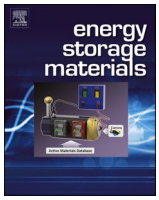
Energy Storage Materials 80 (2025) 104390

Contents lists available at ScienceDirect

Energy Storage Materials

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



A knowledge acquisition automatizing framework from literature exemplified by Na^+ activation energy prediction of NASICON solid-state electrolyte

Yue Liu [a,d], Dahui Liu [a], Zhengwei Yang [a], Xianyuan Ge [a], Wenxuan Yao [a], Jie Wu [a], Maxim Avdeev [e,f], 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] Shanghai Engineering Research Center of Intelligent Computing System, Shanghai 200444, China
[e] Australian Nuclear Science and Technology Organisation, Sydney 2232, Australia
[f] School of Chemistry, The University of Sydney, Sydney 2006, Australia

ARTICLE INFO

Keywords:
Information extraction
Knowledge graph
Machine learning
Materials science

ABSTRACT

Materials science literature contains vast amount of structure-activity relationship knowledge crucial for materials discovery and design. However, automatic extraction of domain knowledge from literature remains challenging due to its unstructured and heterogeneous format. Herein, we propose a framework for automating knowledge acquisition, which involves a materials entity-aware relational extraction model (MatRE) to mine triples, an approach to construct a knowledge graph (KG) for the detection of associations among triples, as well as inference and representation of structure-activity relationships in a machine learning (ML)-compatible format. We demonstrate its application in predicting sodium ion activation energy for the NASICON solid-state electrolyte (SSE) system. MatRE trained on a NASICON SSE dataset, achieves an F1-score of 0.80, and is used to extract 260,475 entity–relation triples from 1,808 scientific publications. Furthermore, embedding 24 knowledge bullets from the KG into data pre-processing and feature engineering stages improves the performance and interpretability of six common ML models by up to 25.7%. This work offers key insights into automatic knowledge acquisition from literature and heralds a new paradigm for AI-assisted materials genome engineering driven by both data and knowledge.

1. Introduction

Recent developments in knowledge-driven computational intelligence are adding momentum to the "AI for Materials Science" research, thereby accelerating optimization of materials performance and design of novel materials [1–4]. Currently, materials domain knowledge primarily stems from two sources: (1) cognitive experience and (2) literature (including scientific journals, textbooks and patents). Generally, the former is typically confined to experts within narrow research fields, which poses a significant barrier to interdisciplinary research. In contrast, the latter comprises vast amount of semi- or un-structured domain knowledge screened by peer review process. Nevertheless, manually searching and extracting domain knowledge from massive body of literature is a time- and resource-intensive endeavor. Even worse, the subjectivity of different experts and the knowledge limitations inherent in specialized fields can significantly impact the extraction process. This challenge is intensifying as publication volumes continue to soar.

Recently, the rapid development of Natural Language Processing (NLP) has provided a long-sought path to automatize the aforementioned search and extraction process [5–7]. With the aid of Named Entity Recognition (NER), researchers have already made impressive progress in knowledge extraction for inorganic materials mentions [8–10], synthesis recipes [11–14], sundry materials information [15, 16], *etc*. Nevertheless, NER is just the first step of the knowledge extraction, and the following Relation Extraction (RE) is the key to

* Corresponding author.
E-mail address: sqshi@shu.edu.cn (S. Shi).

https://doi.org/10.1016/j.ensm.2025.104390
Received 6 January 2025; Received in revised form 25 May 2025; Accepted 8 June 2025
Available online 9 June 2025

explore the potential relationships between materials entities. RE focuses on identifying and classifying the relationships between entity pairs in literature, which include the coappearance among entities and the correlation between entities and corresponding property. In most cases, RE in materials science is solely realized by dependency parsing based methods [17–20] involving pre-defined rules and snowball algorithm. However, the large-scale manual intervention during the operation and the high-ratio negative samples in the extracted results can dramatically hinder practical application of structure-activity relationships. In light of these challenges, Tshitoyan *et al.* [21] proposed an unsupervised manner to embed words with rich information and then deduced the similar relations among entities through vector operations, which works in the recognition of potential thermoelectric materials. Nie *et al.* [22] further confirmed that the word embeddings can be trained by combined domain corpora, which enabled the exploration of the possible relations between materials entities through cosine similarities. Dagdelen *et al.* [23] proposed a method using large language models for joint named entity recognition and relation extraction using large language models, fine-tuned to automatically extract structured entity-relational data from materials science texts. These appealing methods with powerful features are of high referential significance to the RE in materials science.

However, a majority of texts are dependent on materials systems with various unique properties, leading to numerous overlapping relationships. For example, high ionic conductivity is a key factor in the design of NASICON solid-state electrolytes (SSEs), and their application is determined by activation energy, interfacial resistance, and electrochemical stability window. A specific example is provided in the sentence "*Increasing the sintering temperature causes the lattice parameters and the unit volume to increase*", where "*temperature*" acting as the subject entity, has a "*cause*" relation simultaneously with "*lattice parameters*" and "*unit volume*". This overlapping issue arises from the complex grammatical structures in materials science literature, making it challenging to apply methods developed in other fields. One potential solution is to fully recognize semantic information of materials entities to facilitate the relation extraction process from materials corpora. Therefore, there is a critical need for an efficient and convenient approach to mine and represent domain knowledge embedded in unstructured data.

Building on prior work in materials NER [24], we develop a framework for recognizing structure-activity relationships tailored to materials science literature, with the goal of providing an effective method for automating the acquisition of materials domain knowledge. In this framework, **Mat**erials Entity-aware **R**elational **E**xtraction Model (MatRE) is first introduced, which is a neural network composed of three distinctive modules enabling an easy way to accurately extract the entity-relation triples from materials science literature. Based on this, we further propose the algorithm to construct knowledge graph (KG) and acquire corresponding knowledge, which facilitates the detection of deep-seated associations among materials entities, as well as inference and symbolical representation of potential structure-activity relationships in a form amenable to ML. RE datasets from various materials systems are used to evaluate the performance of MatRE and a substantial improvement in the classification accuracy of relational labels can be observed. Overall, the contribution of this work can be summarized as the following three points:

(1) Based on entity-aware word embeddings computed by fine-tuned language model, an RE model named MatRE is developed to automatically extract relationships between materials entities and represent them as triples, thereby overcoming manual retrieval constraints.
(2) A materials KG construction and acquisition method is designed. It involves storing triples in a general graph database, organizing them into KGs, and deducing high-quality and fine-grained structure-activity relationships relevant to target property. This establishes robust knowledge support for materials property prediction bidirectionally driven by data and knowledge.
(3) Taking ionic conductivity activation energy prediction as an example, MatRE trained on a NASICON SSE dataset achieves an F1-score of 0.80 and is used to extract 260,475 triples from 1,808 NASICON SSEs literature sources. With the embedding of 24 relevant knowledge bullets screened from KG into the data preprocessing and feature engineering stages, the prediction accuracy of 6 ML models exhibits an improvement of 25.7%.

The remainder of this paper is organized as follows. The methodology of structure-activity relationship recognizing framework is introduced in section 2. The testing results and corresponding analysis are presented in Section 3. The specific application by an example of knowledge acquisition and performance prediction of NASICON SSEs is shown in Section 4. Finally, the conclusion of this article and an outlook for future work are presented in Section 5.

2. Methods

Our method achieves the extraction of the relation between materials entity and the corresponding application in a pipeline manner and the detailed workflow is illustrated in Fig. 1. Initially, materials entities are obtained using our previously developed automatic recognizer for descriptors based on NER model [24]. Notably, this pipeline is not restricted to our NER model, but can also be applied to other models able to identify materials entities. Then, the target entities are surrounded by predefined entity-aware special tokens to enhance the model's comprehension of materials terms. On this basis, we study the RE method in materials science and innovatively propose MatRE model, which enables **Mat**erials-specific **P**re-training **L**anguage **M**odel (MatPLM) and other components to recognize the relations. We further propose the approaches to construct the materials KG and obtain structure-activity knowledge, including the materials KG construction method based on Neo4j database [25] and the knowledge acquisition algorithm driven by target property. These approaches make it possible to quickly obtain knowledge related to the materials structure-activity relationships and to realize the performance prediction embedded with domain knowledge. Details of our method are introduced in **Section 2.1** and **2.2** below.

2.1. Materials entity-aware relational extraction model

As illustrated in Fig. 2, MatRE consists of three independent modules: the entity-aware word embedding module; the BiGRU-based semantic representation module; and the materials relational classification module. This framework can effectively realize the prediction and classification of the relations among target entities.

