, F. Khomh, Data cleaning and machine learning: a systematic literature review, Autom. Softw. Eng. 31 (2024) 54.

[31] X. Xu, H. Liu, M. Yao, Recent Progress of Anomaly Detection, Complexity 2019 (2019) 2686378.

[32] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.

[33] Y. Liu, X. Zou, Z. Yang, S. Shi, Machine learning embedded with materials domain knowledge, J. Chin. Ceram. Soc. 50 (2022) 863–876.

[34] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C.W. Park, A. Choudhary, A. Agrawal, S.J.L. Billinge, E. Holm, S.P. Ong, C. Wolverton, Recent advances and applications of deep learning methods in materials science, npj Comput. Mater. 8 (2022) 59.

[35] Y. Liu, Z.W. Yang, Z.Y. Yu, Z.T. Liu, D.H. Liu, H.L. Lin, M.Q. Li, S.C. Ma, M. Avdeev, S.Q. Shi, Generative artificial intelligence and its applications in materials science: Current situation and future perspectives, J. Mater. 9 (2023) 798–816.

[36] H. Wickham, in: H. Wickham (Ed.), Data Analysis. in ggplot2: Elegant Graphics for Data Analysis, Springer International Publishing, Cham, 2016, pp. 189–201.

[37] B. He, S. Chi, A. Ye, P. Mi, L. Zhang, B. Pu, Z. Zou, Y. Ran, Q. Zhao, D. Wang, W. Zhang, J. Zhao, S. Adams, M. Avdeev, S. Shi, High-throughput screening platform for solid electrolytes combining hierarchical ion-transport prediction algorithms, Sci. Data 7 (2020) 151.

[38] R. Jacobs, L.E. Schultz, A. Scourtas, K.J. Schmidt, O. Price-Skelly, W. Engler, I. Foster, B. Blaiszik, P.M. Voyles, D. Morgan, Machine learning materials properties with accurate predictions, uncertainty estimates, domain guidance, and persistent online accessibility, Machine Learning Science Technology 5 (2024) 045051.

[39] C. Chen, Y.X. Zuo, W.K. Ye, X.G. Li, Z. Deng, S.P. Ong, A Critical Review of Machine Learning of Energy Materials, Adv. Energy Mater. 10 (2020) 1903242.

[40] H. Wu, A. Lorenson, B. Anderson, L. Witteman, H.T. Wu, B. Meredig, D. Morgan, Robust FCC solute diffusion predictions from ab-initio machine learning methods, Comput. Mater. Sci. 134 (2017) 160–165.

[41] Poltavsky, I., Charkin-Gorbulin, A., Puleva, M., Fonseca, G., Batatia, I., Browning, N.J., Chmiela, S., Cui, M., Frank, J.T., Heinen, S., Huang, B., Käser, S., Kabylda, A., Khan, D., Müller, C., Price, A.J.A., Riedmiller, K., Töpfer, K., Ko, T.W., Meuwly, M., Rupp, M., Csányi, G., von Lilienfeld, O.A., Margraf, J.T., Müller, K.-R. & Tkatchenko, A. Crash testing machine learning force fields for molecules, materials, and interfaces: model analysis in the TEA Challenge 2023. *Chemical Science* **16**, 3720-3737 (2025).

[42] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. 5 (2019) 83.

[43] Y. Liu, T.L. Zhao, W.W. Ju, S.Q. Shi, Materials discovery and design using machine learning, J. Mater. 3 (2017) 159–177.

[44] Y. Wang, Z. Pan, Y. Pan, A Training Data Set Cleaning Method by Classification Ability Ranking for the k -Nearest Neighbor Classifier, IEEE Trans. Neural Netw. Learn. Syst. 31 (2020) 1544–1556.

[45] M.H. Jeong, C.J. Sullivan, Y.Z. Gao, S.W. Wang, Robust abnormality detection methods for spatial search of radioactive materials, Trans. GIS 23 (2019) 860–877.

[46] B. Ma, X. Wei, C. Liu, X. Ban, H. Huang, H. Wang, W. Xue, S. Wu, M. Gao, Q. Shen, Data augmentation in microscopic images for material data mining, npj Comput. Mater. 6 (2020) 125.

[47] A.D. Sendek, B. Ransom, E.D. Cubuk, L.A. Pellouchoud, J. Nanda, E.J. Reed, Machine Learning Modeling for Accelerated Battery Materials Design in the Small Data Regime, Adv. Energy Mater. 12 (2022) 2200553.

