Journal of Power Sources 545 (2022) 231946

Contents lists available at ScienceDirect

Journal of Power Sources

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

An automatic descriptors recognizer customized for materials science literature

Yue Liu [a,b], Xianyuan Ge [a], Zhengwei Yang [a], Shiyu Sun [d], Dahui Liu [a], Maxim Avdeev [e,f], Siqi Shi [c,d,*]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China
[b] Shanghai Engineering Research Center of Intelligent Computing System, Shanghai, 200444, China
[c] State Key Laboratory of Advanced Special Steel, School of Materials Science and Engineering, Shanghai University, Shanghai, 200444, China
[d] Materials Genome Institute, Shanghai University, Shanghai, 200444, China
[e] Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC, NSW, 2232, Australia
[f] School of Chemistry, The University of Sydney, Sydney, 2006, Australia

HIGHLIGHTS

- Proposing a text data augmentation method incorporating materials domain knowledge.
- Customizing an automatic descriptors recognizer from materials literature.
- Constructing activation energy prediction models of NASICON-type electrolytes.

ARTICLE INFO

Keywords:
Materials design
Descriptor
Natural language processing

ABSTRACT

Materials science literature contains domain knowledge about numerous descriptors, which play a critical role in data-driven materials design. However, automatically extracting descriptors from literature remains challenging. Here, we develop an automatic descriptors recognizer based on natural language processing (NLP) to mine latent descriptors, which consists of a conditional data augmentation model incorporating materials domain knowledge (cDA-DK), coarse- and fine-grained descriptors subrecognizers (CGDR and FGDR). cDA-DK conducts augmenting training data of text mining model, which can significantly reduce the cost of manually labeling and enhance the robustness of its model. On this basis, CGDR recognizes coarse-grained descriptor entities automatically, and FGDR performs screening of fine-grained descriptors related to specific materials design. Following this, the activation energy of NASICON-type solid electrolytes, which is influenced by complicated descriptors, is taken as an example to demonstrate the potential utility of our recognizer. CGDR extracts 106896 descriptor entities from 1808 relevant articles with an accuracy ($F1$) of 0.87. Furthermore, with features from 408 descriptors screened by FGDR, six activation energy prediction models are constructed to perform experiments, achieving an optimal prediction performance (R^2) of 0.96. This work provides important insight towards the understanding of structure-activity relationships, thus promoting materials design and discovery.

1. Introduction

Structure-activity relationships are the key to optimizing properties of materials and discovering novel materials. With the progress in materials research, a large amount of data has been accumulated. Accordingly, data-driven machine learning (ML) methods have been widely used to establish structure-activity relationships from historical materials data in recent years [1–3]. Selection of descriptors is critical for the ML process, which determines the upper limit of ML performance [4,5]. Realizing rapid and effective selection of descriptors, consequently, is vital for exploring material structure-activity relationships.

The selection of descriptors in the research for material structure-

* Corresponding author. State Key Laboratory of Advanced Special Steel, 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.jpowsour.2022.231946
Received 5 July 2022; Received in revised form 30 July 2022; Accepted 1 August 2022
Available online 18 August 2022
0378-7753/© 2022 Elsevier B.V. All rights reserved.

activity relationships almost always relies on manual operation with expert knowledge. For instance, Jalem et al. [6] selected chemical components (charge and coordination number of elements) and crystal structure (lattice constants, crystal cell volume, bond lengths and bond angles of polyhedra, and interatomic distances) descriptors to predict the diffusion barrier and cohesive energy of an olivine-type LiMXO_4 by summarizing related knowledge and literature; then, based on their previous work, chemical components (element radius, ionization energy, melting point, boiling point, and vaporization enthalpy) and crystal structure (bond valence parameters) were selected as new descriptors, which yields better prediction results [7]. Sendek et al. [8] manually screened 12831 lithium-containing crystalline solids for those with high structural and chemical stability, low electronic conductivity, and low cost from the Material Project database. After that, 21 descriptors that related to crystal structure and chemical composition were screened by conductivity, which is used to explore the structure-composition-conductivity relationships of solid electrolytes for lithium-ion batteries. Xu et al. [9] constructed a prediction model of ionic conductivity based on the ionic conductivity data of 70 $R3\overline{c}$ space groups of NASICON compounds, which are obtained by extrapolation from Arrhenius equation using the data from a large literature collection. 16 chemical component descriptors and 12 crystal structural descriptors were empirically selected to explore the structure-activity relationships with conductivity. In their study, selecting related descriptors requires laborious and tedious literature search, performed manually by domain experts. For this reason, the subjectivity of various experts and the limitation of the particular expert knowledge seriously affect the selection of descriptors with the growing amount of published materials data. In this context, realizing automatic selection of descriptors from literature can not only greatly improve the rate of research but also alleviate the subjective factors of experts. Moreover, materials informatics studies often make predictions for hundreds or thousands of materials [10–12], thus, it is extremely useful to be able to extract descriptors from materials literature automatically to promote further development of the materials field [13,14].

Named Entity Recognition (NER) [15] provides the capability of extracting information from unstructured text automatically. Typically, this type of task is regarded as a supervised ML problem, namely, a model learns to identify the keywords in a sentence. In the field of materials science, NER has been applied in information extraction tasks including inorganic materials synthesis recipes [16–19], inorganic materials mentions [20–23], and organic materials knowledge [24]. Recently, Ceder et al. [25] captured latent knowledge through unsupervised word embeddings derived from Word2vec approach [26], which points a novel path to extracting materials knowledge and relationships from scientific literature. On this basis, they achieved automatic extraction of large-scale inorganic material information and solid-state synthesis information by manually annotating a large amount of supervised data and then training a deep learning NER model (BiLSTM-CRF) [27,28]. To the best of our knowledge, there have been no attempts to employ NER to mine for descriptors from materials literature. It is worth noting that the annotation of high-quality supervised data is laborious and time-consuming in most cases. Furthermore, Yimam et al. [29] found that it is difficult for the unsupervised Word2vec approach to fully understand the material vocabulary compared to the pre-trained BERT model through a large number of comparative experiments. The reason is that the Word2vec generates word embedding that is context-independent (static embeddings) and does not possess complex characteristics (e.g., syntax, semantics). In summary, NER has become one of the main approaches to mining word or phrase information from materials literature [30–32]. Nevertheless, with the increase of computer computing power, pre-trained models are more and more popular in the field of text mining, which makes a large number of scholars use BERT model to obtain word embedding nowadays [33–36].