2.1.1. Entity-aware word embedding module

Since word embedding can quantitatively represent different materials terminologies in the same local contexts, it is considered as the prerequisite for the extraction of materials relationship. Due to the multi-dependence of targeting properties and the overlapped relations resulted from complicated grammar structure in materials science literature, we add special tokens around entities related to key terminologies to emphasize their importance. Specifically, given a sentence of $S = \{w_1, w_2, \cdots, w_n\}$ and a set of triples from sentence $T = \{(s, r, o)\} \in S$, the subject entity s and the object entity o consist of word $w_i \sim w_j$ and $w_k \sim w_m$, respectively. Here, r represents the classified relation between s and o. First, we insert the special tokens of "[" and "]" at the starting and ending positions of s. Similarly, we use another special token set of "{"and "}"to indicate the starting and ending positions of o. These tokens can increase the sensitivity of MatPLM on the target entity, with more attention paid to the classification and boundary information. In addition, the tokens of "[CLS]"and "[SEP]" are also added at the start and

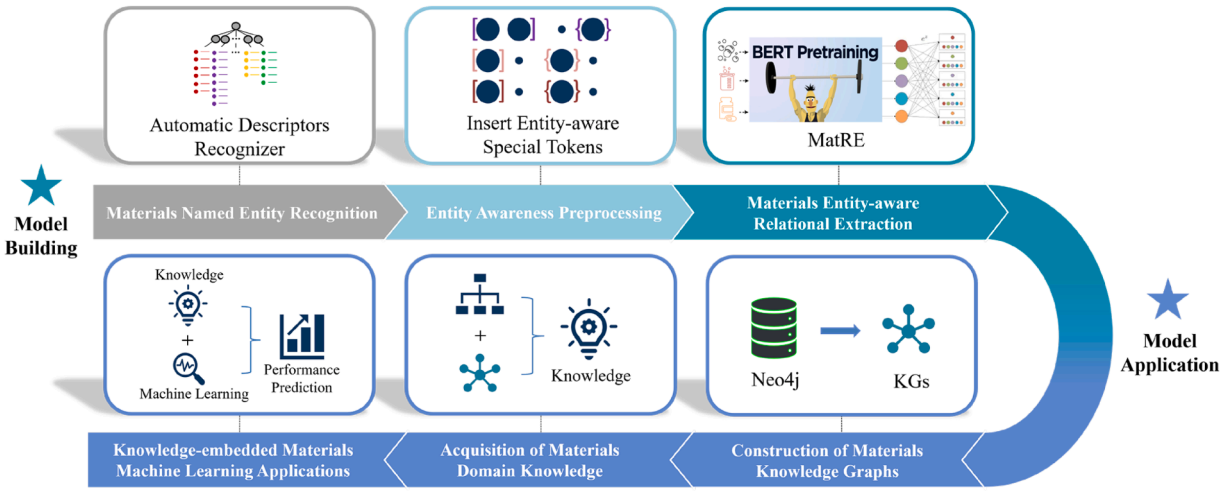


Fig. 1. Overview of the automatic structure-activity relationship recognizing framework.

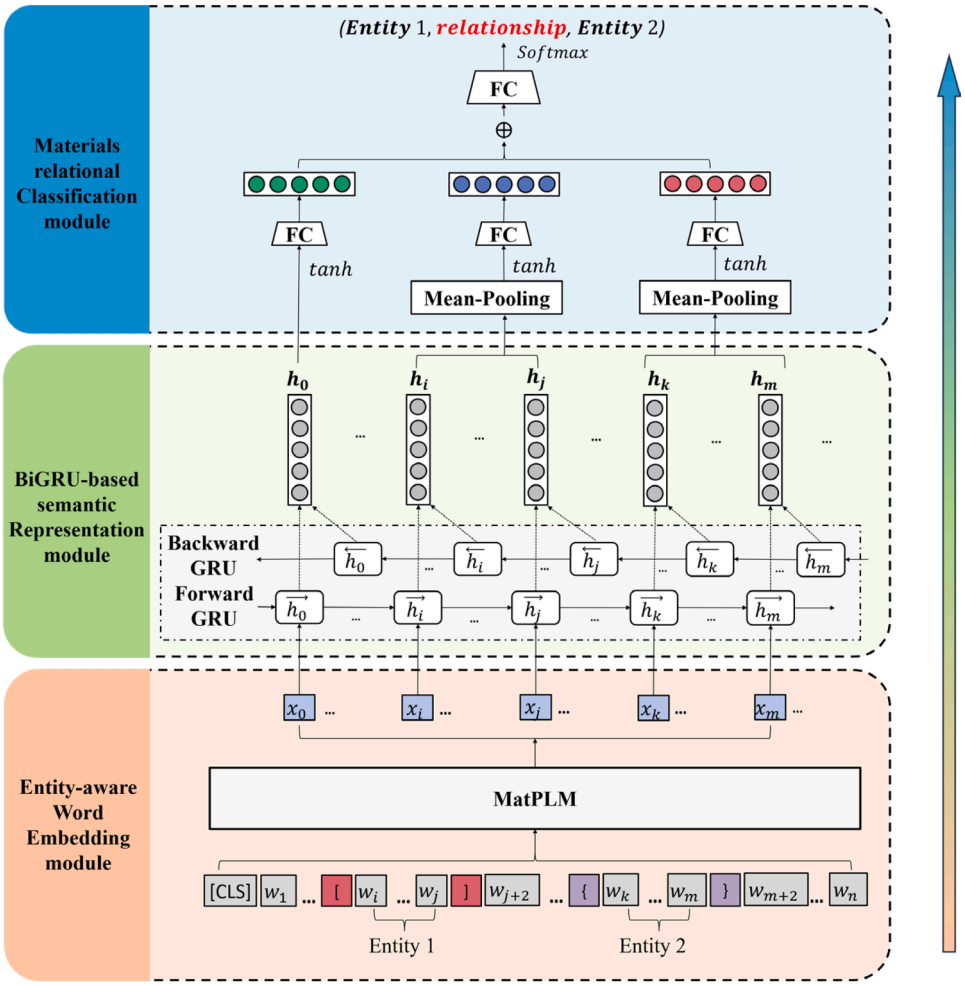


Fig. 2. The architecture of materials entity-aware relational extraction model.

end of sentence, making it feasible to globally represent the semantic information of whole sentence.

By analyzing the materials corpora, we can find that even the same word may possess totally different meaning in different contexts, not to mention the synonyms. Generally, this issue can not be addressed by language model in other fields. Therefore, we introduce MatPLM to generate the word embedding vector for each target word. For given sentence S, we first tokenize it into a word list with inserted entity-aware tokens. After that, the expanded word list $S = \{w_{CLS}, w_1, w_2, \cdots, w_n, w_{SEP}\}$ is transferred to MatPLM as an input attribute. The generated word embedding can be represented as:

$$x_i = MatPLM(w_i), i = \{CLS, 1, \cdots, n, SEP\} \quad (1)$$

Here, x_i is the vector representation of the i-th word. $MatPLM(*)$ is the PLM after fine-tuning, which is a component that can be replaced. Compared to BERT [26], the state-of-the-art MatSciBERT [27] has much more abundant domain vocabulary, which can effectively limit the maximum length of tokenized list and solve the "out of vocabulary" issue. Thus, it is chosen as the candidate PLM.

2.1.2. BiGRU-based semantic representation module

Although MatPLM can effectively capture the word-level semantic features in the sentence, this information may be gradually lost during the transfer process between the internal layers in the model, especially the positional information of the words closely related to the RE task. To avoid the loss of such information and improve the ability to focus on the semantic relation of the target word in the contexts, we introduce the Bidirectional Gated Recurrent Unit (BiGRU) [28]. It is built based on the modeling of sentence sequence and can represent the local contextual semantics of each word. Compared to Recurrent Neural Networks [29], the settings of update gate and reset gate in GRU can be used to screen information needing to be saved to prevent its loss in the previous step. This can mitigate the issues of gradient disappearance or explosion. In addition, this bidirectional setting also makes the performance of GRU match up with that of Long Short-Term Memory networks [30] at a lower computational cost.

BiGRU is made up of the forward and backward GRU. The forward GRU first sequentially encodes the word embedding input $(x_{CLS}, x_1, \cdots, x_n, x_{SEP})$ into the forward hidden states $\left(\overrightarrow{h_{CLS}}, \overrightarrow{h_1}, \cdots, \overrightarrow{h_n}, \overrightarrow{h_{SEP}}\right)$. Then, the backward GRU conversely translates the word embedding input into backward hidden states $\left(\overleftarrow{h_{CLS}}, \overleftarrow{h_1}, \cdots, \overleftarrow{h_n}, \overleftarrow{h_{SEP}}\right)$. At t step, $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are updated based on the following equations.

$$\overrightarrow{h_t} = GRU\left(x_t, \overrightarrow{h_{t-1}}\right) \tag{2}$$

$$\overleftarrow{h_t} = GRU\left(x_t, \overleftarrow{h_{t-1}}\right) \tag{3}$$

Finally, we can obtain the local contextual semantics of each word by connecting $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, as shown in Eq. (4).