[48] M. Álvarez-Moreno, C. de Graaf, N. López, F. Maseras, J.M. Poblet, C. Bo, Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform, J. Chem. Inf. Model. 55 (2015) 95–103.

[49] S.L. Weibel, T. Koch, The Dublin core metadata initiative, D. Lib. Mag. 6 (2000) 1082–9873.

[50] I. Altintas, S. Bhagwanani, D. Buttler, S. Chandra, C. Zhengang, M.A. Coleman, T. Critchlow, A. Gupta, H. Wei, L. Ling, B. Ludascher, C. Pu, R. Moore, A. Shoshani, M. Vouk, A modeling and execution environment for distributed scientific workflows, 15th Int. Conf. Sci. Stat. Database Manag. 2003 (2003) 247–250.

[51] G.S. Sureshrao, H.P. Ambulgekar, MapReduce-based warehouse systems: A survey, *ICAETR-2014*, 2014 Int. Conf. Adv. Eng. Technol. Res. (2014) 1–8.

[52] L. Carata, S. Akoush, N. Balakrishnan, T. Bytheway, R. Sohan, M. Seltzer, A. Hopper, A Primer on Provenance: Better understanding of data requires tracking its history and context, Queue 12 (2014) 10–23.

[53] Q. Sun, Y. Liu, W. Tian, Y.C.F.- Guo, PROV: A Content-Rich and Fine-Grained Scientific Workflow Provenance Model, IEEE Access 7 (2019) 30002–30016.

[54] Y. Liu, X.Y. Ge, Z.W. Yang, S.Y. Sun, D.H. Liu, M. Avdeev, S.Q. Shi, An automatic descriptors recognizer customized for materials science literature, J. Power Sources 545 (2022) 231946.

[55] L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G. J. Snyder, I. Foster, A. Jain, Matminer: An open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69.

[56] S. Lin, Y. Zhuang, K. Chen, J. Lu, K. Wang, L. Han, M. Li, X. Li, X. Zhu, M. Yang, G. Yin, J. Lin, X. Zhang, Osteoinductive biomaterials: Machine learning for prediction and interpretation, Acta Biomater. 187 (2024) 422–433.

[57] M.G. Kenward, J. Carpenter, Multiple imputation: current perspectives, Stat. Methods Med. Res. 16 (2007) 199–218.

[58] R.J. Hathaway, J.C. Bezdek, Fuzzy c-means clustering of incomplete data, IEEE Trans. Syst. Man Cybern. Part B (Cybern. 31 (2001) 735–744.

[59] W.L. ZHANG Jian, X.I.E. WANG Dong, L.U. Guang, Yuzhang, L.O.U. SHEN Jian, Langhong. Recent Progress in Research and Development of Nickel-Based Single Crystal Superalloys, Acta Met. Sin. 55 (2019) 1077–1094.

[60] Y. Li, C.F. Zou, M. Berecibar, E. Nanini-Maury, J.C.W. Chan, P. van den Bossche, J. Van Mierlo, N. Omar, Random forest regression for online capacity estimation of lithium-ion batteries, Appl. Energy 232 (2018) 197–210.

[61] R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse, M. Bokdam, Phase Transitions of Hybrid Perovskites Simulated by Machine-Learning Force Fields Trained on the Fly with Bayesian Inference, Phys. Rev. Lett. 122 (2019) 225701.

[62] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery, Mol. Syst. Des. Eng. 3 (2018) 819–825.

[63] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, B. Meredig, High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with Well-Calibrated Uncertainty Estimates, Integr. Mater. Manuf. Innov. 6 (2017) 207–217.

[64] C. Bandt, B. Pompe, Permutation Entropy: A Natural Complexity Measure for Time Series, Phys. Rev. Lett. 88 (2002) 174102.

[65] A.A.B. Pessa, H. Ribeiro, V. ordpy: A Python package for data analysis with permutation entropy and ordinal network methods, Chaos Interdisciplinary Journal Nonlinear Science 31 (2021).

[66] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (2019) 95–98.

[67] Z.W. Nie, Y.J. Liu, L.Y. Yang, S.N. Li, F. Pan, Construction and Application of Materials Knowledge Graph Based on Author Disambiguation: Revisiting the Evolution of $LiFePO_4$, Adv. Energy Mater. 11 (2021) 2003580.

[68] Y.Q. Liu, Y. Mu, K.Y. Chen, Y.M. Li, J.H. Guo, Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient, Neural Process. Lett. 51 (2020) 1771–1787.