In this study, we develop an automatic descriptors recognizer (ADR) customized for materials literature, which aims to provide an objective and effective method to obtain descriptors. The descriptors recognizer consists of a conditional data augmentation model incorporating materials domain knowledge (cDA-DK), coarse- and fine-grained subrecognizers (CGDR and FGDR, respectively). The cDA-DK model is built to significantly reduce the overhead of manually labeling data and enhance the robustness of text mining model. After that, the CGDR constructs a NER model (MatBERT-BiLSTM-CRF) of multi-level semantic feature extraction which is a neural network trained on obtained dataset. Then, the FGDR screens descriptors by two strategies, which are descriptors co-occurrence with performance features of the target material and the importance of the current descriptor in their corresponding sentences. Overall, our work can be summarized as the following four points:

(1) Through collecting related literature, a small size of materials NER dataset is hand-annotated to train the NER model. On this basis, the cDA-DK model is constructed to augment training data and enhance the robustness of downstream NER model.
(2) Based on word embedding obtained by fine-tuned MatBERT model, a NER model named MatBERT-BiLSTM-CRF is constructed, which is capable of automatically extracting coarse-grained descriptors, including materials composition, structure, property, processing, external condition, and so forth. Meanwhile, a descriptor knowledge base (KB) is built to store these descriptors.
(3) Taking activation energy prediction as an example, two screening strategies are designed to screen fine-grained descriptors from the extracted results of the previous step, which are high-quality and relevant to activation energy prediction.
(4) Two activation energy prediction datasets are constructed with 31 and 45 features selected from screened descriptors, and then ML models are built to execute prediction experiments.

The remainder of the paper is organized as follows. Section 2 proposes the general pipeline of potential descriptors extraction and introduces the methods in detail. Section 3 presents the experimental results and their corresponding analysis. Section 4 gives the specific application by an example of predicting activation energy of NASICON-type solid electrolyte materials. Finally, Section 5 concludes this paper and gives an outlook for future work.

2. Methods

2.1. The pipeline of latent descriptors recognizer

Instead of acquiring descriptors manually, we develop an automatic descriptors recognizer (ADR) to provide an automatic pipeline that is composed of Data Processor, Coarse-Grained Descriptors Subrecognizer (CGDR), and Fine-Grained Descriptors Subrecognizer (FGDR) as shown in Fig. 1, and the details as follows:

Firstly, Data Processor aims to construct sufficient and high-quality learning samples through the process of data collection & preprocessing, tokenization & labeling, and data augmentation. The data collection & preprocessing collects text data from literature and preprocesses them (the detailed collection and processing scheme is outlined in detail in the Supporting Information S.1). Then tokenization & labeling designs labels to decide what types of descriptor information need to be recognized and annotates with the processed data. Especially, a conditional Data Augmentation model incorporating materials Domain Knowledge (cDA-DK) is proposed to augment annotated dataset, which can effectively reduce the workload of manual annotation and greatly improve the generalization ability of the NER model of CGDR.

Secondly, CGDR, which consists of the processes of Token-Level Feature Extraction, Local Context Feature Extraction, and Knowledge Base Construction, plays a critical role in this pipeline. It can extract