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{4}$$

2.1.3. Materials relational classification module

The goal of RE is to find an appropriate relation $r \in R$ for each triple $T = (s, r, o)$ in the sentence sequence. Here, R is the predefined set of relation types. Based on previous modules, we can obtain the word vectors $H = (h_{CLS}, h_1, \cdots, h_n, h_{SEP})$ with rich semantic information in the contexts. To accurately classify the relations between subject entity $s = (w_i, \cdots, w_j)$ and object entity $o = (w_k, \cdots, w_m)$, we need to represent the entity span of subject and object $\left(h_{ij}^s \text{ and } h_{km}^o\right)$:

$$h_{ij}^s = W_s\left[tanh\left(meanpool\left(h_{[}, h_i, \cdots, h_j, h]\right)\right)\right] + b_s \tag{5}$$

$$h_{km}^o = W_o[tanh(meanpool(h_{\{}, h_k, \cdots, h_m, h_{\}}))] + b_o \tag{6}$$

Here, $W_{(*)}$ and $b_{(*)}$ indicate weights and biases that can be studied; $(h_{[}, h])$ and $(h_{\{}, h_{\}})$are vector representations corresponding to the subject and object entity-aware tokens of the subject and object entities; $meanpool(*)$ represents the mean-pooling operation, which uses the mean value of all word vectors to approximate the entity representation. Furthermore, the sentence-level semantic information can also be obtained from the global feature h_{global}, which is calculated based on the word vector h_{CLS} inserted at the beginning of the sentence:

$$h_{global} = W_g[tanh(h_{CLS})] + b_g \tag{7}$$

At last, we concatenate the word-level features (h_{ij}^s and h_{km}^o) and sentence-level global features (h_{global}), and then realize the relation classification through fully connected layers and Softmax classifiers, as shown in Eq. (8).

$$p(r|r \in R) = Softmax\left(W_f\left[h_{ij}^s, h_{km}^o, h_{global}\right] + b_f\right) \tag{8}$$

Here, $p(r|r \in R)$ is the probability that the relation between subject s and object o is r.

2.2. Structure-activity relationship knowledge acquisition method

2.2.1. Materials KG construction

To effectively explore the abundant potential knowledge in the extracted information, we realize the storage of materials triples based on the Neo4j database. Neo4j is an open-access NoSQL graph database based on Java, aiming to optimize the fast management, storage and iteration process of entities and relations [25]. There are only two types of data in Neo4j: nodes and directional edges. Nodes store entity information, while directed edges connect these nodes to depict relationships among entities. By performing reading and writing operations on Neo4j through Python py2neo toolkit, we can realize the storage of triples.

Based on that, we can construct the materials KG, which enables quick searching and deducing of materials knowledge driven by the target property. KG is logically constructed by the data layer and the pattern layer. The former can save a series of factual data, among which knowledge is represented in the unit of fact (*e.g.*, the triples of subject-relation-object or entity-property-value). The latter regulates the facts through ontology and thus are built on top of the data layer. According to the unique features in materials field, we define materials-domain KG particularity (MatKGPtcl) shown in Eq. (9) to support the construction of materials KG.

$$MatKGPtcl \leq F_{kno}, O_{kno}, U_{kno} > \tag{9}$$

Here, F_{kno} is the factual representation of knowledge, which can be specified as triples in this article. O_{kno} is the organization form of the knowledge used to set the structure for materials KG. U_{kno} is the usage demand of the knowledge. In this study, we aim to find out more explicit and implicit knowledge by fully utilizing the deductive ability of materials KG. This can enables visualized knowledge representation for materials research and further provides formalized knowledge support for ML in materials science.

Based on Neo4j and MatKGPtcl, we can construct property driven materials KG. The detailed process is listed below:

(1) Set the organizational structure of KG from top (pattern layer) to bottom (data layer);
(2) Take the target materials property as root node and fill the pattern layer with the abstract information (*e.g.*, Composition, Structure, Processing, Property, Feature, Application, Condition and Characterization);
(3) Fill the data layer with triples of materials entities and relations.

It is worth noting that the information stored in the data layer is not only constrained by the pattern layer, but also needs to be related to the target material property.

2.2.2. Structure-activity relationship knowledge acquisition

To reveal the hierarchical relations between target entities in the materials KG and formalize them into knowledge bullets about the specific materials property and the corresponding influencing factors, we outline the approach of Materials structure-activity relationship Knowledge Acquisition (MatKA) with detailed process listed below:

Step 1: Define the sets of target entity. Iterate through all elements that may be related to the target entity and combine each entity in the materials KG into pairs.

Step 2: Acquire the relations between entities and originate the

corresponding sentences. Search target entity pairs in the materials KG and acquire the relations between entities and the corresponding original sentences. Save the information to the candidate knowledge database.

Step 3: Deduce the knowledge about structure-activity relationship according to defined rules. Iterate through candidate knowledge database, and perform deduction for all saved sentences. If a sentence contains correlation words (*e.g.* positive, negative, increase, decrease) or influencing rules (*e.g.* the increase of A can lead to the decrease of B), it will be included in the final knowledge database since there exists a structure-activity relation among the materials entities in this sentence.

The pseudocode of MatKA is shown in Algorithm 1. $R = \{r_1, r_2, \cdots, r_k\}$ is the predefined set of relation types, which is instantiated to the 7 semantic relations in the application process; $D = \{d_1, d_2, \cdots, d_n\}$ is the target entity set; $r(a, b)$ describes the relation between a and b; S is the sentence containing d_i, d_j and their relation $r(d_i, d_j)$; $Corr_{know}$ is the correlation words.

3. Experiments

3.1. Experimental setups

The following two materials datasets are used to evaluate the performance of MatRE.

- NASICON SSEs RE dataset: This dataset is constructed under the guidance of domain experts. It was derived from 55 randomly selected publications related to "NASICON solid electrolytes", collected via web crawling. We defined 7 semantic and 1 general relationship tags. The former includes "Cause-Effect", "Component-Whole", "Instance-Of", "Located-Of", "Method-Of", "Condition-On", and "Feature-Of". The latter is "Other", namely, other types apart from the above 7 tags. Since these tags are chosen according to the well-known tetrahedral principle of materials science (i.e., structure, properties, processing and performance), they can also be applied to other materials systems. See Supplementary Information S.1 for the detailed construction process (including data collection, text preprocessing, label definition and tagging, etc.) and the exemplified notation of each relationship tags. Finally, a total of 2,434 sentences and 3,376 triples (2,542 for training and 834 for testing) are prepared. Among them, certain sentences without any tag are also kept improving the ability of MatRE to differentiate positive samples and hard negative samples.
- MaiSciRE [31]: This open-access dataset is focused on the materials science research related to batteries. It contains 5 different tags: "Conductivity", "Coulombic Efficiency", "Capacity", "Voltage", and "Energy". In total, 1,255 sentences and 7,027 triples (5,316 for training and 1,711 for testing) are prepared for analysis. It should be noted that one sentence may include more than one relation tags.

To effectively evaluate the capabilities of MatRE, three indicators are taken into consideration, including precision (P), recall (R) and F1-Score ($F1$). As the harmonic mean value of P and R, $F1$ plays dominant role in the evaluation. The detailed calculations are displayed in Eq. (10) ~ (12), where TP, TP and FN is the percentage of true positive, false positive, and false negative samples in the results, respectively. The parameter settings during the model training are provided in Supplementary Information S.2.

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{12}$$

3.2. Experimental results

3.2.1. Comparison of different baselines

We include the strong baselines as comparative models (*i.e.*, Word2vec+CNN+ATT [32], Word2vec+BiLSTM+ATT [33], R-BERT [34], D-BERT [35], SciBERT [36], MatBERT [37] and DeepSeek [38]). For MatRE, we leverage both BERT and MatSciBERT to encode the word embedding, to demonstrate the generalization and weak PLM-dependency. The corresponding models are distinguished by the subscripts (*i.e.*, MatRE_{BERT} and $\text{MatRE}_{MatSciBERT}$). The experimental results of MatRE and all comparative models on two different datasets are listed in Table 1, among which WV and ATT represent the word embedding features obtained through Word2vec [39] and target entity features extracted by attention mechanism [40]. It can be seen that MatRE_{BERT} has much better results compared to all the previous works including models built on top of larger PLMs, regardless of the different datasets. By replacing BERT with a material-specific encoder MatSciBERT, we can further improve the performance.

Compared to attention-based approaches (*e.g.*, WCA and WBA),

Table 1
The overall results on test set of two different datasets.