[69] K. Ali Abd Al-Hameed, Spearman's correlation coefficient in statistical analysis, Int. J. Nonlinear Anal. Appl. 13 (2022) 3249–3255.

[70] S. Solorio-Fernández, J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad, A review of unsupervised feature selection methods, Artif. Intell. Rev. 53 (2020) 907–948.

[71] Z. Zhou, L. Zhang, Y. Yu, B. Wu, M. Li, L. Hong, P. Tan, Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning, Nat. Commun. 15 (2024) 5566.

[72] T. Li, S. Shetty, A. Kamath, A. Jaiswal, X. Jiang, Y. Ding, Y. Kim, CancerGPT for few shot drug pair synergy prediction using large pretrained language models, npj Digit. Med. 7 (2024) 40.

[73] A. Narayan, I. Chami, L. Orr, C. Ré, Can Foundation Models Wrangle Your Data? Proc. VLDB Endow. 16 (2022) 738–746.

[74] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.

[75] K. Roy, S. Kar, R.N. Das, QSAR/QSPR Modeling: Introduction. in *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*, Springer, Cham, 2015.

[76] Y. Liu, J.M. Wu, M. Avdeev, S.Q. Shi, Multi-Layer Feature Selection Incorporating Weighted Score-Based Expert Knowledge toward Modeling Materials with Targeted Properties, Adv. Theory Simul. 3 (2020) 1900215.

[77] Y. Liu, X. Zou, S. Ma, M. Avdeev, S. Shi, Feature selection method reducing correlations among features by embedding domain knowledge, Acta Mater. 238 (2022) 118195.

[78] M.A. Hernández, S.J. Stolfo, The merge/purge problem for large databases, ACM Sigmod Rec. 24 (1995) 127–138.

[79] H.H. Shahri, A.A. Barforush, A Flexible Fuzzy Expert System for Fuzzy Duplicate Elimination in Data Cleaning, in: F. Galindo, M. Takizawa, R. Traunmüller (Eds.), in Database and Expert Systems Applications, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 161–170.

[80] W.L. Low, M.L. Lee, T.W. Ling, A knowledge-based approach for duplicate elimination in data cleaning, Inf. Syst. 26 (2001) 585–606.

[81] J.D. Evans, F.-X. Coudert, Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning, Chem. Mater. 29 (2017) 7833–7839.

[82] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. 2 (2016) 16028.

[83] S. Garcia, J. Luengo, J.A. Sáez, V. Lopez, F. Herrera, A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning, IEEE Trans. Knowl. Data Eng. 25 (2012) 734–750.

[84] X. Rui, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (2005) 645–678.

[85] S.P. Niblett, P. Kourtis, I.-B. Magdău, C.P. Grey, G. Csányi, Transferability of datasets between Machine-Learning Interaction Potentials, arXiv 2409 (2024) 05590.

[86] Q. Yang, X.D. Wu, 10 Challenging problems in data mining research, Int. J. Inf. Technol. Decis. Mak. 5 (2006) 597–604.

[87] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based

approaches, IEEE Trans. Syst. Man Cybern. Part C. (Appl. Rev. 42 (2011) 463–484.

[88] P. Juszczak, D. Tax, R.P. Duin, Feature scaling in support vector data description. in Proc ASCI, Citeseer, 2002.

[89] K.T. Schütt, S.S.P. Hessmann, N.W.A. Gebauer, J. Lederer, M. Gastegger, SchNetPack 2.0: A neural network toolbox for atomistic machine learning, J. Chem. Phys. 158 (2023) 144801.

[90] K.T. Schütt, P. Kessel, M. Gastegger, K.A. Nicoli, A. Tkatchenko, K.R. Müller, SchNetPack: A Deep Learning Toolbox For Atomistic Systems, J. Chem. Theory Comput. 15 (2019) 448–455.

[91] Geiger, M. & Smidt, T. e3nn: Euclidean Neural Networks. arXiv:2207.09453 (2022).

92 I. Batatia, D. Péter Kovács, G.N.C. Simm, C. Ortner, G. Csányi, MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, arXiv 2206 (2022) 07697.

[93] M.A.F. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249.

[94] M. Guin, F. Tietz, Survey of the transport properties of sodium superionic conductor materials for use in sodium batteries, J. Power Sources 273 (2015) 1056–1064.