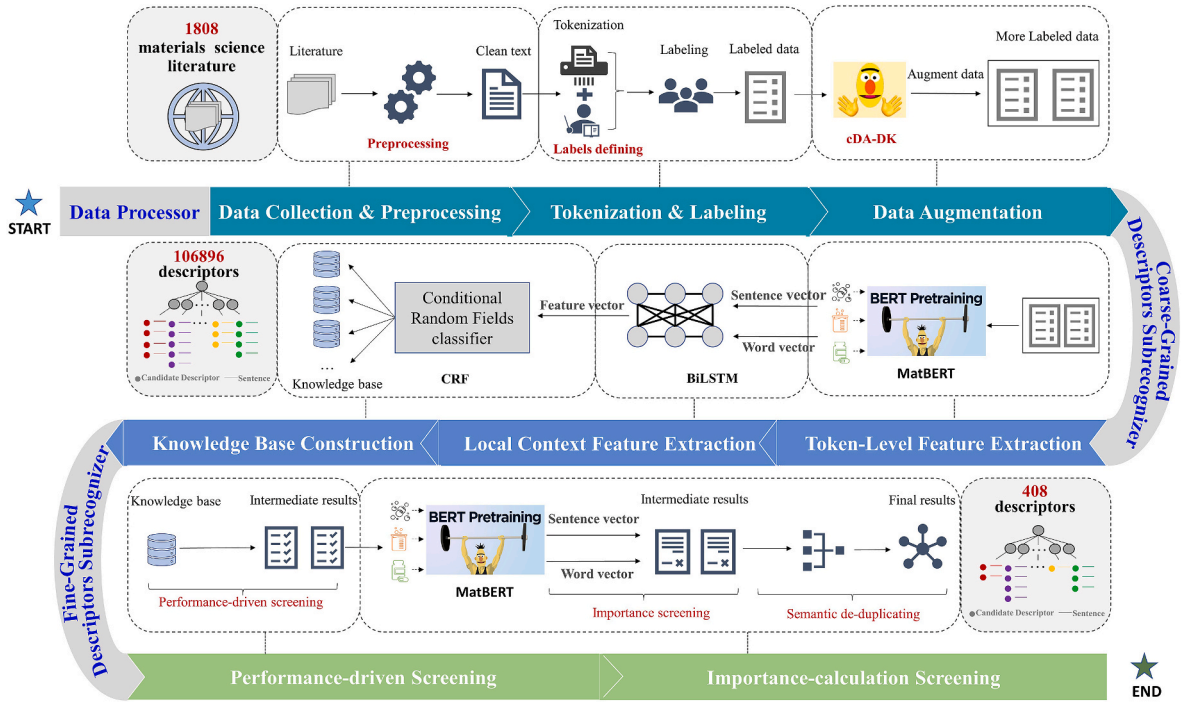

Fig. 1. The overall pipeline of automatic descriptors recognizer. The pipeline involves three stages of Data Processor, coarse-grained descriptors subrecognizer (CGDR), and fine-grained descriptors subrecognizer (FGDR). Data Processor aims to construct sufficient and high-quality learning samples. CGDR constructs a NER model, which can extract numerous descriptor entities from materials literature by training labeled data provided by Data Processor. FGDR performs screening fine-grained descriptors related to the target materials property, which can help to construct samples for further research of materials design or structure-activity relationships.

numerous descriptor words or phrases from materials literature by training labeled data provided by Data Processor. It is worth noting that the CGDR is realized by constructing a NER model of multi-level semantic feature extraction, in which MatBERT-based semantic feature extraction for token level (Material Bidirectional Encoder Representations from Transformers implements semantic extraction for token level) primarily represents materials text as the vector of words with semantic information, BiLSTM-based semantic feature extraction for local context (Bi-directional Long Short-Term Memory implements feature extraction within the local context of sentences) is responsible for capturing the local contextual features of words, and CRF-based descriptors classification (Condition Random Field implements descriptors classification) serves to classify words. The final classification results are stored in the constructed knowledge base.

Finally, to rapidly screen fine-grained descriptors related to the target materials property, the coarse-grained descriptor entities in the knowledge base are fed to FGDR, which is designed with performance-driven screening and importance-calculated screening. Performance-driven screening is performed by retrieving descriptors co-occurrence with the performance features of the target material, whereas importance-calculated screening is executed through calculating the importance of descriptors in the corresponding sentence and deduplicating based on semantic information. Moreover, with the descriptors selected by FGDR, we can construct samples for further research of materials design or structure-activity relationships.

2.2. Data processor with conditional data augmentation incorporating materials domain knowledge

In the study of structure-activity relationships, there are multiple classes of descriptors that influence a particular property. For example, activation energy of NASICON-type solid electrolyte materials is often influenced by descriptors such as composition, structure, process, property, and so on. Therefore, to obtain labeled learning samples for the specific descriptor acquiring task, the design of tags is vital, which determines the descriptor information to be recognized. Here, refer to Ref. [27], we design eight entity tags of descriptor, including “Composition”, “Structure”, “Property”, “Processing”, “Characterization”, “Application”, “Feature” and “Condition”. With these tags, the dataset for training CGDR can be constructed through hand-annotating materials literature. The description and corresponding examples for each tag, and detailed annotation scheme are given in Supporting Information (S.2).

Training the CGDR model requires a large amount of annotated data, however, the process of data annotation is laborious and time-consuming. Data augmentation (DA) [37–40], a technique for improving the performance and accuracy of ML models under data-constrained conditions, is expected to be an effective solution. Textual DA is commonly achieved by replacing words with synonyms based on dictionaries or translating to a different language and back [41]. However, there are no synonym dictionaries in the material research field, and the method of translating inevitably generates more noisy data, which may affect the quality of generated data. Towards this challenge, we propose a conditional data augmentation model incorporating materials domain knowledge (cDA-DK), which can learn the characteristics of materials text data and dynamically generate high-quality data.

The specific process of cDA-DK is shown in Fig. 2. As shown, cDA-DK mainly employs the DistilRoBERTa model (Roberta’s [42] model of knowledge distillation) to massively augment textual data. Herein, instead of using this pre-trained model directly, we first incorporate materials text knowledge to fine-tune it. After that, the fine-tuned DistilRoBERTa model can capture the contextual semantic information of materials text. For example, a sentence to be augmented, “*The ionic conductivity decreases with increasing activation energy*” and the corresponding label of each token in this sentence is fed into fine-tuned DistilRoBERTa model. Through the fine-tuned DistilRoBERTa model, some words of the sentence are randomly masked. Here, the masked

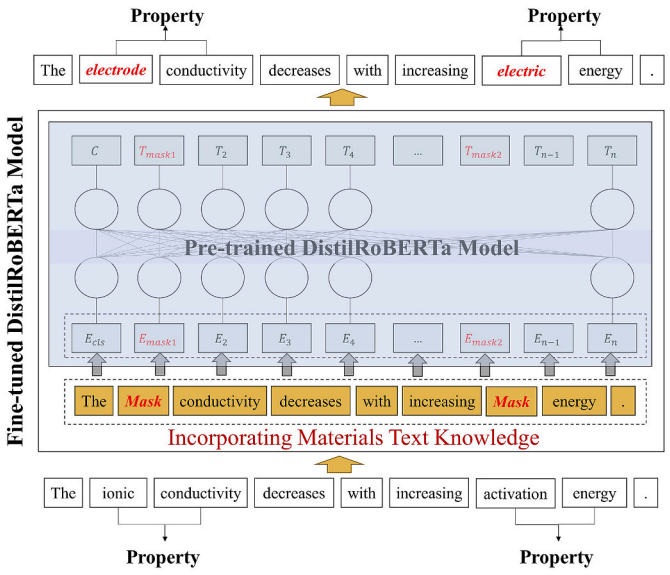

Fig. 2. Material data augmentation based on cDA-DK. The hand-annotated NER dataset is augmented to address the problem of insufficient training samples and improve the robustness of our model.