Models	NASICON dataset			MatSciRE dataset		
	P	R	$F1$	P	R	$F1$
WV+CNN+ATT(WCA) [32]	0.532	0.511	0.521	0.614	0.572	0.592
WV+BiLSTM+ATT(WBA) [33]	0.610	0.580	0.594	0.648	0.659	0.653
R-BERT [34]	0.647	0.673	0.660	0.737	0.712	0.724
D-BERT [35]	0.683	0.661	0.671	0.754	0.717	0.735
SciBERT [36]	0.767	0.713	0.729	0.783	0.818	0.800
MatBERT [37]	0.782	0.781	0.781	0.761	0.870	0.812
DeepSeek [38]	0.259	0.259	0.259	0.822	0.822	0.822
MatRE_{BERT}	0.798	0.793	0.791	0.890	0.878	0.880
$\text{MatRE}_{MatSciBERT}$	0.814	0.806	0.799	0.894	0.884	0.888

Algorithm 1
MatKA: Materials structure-activity relationship Knowledge Acquisition.

Input: Materials knowledge graph KG_{mat}; Relational type set $R = \{r_1, r_2, \cdots, r_k\}$; Target entity set $D = \{d_1, d_2, \cdots, d_n\}$; Candidate knowledge base $Cand_{know}$; Knowledge base KB

Output: Knowledge base KB
1: Initialize $KB = \{\}$ and $Cand_{know} = \{\}$
2: for $d_i, d_j \in D$ $(i \neq j)$ do
3: if d_i, d_j in KG_{mat} and $\mathrm{r}(d_i, d_j) \in R$:
4: $Cand_{know} = Cand_{know} \cup \{S : [d_i, d_j, \mathrm{r}(d_i, d_j)]\}$
5: end for
6: for $S_i \in Cand_{know}$ do
7: if $\exists Corr_{know}$ in S_i:
8: $KB = KB \cup \{Corr_{know} : [d_i, d_j]\}$
9: end for
10: return KB

MatRE$_{MatSciBERT}$ leads to an absolute $F1$ improvement of +27.8% and +20.5% on NASICON dataset. For MatSciRE dataset, the corresponding improvements are +29.6% and +23.5%, respectively. A major difference between MatRE and attention-based approaches is that the former not only incorporates the semantic information of the contexts based on attention mechanism, but also extracts the type and boundary information of the target entity through the introduction of entity-aware tokens. Therefore, this amelioration validates the advantage of such a dynamic entity-aware word embedding generation process over the simple combination of "static word embedding + attention".

Compared to the PLM based methods (*e.g.*, R-BERT, D-BERT, SciBERT and MatBERT), MatRE$_{MatSciBERT}$ also exhibits an average $F1$ improvement of +1.8% and 7.6% on the two datasets, respectively. As far as we know, PLM's representation on long sequence is not sufficient, since it may lose the semantic information of the local contexts. We believe the long-distance dependency of the sequence in BiGRU can compensate such semantic information loss and the early incorporation of entity information in the entity-aware process is also beneficial.

Compared to large language model methods (e.g., DeepSeek), MatRE$_{MatSciBERT}$ achieves $F1$ improvement of 54.0% and 6.6% on the two datasets respectively. Despite their strong generalization and contextual understanding capabilities, large language models cannot be directly applied in specialized materials science domains, as evidenced by their significantly lagged performance on the NASICON dataset. This further underscores the superior performance of this method.

To comprehensively evaluate the performance of MatRE on each relation, we list the confusion matrix of the models on two datasets in Fig. 3. The confusion matrix summarizes the predicted classification results for each relationship in the ground truth. For NASICON, nearly all samples in every relation type are correctly predicted (The darker the color in the diagonal, the better the predicted results). The light color of "Instance of" in the diagonal is mainly due to the small number of supporting samples. Similar cases also occur in MatSciRE dataset. To be specific, only a small number of samples contain "Conductivity" and "Energy", and certain entities do not belong to any relations. They have the same unit (*e.g.*, $S{\cdot}cm^{-1}$, V, and %) as the labelled entity, but cannot be labelled.

More straightforward conclusions can be drawn from Fig. 3b. MatRE achieves an average $F1$ of 0.830 on the 5 relation types in MatSciRE dataset. Except "Instance-Of" type, the F1-scores on other types in NASICON are all over 0.710 and the core type "Cause-Effect" display an $F1$ of 0.854. This indicates that MatRE can enable an accurate recognition of various structure-activity relationships in different materials systems.

The relatively low F1-Score on "Instance-Of" can be mainly attributed to the two following aspects. Firstly, positive and negative samples are not evenly distributed. As already discussed in Fig. 3a, small number of mispredictions can dramatically influence the overall results due to

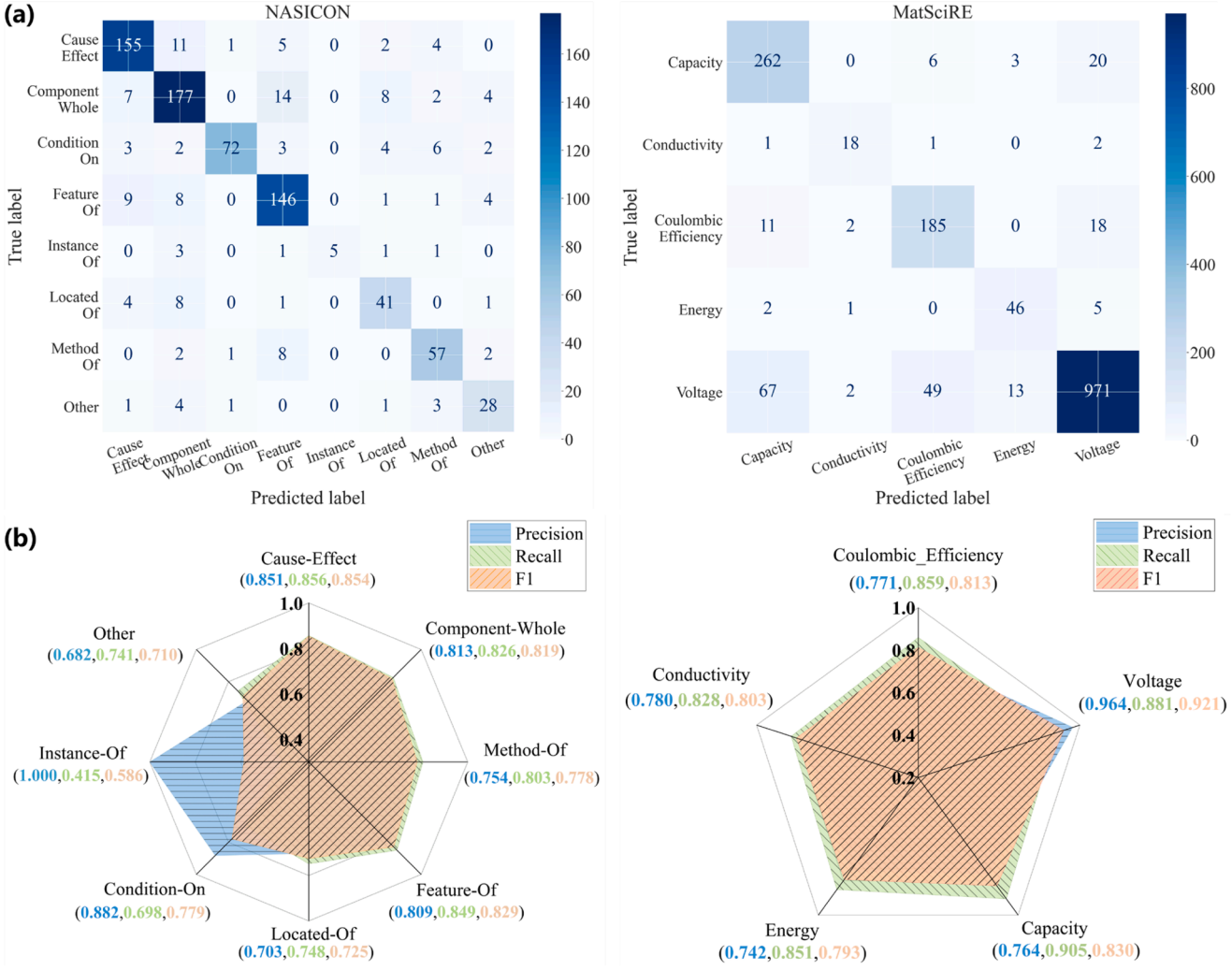


Fig.. 3. Confusion matrix and category analysis on the test set of two datasets. (a) the confusion matrix for both datasets, which is generated by comparing the relation types predicted by MatRE model (labels on the *x*-axis) with the actual types of the ground truth (labels on the *y*-axis). (b) the prediction results for each relation types in two datasets.

the limited number of "Instance-Of" relations in the test sets. Secondly, "Instance-Of" relation may suffer from noise samples. By definition, "Instance-Of" can be easily confused with other relation types, leading to plenty of mislabeled samples. Thus, it becomes difficult to construct high-quality negative triples to learn the accurate representation of "Instance-Of". In contrast, the similar relation of "Component-Whole" has a clear definition and the corresponding F1-score reaches 0.819.