[95] H. Park, R. Mall, F.H. Alharbi, S. Sanvito, N. Tabet, H. Bensmail, F. El-Mellouhi, Exploring new approaches towards the formability of mixed-ion perovskites by DFT and machine learning, Phys. Chem. Chem. Phys. 21 (2019) 1078–1088.

[96] A. Klein, S. Falkner, S. Bartels, P. Hennig, F. Hutter, Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. in Proc of the 20th International Conference on Artificial Intelligence and Statistics, PMLR, 2017.

[97] Y. Liu, S. Wang, Z. Yang, M. Avdeev, S. Shi, Auto-MatRegressor: liberating machine learning alchemists, Sci. Bull. 68 (2023) 1259–1270.

[98] B. He, P.H. Mi, A.J. Ye, S.T. Chi, Y. Jiao, L.W. Zhang, B.W. Pu, Z.Y. Zou, W. Q. Zhang, M. Avdeev, S. Adams, J.T. Zhao, S.Q. Shi, A highly efficient and informative method to identify ion transport networks in fast ion conductors, Acta Mater. 203 (2021) 116490.

[99] M.R. Johan, S. Ibrahim, Optimization of neural network for ionic conductivity of nanocomposite solid polymer electrolyte system (PEO-LiPF$_6$-EC-CNT), Commun. Nonlinear Sci. Numer. Simul. 17 (2012) 329–340.

[100] Q. Zhao, M. Avdeev, L. Chen, S. Shi, Machine learning prediction of activation energy in cubic Li-argyrodites with hierarchically encoding crystal structure-based (HECS) descriptors, Sci. Bull. 66 (2021) 1401–1408.

[101] X. Jiang, H.Q. Yin, C. Zhang, R.J. Zhang, K.Q. Zhang, Z.H. Deng, G.Q. Liu, X. H. Qu, An materials informatics approach to Ni-based single crystal superalloys lattice misfit prediction, Comput. Mater. Sci. 143 (2018) 295–300.

[102] C. Wen, C.X. Wang, Y. Zhang, S. Antonov, D.Z. Xue, T. Lookman, Y.J. Su, Modeling solid solution strengthening in high entropy alloys using machine learning, Acta Mater. 212 (2021) 116917.

[103] A. Hemmati-Sarapardeh, M. Tashakkori, M. Hosseinzadeh, A. Mozafari, S. Hajirezaie, On the evaluation of density of ionic liquid binary mixtures: Modeling and data assessment, J. Mol. Liq. 222 (2016) 745–751.

[104] L.E. Eberly, Multiple linear regression, Methods Mol. Biol. 404 (2007) 165–187.

[105] G.C. McDonald, Ridge regression, WIREs Comput. Stat. 1 (2009) 93–100.

[106] A. Ishikawa, K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, M. Okada, Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents, Phys. Chem. Chem. Phys. 21 (2019) 26399–26405.

[107] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.

[108] V.L. Deringer, A.P. Bartók, N. Bernstein, D.M. Wilkins, M. Ceriotti, G. Csányi, Gaussian Process Regression for Materials and Molecules, Chem. Rev. 121 (2021) 10073–10141.

[109] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.

[110] O. Kramer, Scikit-Learn. ed. Kramer, O., in: Machine Learning for Evolution Strategies, Springer International Publishing, Cham, 2016, pp. 45–53.

[111] Z. Wang, M. Zoghi, F. Hutter, D. Matheson, N.D. Freitas, Bayesian optimization in high dimensions via random embeddings. in Proceedings of the Twenty-Third international joint conference on Artificial Intelligence 1778–1784, AAAI Press, Beijing, China, 2013.

[112] D.T. Qui, J. Capponi, J. Joubert, R. Shannon, Crystal structure and ionic conductivity in Na$_4$Zr$_2$Si$_3$O$_{12}$, J. Solid State Chem. 39 (1981) 219–229.

[113] Z.Y. Zou, N. Ma, A.P. Wang, Y.B. Ran, T. Song, Y. Jiao, J.P. Liu, H. Zhou, W. Shi, B. He, D. Wang, Y.J. Li, M. Avdeev, S.Q. Shi, Relationships Between Na$^+$ Distribution, Concerted Migration, and Diffusion Properties in Rhombohedral NASICON, Adv. Energy Mater. 10 (2020) 2001486.

[114] K. Atz, F. Grisoni, G. Schneider, Geometric deep learning on molecular representations, Nat. Mach. Intell. 3 (2021) 1023–1032.

[115] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, M. Parmar, A review of convolutional neural networks in computer vision, Artif. Intell. Rev. 57 (2024) 99.