sentence is represented as "*The* ***Mask*** *conductivity decreases with increasing* ***Mask*** *energy*", and then is transformed into vectors as E_{cls}, E_{mask1}, E_2, E_3, $E_4, \ldots, E_{mask2}$, E_{n-1}, E_n for learning contextual semantic information by the fine-tuned DistilRoBERTa model. Finally, fine-tuned DistilRoBERTa model generates semantic vectors C, T_{mask1}, T_2, T_3, T_4, $\ldots, T_{mask2}$, T_{n-1}, T_n through contextual semantic information it learned and translates into text data "*The electrode conductivity decreases with increasing electric energy*". Note that, the augmented data is entirely dependent on the semantic knowledge of materials learned by the fine-tuned DistilRoBERTa model which can greatly decrease the noisy data.

2.3. Mining descriptor entities by coarse-grained descriptors subrecognizer

The selection of descriptors in current studies for specific materials is mostly based on empirical knowledge of experts, which has certain subjectivity and limitations. To this end, we design CGDR to train a NER model in such a way that descriptor words or phrases of diverse types can be automatically recognized from materials text. For example, this model is expected to learn that the phrases "activation energy" and "ionic conductivity" represent coarse-grained descriptors of the "Property" class, whereas "tetrahedra" and "polyhedra" represent the "Structure" class. Therefore, we propose a NER model of multi-level semantic feature extraction to address the above problems as shown in Fig. 3, including MatBERT-based semantic feature extraction for token level, BiLSTM-based semantic feature extraction for local context, and CRF-based classification for words or phrases. This NER model can accurately recognize which words or phrases correspond to a specific descriptor entity type through training.

For MatBERT-based semantic feature extraction for token level, the MatBERT is utilized to extract token-level features, and then these features are represented as semantic vectors. The MatBERT model is derived from the pre-trained Bidirectional Encoder Representation from Transformers (BERT) model [33] fine-tuned with materials text, and it can learn the characteristics of materials text as well as get vector representation of words with abundant semantic information in this process. Through analyzing materials text, it is found that the same words or similar meaning words in different contexts may express very different meanings. For example, the word "bottleneck" not only means limited development but also can indicate the structural information of materials crystal, which shows that the context of the word is very important. Therefore, the MatBERT is used to encode materials text since it can fully capture contextual information of words (i.e., word embedding, segment embedding, and position embedding), then, get the vector representation of words with richer semantic information.

For BiLSTM-based semantic feature extraction for local context, the BiLSTM model is employed to perform the capture of contextual features of materials text. As a sequence tagging problem, NER belongs to the token-level classification task, which means that each word in a sentence needs to be classified. Therefore, it is necessary to take the local context of each word in a sentence into account. For example, in the sentence "*The overall _____ is near to* 10^{-5} *S·cm*$^{-1}$ *at 200°C*", it is presumed that the missing word is likely to be "conductivity" (a coarse-grained descriptor of the "Property" class) based on the contextual information of the missing position. Although the position information introduced by MatBERT has made up for the local context information, the self-attention mechanism of MatBERT weakens the position information during fine-tuning [43]. To this end, the recurrent neural network (RNN) is employed for settling the issues mentioned above, which has the capability of capturing timing information for sequence-to-sequence classification. Nevertheless, RNN often suffers from vanishing gradient and exploding gradient problems in the propagation process of timing information, and thus we introduce a variant of the RNN called long short-term memory (LSTM) [44]. Here, the bidirectional LSTM (Bi-LSTM) is used to capture both forward and backward context information and reinforce position information.

For CRF-based descriptors classification, the CRF predicts the optimal tag sequence to achieve more accurate entity classification by learning the dependencies between tags. Conditional random fields (CRF), as a classifier of sequence tagging problems, can capture the strong interdependence of output labels, so that the optimal label sequence can be obtained. The entity labels of each word in a sentence need to be predicted through the classifier, and there are often certain transfer relations between neighboring entity labels. Therefore, it is useful to consider the correlation between labels in the neighborhood and decode the best label chain for a given input sentence. To this end, the CRF model is set as the classifier layer of NER, rather than a typical

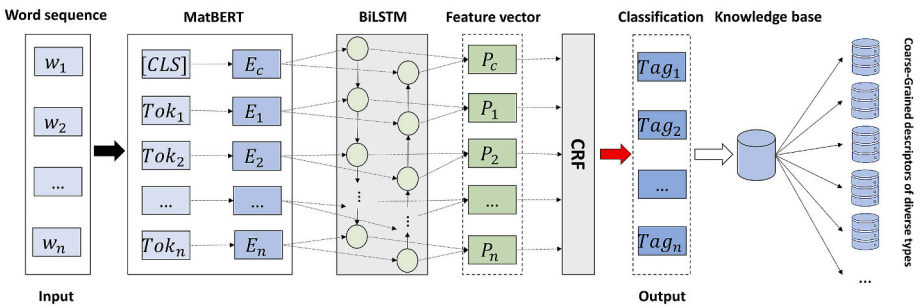

Fig. 3. NER model of CGDR. It receives a sequence of sentences and then extracts the coarse-grained descriptors entities from the sequence to save them in the knowledge base.

softmax layer.

Moreover, in order to facilitate researchers to rapidly obtain relevant descriptors used for the study of materials property prediction or structure-activity relationships, we built a knowledge base to store the recognized result of CGDR as shown in Fig. 4. The knowledge base is composed of eight sub-bases corresponding to eight diverse types of coarse-grained descriptors. The content stored in each sub-base is coarse-grained descriptors of the corresponding type and the sentences where the descriptor appears. Furthermore, dynamically adding descriptors and their corresponding sentences to the knowledge base is also implemented.