3.2.2. Contribution of different modules

As mentioned in Section 3, the MatRE method consists of three modules. Ablation experiments are conducted to assess each module's contribution. Table 2 presents a comparison between our model and variants on NASICON dataset. For fair comparison, the module settings for all variants are kept the same. With the removal of the word embedding module, v1 yields the lowest F1-Score among all variants, highlighting its critical role. In comparison, v2, which restores the embedding while keeping other components unchanged, shows a substantial improvement, confirming the effectiveness of MatPLM. Unlike static embeddings such as Word2Vec, MatPLM is pre-trained on large-scale materials corpora and fine-tuned on relation extraction tasks, enabling it to capture domain-specific semantics and generate context-aware representations for material-related terms and synonyms. *F*1 drops by 4.4 % upon the removal of entity-aware tokens (v2), as these tokens enhance sensitivity to target entities by emphasizing critical information, such as types and boundaries. Furthermore, the different tokens for subject and object can clarify the direction in the relation, which significantly reduces false positives due to misleading directions. Excluding BiGRU (v3) further decreases performance, as BiGRU captures long-term dependencies and contextual semantics, addressing gaps left by MatPLM. BiGRU can be considered as a key to improve the performance, especially for long material sentences with complicated grammatical structures. By combining global features h_{global} in the relation classification stage, MatRE achieves more accurate results than v4. This is mainly because the fusion of word-level and sentence-level semantic information can not only increase the abundance of the triple representation, but also promote the prediction accuracy

4. Applications

Due to excellent thermal/chemical stability and feasible synthesis process, NASICON-type compounds remain in focus in the research of solid-state electrolyte and electrode materials used in rechargeable batteries [41–43]. Taking NASICON SSEs as an example, we investigate the application of the proposed method in the materials field.

4.1. Construction of NASICON-type materials KG

Using python web crawlers and Web of Science website 1808 NASICON SSE publications were selected and MatRE extracted 260,475 materials triples related to NASICON SSEs. As illustrated in Fig. 4, these triples are saved as Neo4j graph database.

As shown in Fig. 4, the constructed graph database contains plentiful information of materials entities (*i.e.,* nodes) and relations (*i.e.,* edges). The massive graph knowledge database can be constructed by connecting related material entities with the edges. Thus, the target entity and its adjacent nodes can be quickly located, that facilitates deduction of structure-activity relationships. Fig. 4b and 4d display the sub-graph structure and the deduction process related to the typical example of entity "*room temperature*". It can be seen that "*room temperature*" is connected with "*high mobility*", "*phase*", and "*ionic conductivity*" by the "Cause-Effect" relation (*i.e.*, gray edge). This demonstrates that room temperature is closely associated with the migration rate, phase, and ionic conductivity of the target material. Similarly, we also observe that the lattice parameters can influence the volume of the unit cell but are also dependent on the position of M element.

Driven by MatKGPtcl, we choose the ionic migration activation energy of NASICON SSEs as the target property and construct the logical relationship graph shown in Fig. 5. The data layer contains various types of materials entities sharing certain relations with the target property. The relations among different entities are also represented in this figure and ready to be extracted. We fill the graph based on the fast-searching function in Neo4j graph database and screen the triples in certain relation to activation energy driven by subject entity and object entity. The sub-graph is shown in Fig. 6. Similarly, we can also reveal all information connected with the target entity and set the foundations for the following knowledge acquisition of structure-activity relationship.

4.2. Knowledge acquisition and application of NASICON SSEs

To evaluate the ability of the acquired materials domain knowledge to aid ML models in reasonable decision-making, we establish two activation energy prediction datasets with 31 and 45 features (labelled as $Dataset_{31}$ and $Dataset_{45}$ [24]). The two datasets share identical samples but differ in the features employed. Detailed sample information is provided in Supplementary Section S.6.3. These datasets are used as research objects to extract the structure–activity relationship knowledge embedded across the features.

As shown in Table 3, we first take the union set of the features in $Dataset_{31}$ and $Dataset_{45}$ as the searching criterion. Based on the procedure of Algorithm 1, we search other entities in certain relation with these features from the materials KG of NASICON SSEs by treating the union set as the target entity set. The obtained relations and corresponding original sentences are incorporated into the candidate knowledge base and saved in the quadruple form of (*Entity 1, Entity 2, Relation, Source*). Certain examples are displayed in Table 4 and the complete information can be seen in Supplementary Information S.3. After that, we perform deduction on sentences in the candidate knowledge base: if there is any term from the vocabulary regarding relation or influencing rules, the sentence is transferred to the final knowledge base. Under the guidance of domain experts, we convert the entity-relation information saved in the final knowledge base into 24 structure-activity knowledge bullets about NASICON SSEs, which are listed in Table 5.

The 24 knowledge bullets mentioned above provide valuable insight for ML models embedded with materials domain knowledge and can be applied throughout the entire life cycle of ML. In this paper, we select two methods developed by our group for the data pre-processing and feature engineering stages in ML as the object of knowledge application:

Table 2
Comprehensive comparison of different variants of MatRE on NASICON dataset.

Variants	Word Embedding		BiGRU Module	Global Feature	*P*	*R*	*F*1
	MatPLM	Entity-aware Tokens					
v1			✓	✓	0.614	0.603	0.608
v2	✓		✓	✓	0.769	0.742	0.755
v3	✓	✓		✓	0.786	0.771	0.778
v4	✓	✓	✓		0.761	0.728	0.744
MatRE	✓	✓	✓	✓	0.814	0.806	0.799

* The word embeddings of v1 are derived from Word2vec.

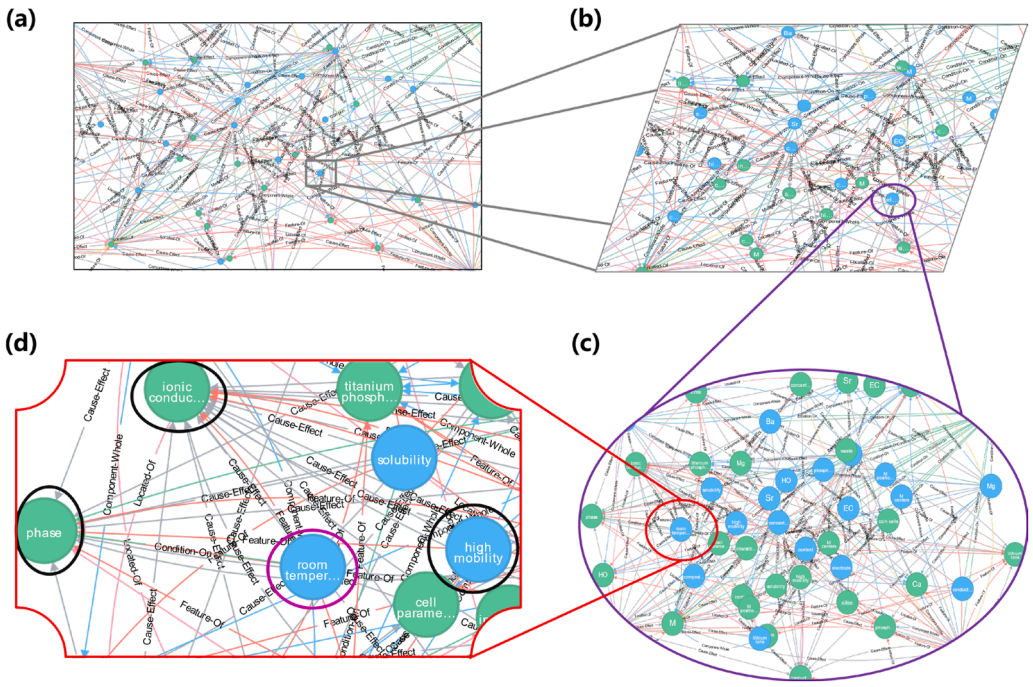


Fig.. 4. The visualization of NASICON SSEs triples based on Neo4j. (a) an overview of the materials KG. (b), (c) and (d) represent the sub-graph centered on "room temperature" in the graph step-by-step.