[116] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A Survey on Vision Transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 87–110.

[117] Y. Li, V. Gupta, M.N.T. Kilic, K. Choudhary, D. Wines, W.-k Liao, A. Choudhary, A. Agrawal, Hybrid-LLM-GNN: integrating large language models and graph neural networks for enhanced materials property prediction, Digit. Discov. 4 (2025) 376–383.

[118] J. Xu, Z. Wu, M. Lin, X. Zhang, S. Wang, LLM and GNN are Complementary: Distilling LLM for Multimodal Graph Learning, arXiv 2406 (2024) 01032.

Yue Liu obtained her B.S. and M.S. in computer science from Jiangxi Normal University in 1997 and 2000, respectively. She got her Ph.D. in control theory and control engineering from Shanghai University (SHU) in 2005. She was 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 September 2012 to September 2013. At present, she is a professor at SHU. Her current research interest focuses on machine learning, data mining, data quality governance, and AI for materials science.

Zhengwei Yang received the M.S. degree from Shandong University of Technology in 2021. He is currently pursuing Ph. D. degree in Shanghai University. His research interest focuses on AI for materials science and data quality governance.

Xinxin Zou received the M.S. degree from Shanghai University in 2021. Her research interest focuses on AI for materials science, with dedicated efforts in advancing AI applications within this field.

Yuxiao Lin is currently a Distinguished Professor of Jiangsu Province, focusing on understanding the mechanism of electrochemical energy storage devices and materials based on computational materials science. He got his bachelor and master degree from Tsinghua University, and PhD from Michigan State University. After finishing the postdoc program in Idaho National Lab, he joined Jiangsu Normal University. He has published over 40 papers, including Nat. Commun., Angew. Chem., and Adv. Mater. He also received multiple awards, including Exceptional Contributions Awards in Idaho National Lab, Outstanding Research Presentation in EFRC-NEES Accomplishment Meeting, and Youth May Fourth Medal in Jiangsu Normal University.

Shuchang Ma received the the M.S. degree from Shanghai University in 2024. Her research interest focuses on data quality governance.

Wei Zuo is currently pursuing Ph.D. degree in Shanghai University. His main research interest focuses on AI for materials science and data quality governance.

Zheyi Zou received his Ph.D. in 2020 from Shanghai University. He is currently a lecturer at the School of Materials Science and Engineering, Xiangtan University. His research mainly focuses on the understanding of ion transport in fast ion conductors by first-principles calculations and molecular dynamics simulations.

Hong Wang received his Ph.D. in Mateials Science and Engineering from the University of Illinois at Urbaba-Champaign, USA in 1994. He is a "Zhiyuan" Chair Professor of Materials Science and Engineering and the Director of Materials Genome Initiative Center, Shanghai Jiao Tong University. He also serves as the Director of the Materials Genome Engineering Field Committee, Chinese Society of Materials Testing (CSTM). He conducted R & D on semiconductor, plat panel displays, coated glass and smart windows for energy efficient buildings, as well as solar heat coatings with companies such as SONY, Panasonic, Guardian Industries Corp. in USA and China Building Materials Academy, Beijing. His current research focuses on the theory of materials genome engineering, high-throughput material preparation and characterization techniques, and materials informatics.

Maxim Avdeev received Ph.D. from the Rostov State University in 1999 for the work in synthesis and crystal structural studies of Na-superionic conductors. After postdoctoral position at the Argonne National Laboratory (USA), in 2005 he joined the Australian Nuclear Science and Technology Organisation (ANSTO) as a full-time researcher, where he now leads Neutron Diffraction Group at the Australian Centre for Neutron Scattering. He also holds a professor position at the University of Sydney. His main research interests are studies of crystal and magnetic structure of inorganic materials using X-ray and neutron diffraction and atomistic simulations.

Siqi Shi is Ministry of Education Yangtze River Scholar Professor (Materials Genome Engineering), Winner of National Excellent Youth Science Foundation and Shanghai Leading Talent Program, Ph.D. Supervisor. He 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. 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. In 2001, he was among the first in China to carry out first-principles calculations on electrochemical energy storage materials. He authored the monograph Computation, Modeling, and Simulation in Electrochemical Energy Storage, and developed the open-access platform for electrochemical energy storage materials design (www.bmaterials.cn). His current research interests focus on establishing a new paradigm for energy materials design that promotes mutual reinforcement among algorithms, data and knowledge, while integrating closely with experimental validation.