2.4. Performance-driven and importance-calculated descriptors screening by fine-grained descriptors subrecognizer

Through CGDR, we extract a large number of coarse-grained descriptors of diverse types from materials literature and save them in a knowledge base. Then, FGDR is constructed to further realize the prediction of specific material property and the research of their structure-activity relationships through selecting relevant descriptors from the knowledge base. The method and workflow of FGDR are shown in Fig. 5. As shown, it mainly works by combining performance-driven screening and importance-calculated screening for acquiring fine-grained descriptors. The former aims to determine the property of the studied target material, and then screens descriptors co-occurred in the same sentence with it; whereas the latter mainly aims to screen high-quality descriptors related to it for the result of the previous step as far as possible. The screened result can help researchers to construct the dataset of descriptors. Using this dataset, ML models can perform materials property prediction or establish structure-activity relationships.

To screen descriptors associated with the studied target materials property, we develop a performance-driven screening rule as follows:

$$R = \{ < D_i, S_i > | (S_i \in KB) \vee ((D_T, D_i) \in S_i) \vee (D_T \neq D_i) \} \quad (1)$$

where R represents the screened result, D_T represents descriptor of target materials property, KB is the knowledge base built in Section 2.3, S_i represents *ith* sentence containing D_T in KB, D_i represents *ith* descriptor in S_i.

Commonly, the performance-driven screening rule needs first to determine the property of the studied target material; then, this rule is used to query the knowledge base and screens descriptors and corresponding sentences that co-occur in the same sentence with it. This way, more fine-grained descriptors can be further screened from the knowledge base.

However, through analyzing the results of performance-driven screening, we find descriptors that co-occurred in the same sentence are not necessarily related, which can affect the quality of screened descriptors. For improving the quality of screened descriptors, we develop an importance-calculated screening strategy. Specifically, the descriptors are screened by the importance relative to their sentence, as shown in Fig. 5. To calculate the importance of each descriptor in the corresponding sentence, we take out the words and corresponding sentence vector of the last layer of MatBERT to do the inner product; and then perform normalization for the result of the previous step with the softmax function to get the final importance. The calculation formula of normalization is shown in Eq. (2). Note that, the MatBERT model here is the same as that in Section 2.3, except that the latter does not need to output the vector of words and corresponding sentence, but feeds to the downstream model for further feature extraction.

$$I_i = \frac{E_i \cdot S_{[CLS]}}{\sum_{i=0}^{n+1} E_i \cdot S_{[CLS]}} \quad (2)$$

where I_i represents the importance of *ith* word, E_i is the embedding vector of *ith* word output by MatBERT, and $S_{[CLS]}$ is the corresponding sentence embedding vector. After that, an importance threshold T_I is set by us to screen descriptor w_i, as shown in Fig. 5. Furthermore, to deduplicate descriptors that have similar semantics in the previous screened results as much as possible, the cosine similarity of any two descriptors is calculated, and if it is less than the T_S (preset similarity threshold) both are retained, otherwise only of them is retained. Based on the above strategy, we can finally screen relatively high-quality fine-grained descriptors.

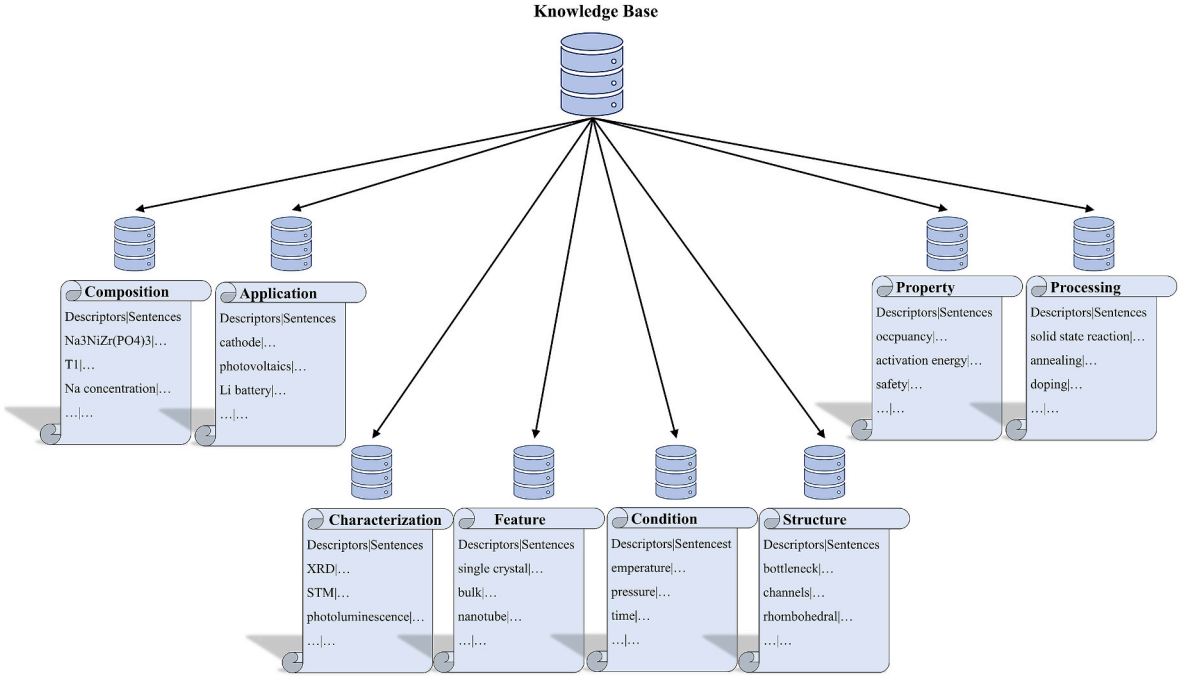

Fig. 4. The structure of the knowledge base. It consists of eight sub-bases corresponding to eight diverse types of coarse-grained descriptors.