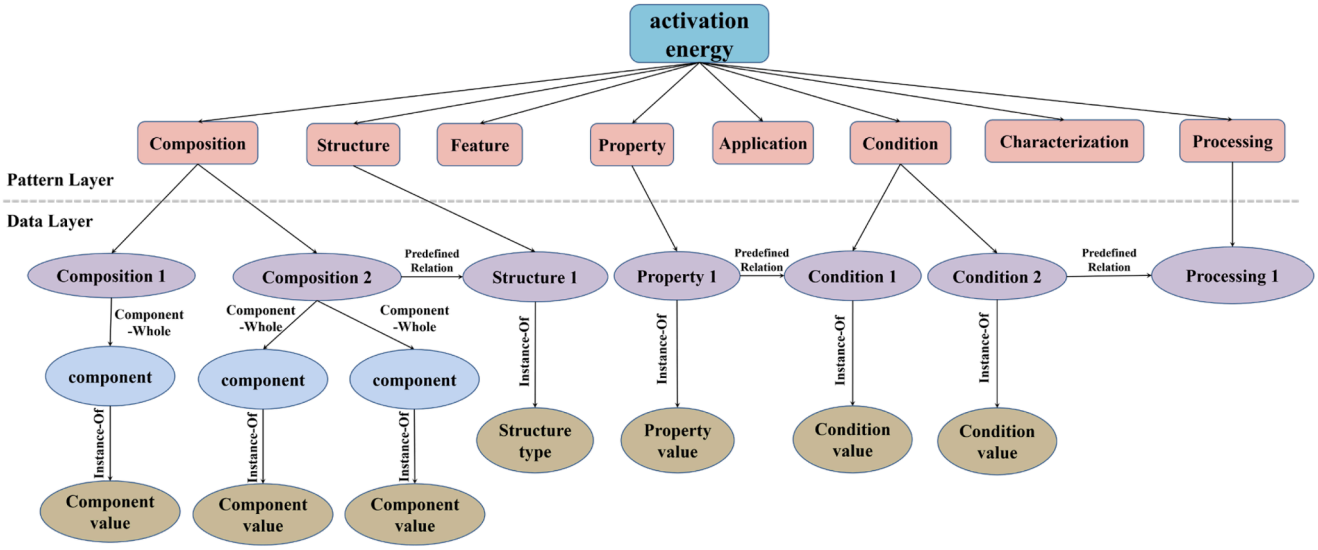


Fig.. 5. Logical relationship graph of NASICON SSEs driven by activation energy.

Data Accuracy Detection Method Incorporating Materials Domain Knowledge ($DADM_{mdk}$) [44] and Feature Selection method embedded with Non-Co-Occurrence Rules (NCOR-FS) [45]. For $DADM_{mdk}$, we employ this knowledge as a source of the relationships between descriptors and target attributes that are necessary in the second phase of the method and use Pearson correlation coefficients to complete multi-dimensional correlation detection of the materials datasets to identify outliers. For NCOR-FS, we transfer the 24 knowledge bullets listed in Table 5 into 27 non-co-occurrence rules among features (details are shown in Supplementary Information S.4), which are embedded into the modeling of NCOR-FS to enlarge the feature selection effects. The process of embedding materials knowledge into the machine-learning model for diffusion barrier prediction is exemplified by NCOR-FS. It combines the relationship between descriptors in the field of materials knowledge with data-driven correlation analysis techniques to obtain non-co-occurrence relationships between descriptors. For example, it is known that there exists a positive correlation between the “lattice parameter” and “unit cell volume”. The combined use of “lattice parameter” and “cell volume” as candidate descriptors yields a non-co-occurrence rule, $R = \{F_1, F_2\}$, where $F_1 = \{lattice\,parameter\}$ and $F_2 = \{cell\,volume\}$. Complete implementation can be found in the Supplementary Information S.4. During the process, 6 ML models (*i.e.*, LASSO, GPR, Ridge, SVR, KNN, and RF) are used to predict the activation energy. Moreover, Root Mean Square Error (*RMSE*), Mean Absolute Percentage Error (*MAPE*) and R^2 are used to evaluate the prediction accuracy before and after knowledge embedding on two datasets, respectively.

Herein, we show the results of feature selection on $Dataset_{31}$

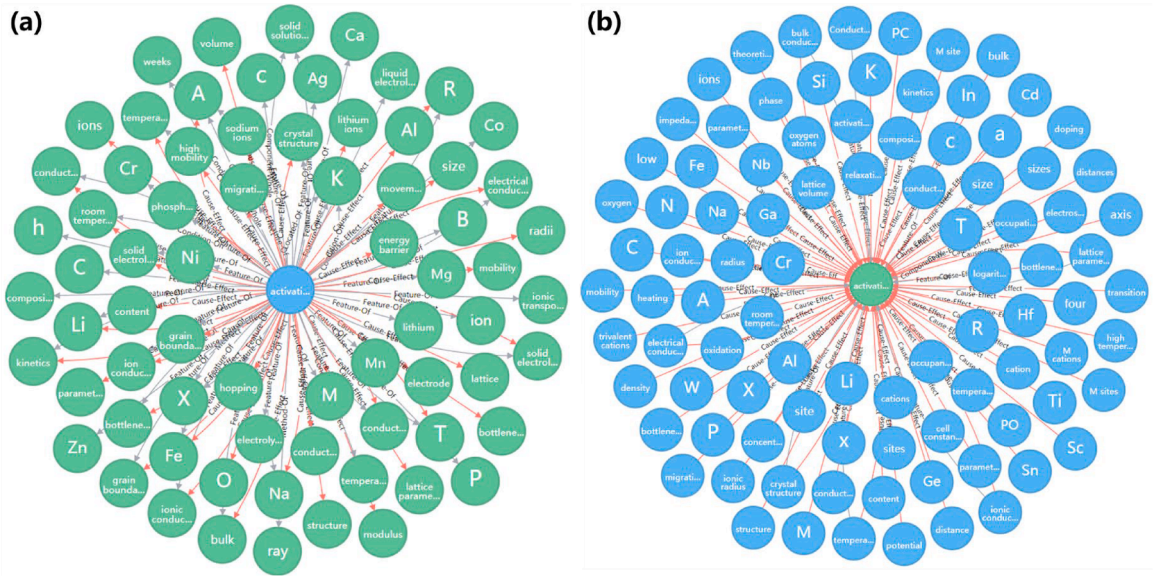


Fig.. 6. Entities and relations related to activation energy in the KG. (a) the subgraph with activation energy as the subject entity. (b) the subgraph with activation energy as the object entity.

Table 3
The union set of features in $Dataset_{31}$ and $Dataset_{45}$.

No.	The union set of features	Description
1	*a*, c, a/c, d, h	Lattice parameter
2	V	Cell volume
3	R_M1, R_M2, Avg_M_R, R_X1, R_X2, Avg_X_R, RT	Ionic radius
4	EN_M1, EN_M2, Avg_M_EN, EN_X1, Avg_X_EN	Electronegativity
5	V_MO_6, V_XO_4, V_Na(1)O_6, V_Na(2)O_8, V_Na(3)O_5	Volume of polyhedron
6	D2_stoich, D3_stoich, Na_stoich, X1_stoich, X2_stoich	Stoichiometric number
7	O_Na1, O_Na2, O_Na3, O_M1, O_M2, O_X1, O_X2	Occupancy ratio
8	C_{Na}	Na^+ concentration
9	V_M1, V_M2, Avg_M_V, V_X1, V_X2, Avg_X_V	Valence of element
10	BT1, BT2, Min_BT	Minimum bottleneck
11	E_Na(1), E_Na(2), E_Na(3), E_Na, E_M, E_X	Configurational entropy
12	T	Temperature

Table 4
Examples of the candidate knowledge base.

Entity 1	Entity 2	Relation	Source
lattice parameters	cell volume	Cause-Effect	1. As the Mo6+ doping content increasing, less Na ions are introduced into the crystal lattice that the lattice parameters of a and c and cell volume increase, as seen in Fig b. 2. There was a very slight increase in lattice parameter a, and hence an increase in lattice volume after immersion for all stability tests.
lattice parameters	bottleneck	Cause-Effect	1. The lattice parameters a and c, and the unit cell volume increased with as the X value increased for this system, which likely increased the bottleneck size in the ionic pathway.
occupancy	activation energy	Cause-Effect	1. At still higher lithium contents, the increase in M site occupancy is accompanied by a gradual rise in activation energy, up to num kJ / mol.

(Table 6) and data accuracy detection on $Dataset_{45}$ (Table 7). The remaining results are shown in Supplementary Information S.5. It can be seen that a better prediction result can be achieved after knowledge embedding, regardless of the ML models and the datasets. For feature selection, knowledge extracted by MatRE enables NCOR-FS to identify more highly correlated features that effect model prediction accuracy, thereby enhancing the performance of machine learning models [45]. KNN behaves the best for $Dataset_{31}$ with *RMSE*, *MAPE*, and R^2 of 0.068, 0.052, and 0.928. These values are improved by 13.9%, 25.7%, and 1.6% comparing to the ones without knowledge embedding. For data accuracy detection, the knowledge obtained through MatRE is used to eliminate five samples with obvious annotation errors and correct three samples containing input anomalies. This process improves the overall quality of the dataset, leading to more accurate prediction of material properties [44]. The optimum model is Random Forest (RF). The *RMSE*, *MAPE*, and R^2 are improved by 15.0%,10.3% and 1.9% after knowledge embedding. In both tasks, the performance improvement brought by knowledge embedding is multi-dimensional. The *RMSE* metric achieves an improvement of over 10% in both tasks, indicating a significant reduction in the overall prediction error magnitude and enhanced robustness in error control. The *MAPE* sees a maximum improvement of 25.7%, reflecting increased consistency in the model's relative prediction error. The increase in R^2 indicates stronger explanatory power for true-value variations, though the improvement is limited due to the already high original R^2 baseline. The overall optimization of total error, relative error, and the explanatory power for data variation further supports the meaningfulness of the obtained structure-activity relationships for NASICON SSEs. The application scenario of this knowledge on the ML assisted materials science research co-driven by both data and knowledge are also investigated. As the last point, we need to emphasize that MatRE and the method for KG construction and knowledge acquisition are universal and can be applied to other classes of materials. They are of great significance to explore interpretable relations among materials and to avoid misleading KG from black box.

5. Conclusion and outlook

A substantial body of structure-activity relationship (SAR) knowledge exists within materials science literature, but its extraction remains time-consuming and labor-intensive. To address this, we propose an automated framework to extract SAR knowledge tailored to materials

Table 5
List of 24 structure-activity knowledge bullets of NASICON SSEs.

No.	Relational triples	Structure-activity relationships
1	("lattice parameter", Cause-Effect, "cell volume")	(lattice parameter, positive correlation, cell volume)
2	("lattice parameter", Cause-Effect, "bottleneck")	(lattice parameter, negative correlation, bottleneck)
3	("lattice parameter", Cause-Effect, "occupancy ratio")	-
4	("bottleneck", Cause-Effect, "activation energy")	(bottleneck, negative correlation, activation energy)
5	("ionic radius", Cause-Effect, "occupancy ratio")	-
6	("temperature", Condition-On, "activation energy")	(temperature, negative correlation, activation energy)
7	("ionic radius", Cause-Effect, "volume")	(ionic radius, positive correlation, volume)
8	("ionic radius", Feature-Of, "activation energy")	(the ionic radius of *M* element, negative correlation, activation energy)
9	("bottleneck", Cause-Effect, "occupancy ratio")	-
10	("ionic conductivity", Cause-effect, "activation energy")	(ionic conductivity, negative correlation, activation energy)
11	("ionic concentration", Cause-Effect, "bottleneck")	-
12	("channel size", Cause-Effect, "activation energy")	(channel size, negative correlation, activation energy)
13	("occupancy ratio", Cause-Effect, "electronegativity")	-
14	("temperature", Condition-On, "volume")	(temperature, positive correlation, volume)
15	("ionic radius", Cause-Effect, "bottleneck")	-
16	("Na^+ concentration", Cause-Effect, "activation energy")	(Na^+ concentration, negative correlation, activation energy)
17	("temperature", Condition-On, "ionic radius")	-
18	("occupancy ratio", Cause-Effect, "activation energy")	(the occupancy of *X* element, negative correlation, activation energy)
19	("configuration entropy", Cause-Effect, "occupancy ratio")	-
20	("temperature", Condition-On, "occupancy ratio")	(temperature, positive correlation, occupancy ratio)
21	("valance", Cause-Effect, "occupancy ratio")	-
22	("volume", Feature-Of, "activation energy")	(cell volume, negative correlation, activation energy)
23	("configuration entropy", Cause-Effect, "activation energy")	(configuration entropy, negative correlation, activation energy)
24	("electronegativity", Cause-Effect, "activation energy")	-

* "-" represents that no explicit correlation knowledge is deduced from source sentence.

Table 6
Feature selection results before and after knowledge embedding on $Dataset_{31}$.

ML models	FS without knowledge embedding			FS with knowledge embedding		
	RMSE	*MAPE*	R^2	*RMSE*	*MAPE*	R^2
LASSO	0.079	0.071	0.910	0.074	0.058	0.925
GPR	0.097	0.070	0.842	0.102	0.094	0.851
Ridge	0.080	0.073	0.894	0.076	0.067	0.922
SVR	0.085	0.076	0.889	0.084	0.072	0.903
KNN	0.079	0.070	0.913	0.068	0.052	0.928
RF	0.081	0.072	0.906	0.078	0.069	0.920

tasks. Firstly, a materials entity-aware relational extraction model (MatRE) is proposed to extract entity-relation triples in a simple yet effective manner. Subsequently, the structure-activity relationship knowledge acquisition method is designed to deduce implicit knowledge, which is then represented as a KG in the database. Experimental results on two distinct materials RE datasets showcase the exceptional performance of MatRE, achieving an up to 16% improvement in F1-Score compared to conventional methods. To validate the effectiveness of proposed methods, we present an application example of NASICON SSEs. Thereinto, 260,475 triples are identified from 1,808 relevant literature sources, and then 24 structure-activity knowledge bullets are derived from the constructed NASICON-type materials KG. The knowledge is transformed into rules for easily embedding into ML modelling, which enables 6 ML models to gain impressive accuracy and interpretability. This framework can be adapted to extraction of any other relational information from text, given that new training data or the definition of new relation types is provided to cover other domains adequately.

Table 7
Data accuracy detection results before and after knowledge embedding on $Dataset_{45}$.

ML models	Original dataset			Corrected dataset with knowledge		
	RMSE	*MAPE*	R^2	*RMSE*	*MAPE*	R^2
LASSO	0.065	0.041	0.928	0.058	0.035	0.943
GPR	0.057	0.039	0.945	0.052	0.037	0.954
Ridge	0.058	0.040	0.943	0.051	0.033	0.956
SVR	0.070	0.060	0.915	0.071	0.057	0.916
KNN	0.100	0.066	0.830	0.079	0.051	0.894
RF	0.060	0.039	0.939	0.051	0.035	0.957

Improving automatic extraction of structure-activity relationships is an ongoing effort, involving retrieval of the corpora to generate high-quality data for pretraining the deep-learning model and annotation of diversiform relation types for refining it. In near future, the utilization of large-scale language models (*e.g.*, GPT-4) to compute word embeddings or conduct extraction with "*Prompt-Completion*" manner directly holds immense promise in further improving the acquisition of scientific knowledge. Furthermore, in scenarios of material text mining where entities are not pre-specified in the text, pipelined NER and RE approaches are prone to cascading impacts of errors. Future research will focus on joint named entity recognition and relation extraction methods to enhance performance in such material knowledge acquisition contexts. In addition, the model can extract relations across sentences by linking entities that co-occur in different sentences. Incorporating knowledge reasoning techniques may further enhance the ability to infer such cross-sentence relations in the future. This enhancement could in turn facilitate more accurate acquisition of knowledge required by downstream tasks, thereby addressing the current situation of limited R^2 improvement and ultimately boosting their performance. We hope this work will pave the way towards making the vast amount of information accumulated in scientific literature more accessible, and we believe is beneficial and essential for machine-assisted breakthroughs in materials science.

Data & code availability

The *NASICON SSEs RE* dataset, which was constructed as part of this study, is available at https://github.com/orphanwu/DataSet/. This dataset includes material-related texts, the entities extracted from these texts, and the corresponding relationships between the entities. Further details can be found at the link provided above. For the source of MatSciRE dataset, please visit study [31].

CRediT authorship contribution statement

Yue Liu: Writing – review & editing, Project administration, Methodology, Investigation, Conceptualization. **Dahui Liu:** Writing – original draft, Software, Methodology, Data curation. **Zhengwei Yang:** Writing – review & editing, Investigation. **Xianyuan Ge:** Writing – original draft, Validation. **Wenxuan Yao:** Validation, Investigation. **Jie Wu:** Investigation. **Maxim Avdeev:** Writing – review & editing. **Siqi**

Shi: Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

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

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 92270124, 92472207 and 52073169) and the National Key Research and Development Program of China (Grant No. 2021YFB3802101). We appreciated the High-Performance Computing Center of Shanghai University, Shanghai Engineering Research Center of Intelligent Computing System and Key Laboratory of Silicate Cultural Relics Conservation (Shanghai University), Ministry of Education for providing the computing resources and technical support.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ensm.2025.104390.

Data availability

Data will be made available on request.