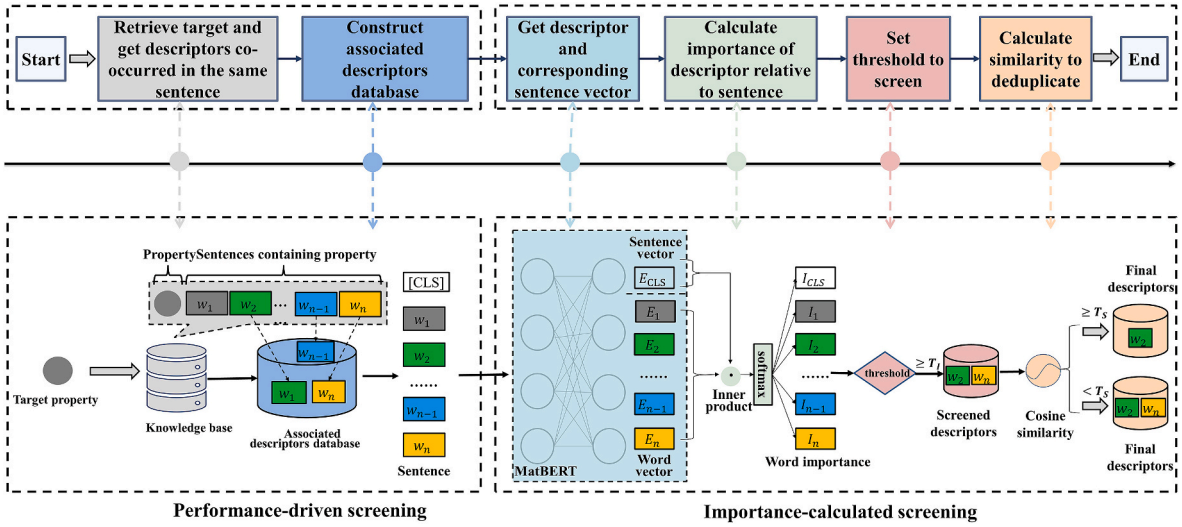

Fig. 5. The method and workflow of FGDR for screening fine-grained descriptors.

3. Experiments

3.1. Experimental dataset

In this study, the NER dataset for the research of automatic descriptors extraction is constructed by hand-annotating due to the lack of such datasets in materials science. The details are provided below: Firstly, through retrieving Crystallographic Information Files (CIFs), 55 materials literature sources, containing a large number of descriptors, are identified. Secondly, through Data Processor (shown in Fig. 1), literature text is extracted and preprepared. After that, the preprepared text is annotated to form NER dataset which includes eight descriptor entity types, 65690 tokens, and 2434 sentences. Fig. 6 shows the distribution of sentence lengths. As can be seen that sentences in materials text are generally long (more than 20 words), which demonstrates the complexity of materials text. Finally, to further improve the generalization ability of the NER model, the hand-annotated dataset is augmented with cDA-DK, and then further experiments are performed on combining the hand-annotated and the augmented datasets.

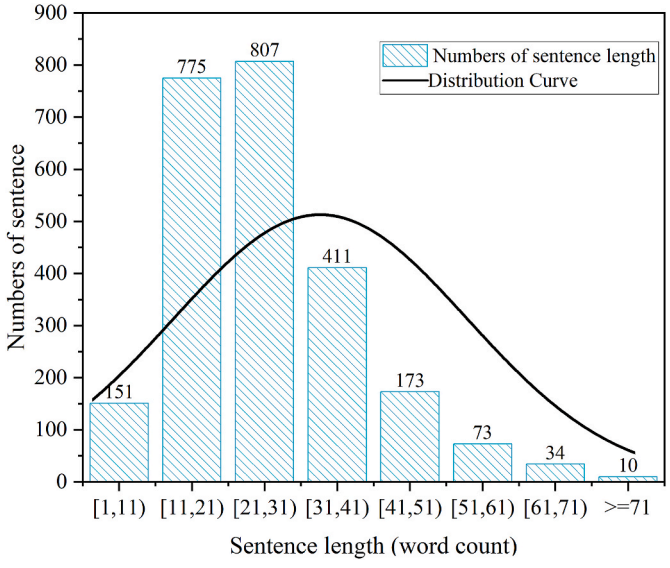

Fig. 6. The sentence distribution of our NER dataset.

3.2. Experimental setup

For valid evaluation of the model, Precision (P), Recall (R), and F1 score are used as the indexes of our experiment. Among them, the F1 score, which is the harmonic mean of *P* and *R*, plays a major role. The *P*, *R*, and *F*1 are calculated as shown in Eqs. (3)–(5), where *TP*, *FP*, and *FN* represent the rates of true positives, false positives, and false negatives, respectively. Moreover, information on parameter configuration for training the relevant model can be found in Supporting Information S.3.

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2P^{*}R}{P + R} \tag{5}$$

3.3. Experimental results and analysis

3.3.1. Performance of cDA-DK

In this subsection, two types of datasets are constructed to simulate a low-resource scenario for validating the performance of cDA-DK. Specifically, Dataset 1 consists of Top 2K sentences from the hand-annotated dataset while Dataset 2 consists of Top 1K sentences combined with augmented 1K sentences data. In addition, Dataset 3 is constructed to further demonstrate the effectiveness of cDA-DK. All experiments are carried on MatBERT-BiLSTM-CRF. The experimental results are shown in Table 1. As shown, the performance of the model trained on Dataset 2 outperforms that on Dataset 1, in which Precision, Recall, and F1 score are improved by 5%, 9%, and 8%, respectively. This phenomenon may be caused by the fact that cDA-DK has a powerful ability to learn and capture contextual semantic information of complex materials text, and then generate high-quality data. Specifically, the pre-trained DistilRoBERTa model in cDA-DK is fine-tuned through incorporating materials text knowledge, so that it learns the complex characteristic of

Table 1
The comparative accuracy metrics on different data conditions.