References

[1] S.V. Kalinin, D. Mukherjee, K. Roccapriore, B.J. Blaiszik, A. Ghosh, M.A. Ziatdinov, A. Al-Najjar, C. Doty, S. Akers, N.S. Rao, J.C. Agar, Machine learning for automated experimentation in scanning transmission electron microscopy, Npj Comput. Mater. 9 (2023) 227–242, https://doi.org/10.1038/s41524-023-01142-0.
[2] Y. Liu, Z. Yang, X. Zou, S. Ma, D. Liu, M. Avdeev, S. Shi, Data quantity governance for machine learning in materials science, Natl. Sci. Rev. 10 (2023) nwad125, https://doi.org/10.1093/nsr/nwad125.
[3] P. Xu, X. Ji, M. Li, W. Lu, Virtual sample generation in machine learning assisted materials design and discovery, J. Mater. Inf. 3 (2023) 16, https://doi.org/10.20517/jmi.2023.18.
[4] L. Zhang, J. Zhou, X. Chen, Data-driven exploration and first-principles analysis of perovskite material, J. Mater. Inf. 4 (2024) 13, https://doi.org/10.20517/jmi.2024.20.
[5] O. Kononova, T. He, H. Huo, A. Trewartha, E.A. Olivetti, G. Ceder, Opportunities and challenges of text mining in materials research, Science 24 (2021) 102155, https://doi.org/10.1016/j.isci.2021.102155.
[6] Y. Liu, D. Liu, X. Ge, Z. Yang, S. Ma, Z. Zou, S. Shi, A high-quality dataset construction method for text mining in materials science, Acta. Phys. Sin. 72 (2023) 070701, https://doi.org/10.7498/aps.72.20222316.
[7] Z. Pei, J. Yin, P.K. Liaw, D. Raabe, Toward the design of ultrahigh-entropy alloys via mining six million texts, Nat. Commun. 14 (2023) 54–61, https://doi.org/10.1038/s41467-022-35766-5.
[8] R. Leaman, C.H. Wei, Z. Lu, tmChem: a high performance approach for chemical named entity recognition and normalization, J. Cheminf. 7 (2015) 1–10, https://doi.org/10.1186/1758-2946-7-S1-S3.
[9] M.C. Swain, J.M. Cole, ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature, J. Chem. Inf. Model. 56 (2016) 1894–1904, https://doi.org/10.1021/acs.jcim.6b00207.
[10] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal, A. Valencia, Information retrieval and text mining technologies for chemistry, Chem. Rev. 117 (2017) 7673–7761, https://doi.org/10.1021/acs.chemrev.6b00851.
[11] E. Kim, K. Huang, S. Jegelka, E. Olivetti, Virtual screening of inorganic materials synthesis parameters with deep learning, Npj Comput. Mater. 3 (2017) 53–61, https://doi.org/10.1038/s41524-017-0055-6.
[12] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, Materials synthesis insights from scientific literature via text extraction and machine learning, Chem. Mater. 29 (2017) 9436–9444, https://doi.org/10.1021/acs.chemmater.7b03500.
[13] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, Machine-learned and codified synthesis parameters of oxide materials, Sci. Data 4 (2017) 1–9, https://doi.org/10.1038/sdata.2017.127.
[14] A.C. Vaucher, F. Zipoli, J. Geluykens, V.H. Nair, P. Schwaller, T. Laino, Automated extraction of chemical synthesis actions from experimental procedures, Nat. Commun. 11 (2020) 3601, https://doi.org/10.1038/s41467-020-17266-6. -1611.
[15] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K.A. Persson, G. Ceder, A. Jain, Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, J. Chem. Inf. Model. 59 (2019) 3692–3702, https://doi.org/10.1021/acs.jcim.9b00470.
[16] T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, G. Ceder, Similarity of precursors in solid-state synthesis as text-mined from scientific literature, Chem. Mater. 32 (2020) 7861–7873, https://doi.org/10.1021/acs.chemmater.0c02553.
[17] L. Hawizy, D.M. Jessop, N. Adams, P. Murray-Rust, ChemicalTagger: A tool for semantic text-mining in chemistry, J. Cheminf. 3 (2011) 1–13, https://doi.org/10.1186/1758-2946-3-17.
[18] C.J. Court, J.M. Cole, Auto-generated materials database of curie and néel temperatures via semi-supervised relationship extraction, Sci. Data 5 (2018) 1–12, https://doi.org/10.1038/sdata.2018.111.
[19] F. Kuniyoshi, K. Makino, J. Ozawa, M. Miwa, Annotating and extracting synthesis process of all-solid-state batteries from scientific literature, Proc. Twelfth Lang. Resour. Eval. Conf. (2020) 1941–1950. https://aclanthology.org/2020.lrec-1.239.
[20] W. Wang, X. Jiang, S. Tian, P. Liu, D. Dang, Y. Su, T. Lookman, J. Xie, Automated pipeline for superalloy data by text mining, Npj Comput. Mater. 8 (2022) 9–20, https://doi.org/10.1038/s41524-021-00687-2.
[21] 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, https://doi.org/10.1038/s41586-019-1335-8.
[22] Z. Nie, S. Zheng, Y. Liu, Z. Chen, S. Li, K. Lei, F. Pan, Automating materials exploration with a semantic knowledge graph for Li-ion battery cathodes, Adv. Funct. Mater. 32 (2022) 2201437, https://doi.org/10.1002/adfm.202201437.
[23] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A.S. Rosen, G. Ceder, K.A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nat. Commun. 15 (2024) 1418–1431, https://doi.org/10.1038/s41467-024-45563-x.
[24] Y. Liu, X. Ge, Z. Yang, S. Sun, D. Liu, M. Avdeev, S. Shi, An automatic descriptors recognizer customized for materials science literature, J. Power Sources 545 (2022) 231946, https://doi.org/10.1016/j.jpowsour.2022.231946.
[25] D. Fernandes, J. Bernardino, Graph databases comparison: Allegrograph, arangoDB, Infinitegraph, neo4j and orientD, Data 18 (2018) 373–380, https://doi.org/10.5220/0006910203730380.
[26] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/n19-1423.
[27] T. Gupta, M. Zaki, N.A. Krishnan, Mausam, MatSciBERT: A materials domain language model for text mining and information extraction, Comput. Mater. 8 (2022) 102, https://doi.org/10.1038/s41524-022-00784-w.
[28] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proc. 2014 Conf. Empirical Methods Nat. Lang. Process, 2014, pp. 1724–1734, https://doi.org/10.3115/v1/D14-1179.
[29] Z.C. Lipton, J. Berkowitz, C. Elkan, A critical review of recurrent neural networks for sequence learning, ArXiv Preprint (2015) 150600019, https://doi.org/10.48550/arXiv.1506.00019.
[30] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[31] A. Mullick, A. Ghosh, G.S. Chaitanya, S. Ghui, T. Nayak, S.C. Lee, S. Bhattacharjee, P. Goyal, Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction, Comput. Mater. Sci. 233 (2024) 112659, https://doi.org/10.1016/j.commatsci.2023.112659.
[32] Y. Shen, X.J. Huang, Attention-based convolutional neural network for semantic relation extraction, 26th Int, Conf. Comput. Linguist. (2016) 2526–2536. https://aclanthology.org/C16-1238.
[33] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proc. 54th Annu. Meet. Assoc. Comput. Linguist, 2016, pp. 207–212, https://doi.org/10.18653/v1/P16-2034.
[34] S. Wu, Y. He, Enriching pre-trained language model with entity information for relation classification, Proc. ACM Int. Confer. Inf. Knowl. Manage. (2019) 2361–2364, https://doi.org/10.1145/3357384.3358119.
[35] Y. Huang, Z. Li, W. Deng, G. Wang, Z. Lin, D-BERT: Incorporating dependency-based attention into BERT for relation extraction, CAAI Trans. Intell. Technol. 6 (2021) 417–425, https://doi.org/10.1049/cit2.12033.
[36] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proc. 2019 Confer. Empirical Methods Nat. Lang. Process. 9th Int. Jt. Confer. Nat. Lang. Process, 2019, pp. 3613–3618, https://doi.org/10.18653/v1/D19-1371.
[37] A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder, A. Jain, Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science, Patterns 3 (2022) 100488, https://doi.org/10.1016/j.patter.2022.100488.
[38] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint (2025) 250112948, https://doi.org/10.48550/arXiv.2501.12948.
[39] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, ArXiv preprint (2013) 13013781, https://doi.org/10.48550/arXiv.1301.3781.
[40] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, Int. Confer. Learning Representations, ArXiv preprint (2014), https://doi.org/10.48550/arXiv.1409.0473, 1409.0473.

[41] R. Thirupathi, V. Kumari, S. Chakrabarty, S. Omar, Recent progress and prospects of NASICON framework electrodes for Na-ion batteries, Prog. Mater. Sci. 137 (2023) 101128, https://doi.org/10.1016/j.pmatsci.2023.101128.

[42] P. Wu, W. Zhou, X. Su, J. Li, M. Su, X. Zhou, B.W. Sheldon, W. Lu, Recent advances in conduction mechanisms, synthesis methods, and improvement strategies for $Li_{1+x}Al_xTi_{2-x}(PO_4)_3$ solid electrolyte for all-solid-state lithium batteries, Adv. Energy Mater. 13 (2023) 2203440, https://doi.org/10.1002/aenm.202203440.

[43] Y. Yang, S. Yang, X. Xue, X. Zhang, Q. Li, Y. Yao, X. Rui, H. Pan, Y. Yu, Inorganic all-solid-state sodium batteries: electrolyte designing and interface engineering, Adv. Mater. 36 (2023) 2308332, https://doi.org/10.1002/adma.202308332.

[44] S. Shi, S. Sun, S. Ma, X. Zou, Q. Qian, Y. Liu, Detection method on data accuracy incorporating materials domain knowledge, J. Inorg. Mater. 37 (2022) 1311–1320. https://www.jim.org.cn/EN/10.15541/jim20220149.

[45] 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, https://doi.org/10.1016/j.actamat.2022.118195.