Dataset	Description	Precision	Recall	F1
Dataset 1	Top 2K	0.77	0.78	0.77
Dataset 2	Top 1K + Augmented 1K	0.82	0.87	**0.85**
Dataset 3	Original all + Augmented all	0.86	0.87	**0.87**

materials text and can dynamically generate high-quality domain text data. The above results demonstrate the effectiveness of cDA-DK for text data augmentation. Therefore, we augment all the original data with cDA-DK and perform experiments on this basis. The performances (Precision, F1 score) on Dataset 3 are improved by 4% and 2%, respectively.

3.3.2. CGDR recognition performance

Fig. 7 shows the overall experimental performance of the NER model on the test set. As shown that the total F1 score of our model is 0.87, which is fairly close to the state-of-the-art (SOTA) NER model (F1 score of 0.92) [33]. However, we cannot directly compare the above two tasks because the datasets for training and evaluating have significant differences. The dataset for the SOTA model is constructed based on common newspaper articles with only three entity labels, and our model is constructed on complex material articles with eight entity labels. As a result, our task is obviously much more challenging. It is worth noting that the results of the cDA-DK model on Dataset 3 are the same as the NER model. The reason is that they are the same experiment, which plays different roles. The former is to further investigate if the cDA-DK model works better on the entire dataset, so that demonstrates its effectiveness. However, the latter is to show the validity of the NER model of CGDR. In addition, the F1 score of each entity class is all beyond 0.80 except for the "Application" class, which indicates that the model performs well in recognizing descriptors of diverse types. The model gets poor performance for discriminating the class of "Application", which is because the model cannot capture the plenty of features from insufficient samples of "Application" to discriminate it.

The MatBERT-BiLSTM-CRF model, then, is compared with the current mainstream NER model that is BiLSTM-CRF [45], BiLSTM-CNNs-CRF [27], and BERT [33], of which performances can be seen in Fig. 8(a). As shown, the F1 score of our model is 0.87, which is higher than other existing models. For further demonstrating the ability of MatBERT to dynamically capture the semantic information of complex materials text compared to the Word2vec, meanwhile removing the bias that our model performs well only on a single dataset (55 hand-annotated materials literature by us), we conduct comparative experiments on a public dataset (800 hand-labeled abstracts [27]). The results of comparison experiment are displayed in Table 2, it is quite clear that the F1 score of our model is 4% higher than BiLSTM-CNNs-CRF [27]. As a result, the above experimental results prove that the introduction of MatBERT can extract more sufficient contextual semantic features of words, and then better represent the semantic information of words. Furthermore, the BiLSTM can fully capture the local context semantic information in the sentence sequence.

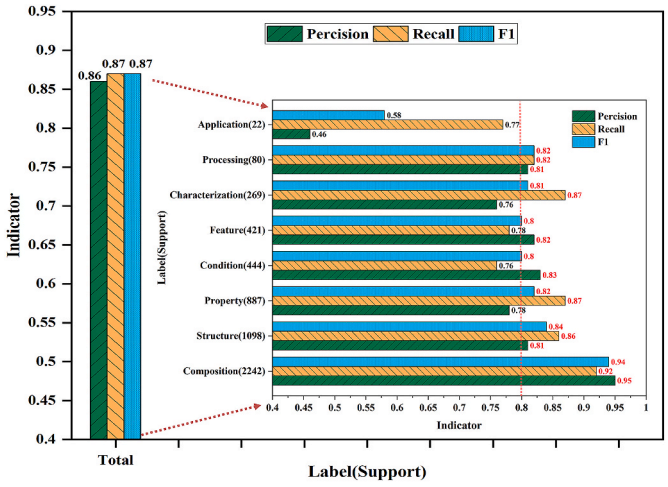

Fig. 7. The overall accuracy metrics for our model on the test set.

Therefore, our model demonstrates better entity recognition performance.

Moreover, to evaluate the components of the MatBERT-BiLSTM-CRF (MB-BC) model, ablation experiments are designed. As shown in Fig. 8 (b), the MB-BC is divided into two situations that are BC (Word vectors are obtained by traditional Word2vec training instead of MatBERT) and MB-C (The BiLSTM network is missing) model, which tests the contribution of MatBERT and BiLSTM to MB-BC, respectively. As a result, the evaluation metrics of MB-BC are affected to varying degrees. For BC, the F1 score decreases by 16%, which indicates that MatBERT can dynamically generate embedding vectors with richer semantics based on context information. For another, the F1 score decreases by 8%, which indicates that BiLSTM can improve the long-range dependence of words and capture more sufficient local context semantic information.

Through the trained NER model of CGDR, descriptor information is extracted from materials text accurately. Fig. 8(c) displays examples of the comparison of prediction results. As shown, in the sentence "*For those NASICON materials which show a phase transition, the activation energy differed at low temperature (LT) and high temperature (HT).*", it can be seen that "NASICON materials" is a descriptor of "Feature" class, "phase transition" and "activation energy" belong to "Property" class, and "low temperature" and "high temperature" belong to "Condition" class. Almost all entities are correctly predicted by our model, in contrast to the other models. For example, as for the BiLSTM-CNNs-CRF, "NASICON materials" and "phase transition" are incorrectly recognized as "Application" and "Structure" classes respectively by the comparison model. For the sentence "*Hence the effect of pressure on NaZr2(PO4)3 is found to be primarily due to the size of alkali cation.*", it can be seen that "pressure" is a descriptor of the "Property" class, "NaZr2(PO4)3" belongs to "Composition" class, and "alkali cation" belongs to "Feature" class. The "alkali cation" is incorrectly recognized as the "Composition" class by the comparison model and "pressure" has not been recognized. The above result demonstrates that our method is more suitable for the recognition of material entities.

To validate the robustness of the NER model of CGDR, a new test set was constructed with additional hand-annotated 35 materials literature sources. Their overall experimental performances on the initial and new test sets are shown in Table 3. As shown, the total F1 score on the new test set is 0.86, which is comparable to the NER model training on the initial test set. Moreover, the F1 score on the new test set achieves significant improvement in the "Application" and "Condition" classes, by 10% and 3% respectively. For other entity classes, the F1 score fluctuates within the normal range (±3%), which shows that the model owns ideal classification performance even though the distribution of datasets differs from the training set. That is, this model still maintains optimal performance during encountering other NER datasets in the materials field.

The confusion matrix of our model on the initial and new test set is illustrated in Fig. 9(a and b). As shown, the prediction results of all the samples in each entity class are almost accurate, with more correct samples of prediction in the "Characterization" and "Condition" classes for the initial test set, 2507 and 1811 respectively. Whereas for the new test set, "Characterization" and "Condition" classes have more correct samples of prediction, with 2484 and 1855 respectively. Note that the predicted correct results of the "Other" class are not considered, since the prediction is focused on descriptor classes. However, 70 samples and 67 samples of the initial and new test set respectively originally belonging to the "Structure" class are misclassified as the "Condition" class, which may be caused by the fact that there are many bond values for elements in the chemical formula of materials in the new test set (materials text) that can easily be mistaken for the values of external condition (e.g., values for temperature and pressure, etc.). Moreover, 69 samples of the initial test set originally belonging to the "Characterization" class are misclassified as the "Condition" class. This may be caused by the fact that the "Characterization" and "Condition" classes have similar contextual information in the material text (i.e., they are usually

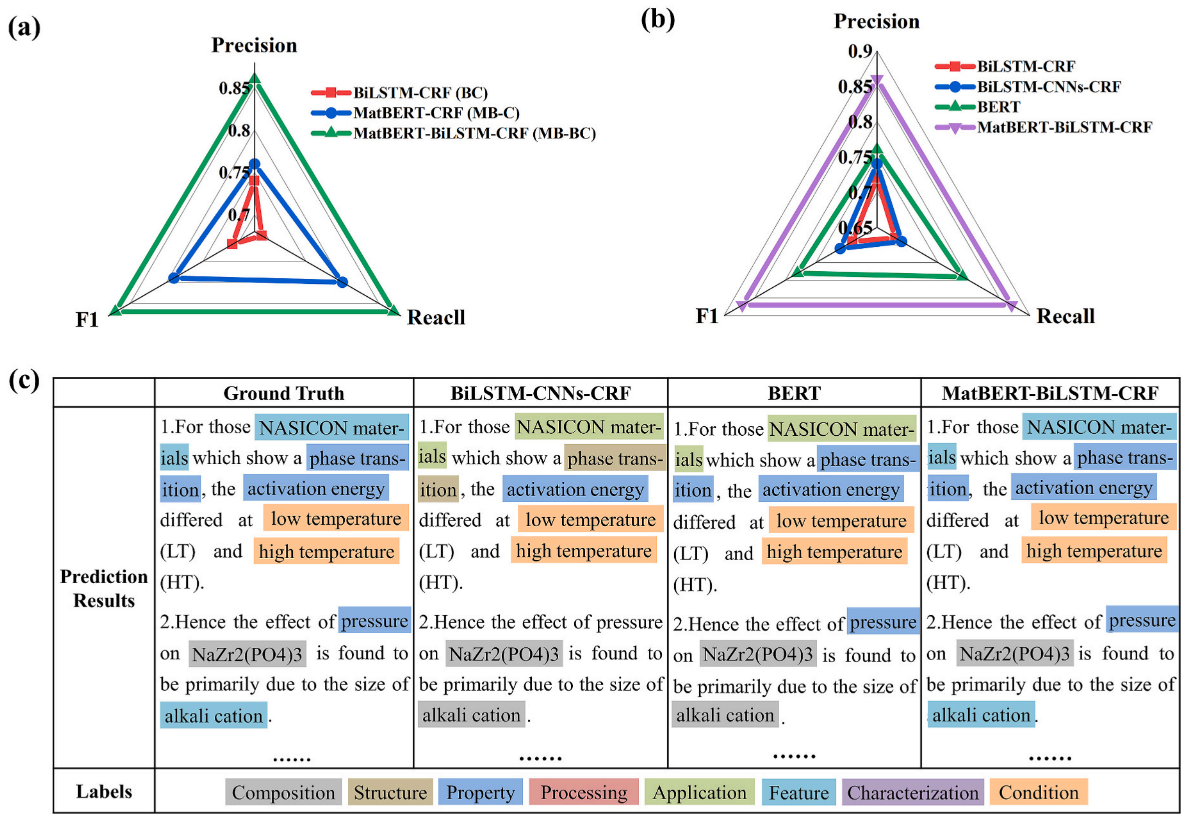

Fig. 8. (a) The comparative accuracy metrics for benchmark and our model. (b) Comparison of part contribution in MB-BC model. (c) The comparison of example prediction results between the benchmark and our model. The highlighting indicates regions of text that the corresponding model has associated with a particular descriptor entity type.

Table 2
The comparative accuracy metrics for different models on 800 hand-annotated abstracts.

Model	F1 (800 hand-annotated abstracts [27])
BiLSTM-CNNs-CRF [27]	0.87
MatBERT-BiLSTM-CRF (Ours)	0.91

Table 3
The overall accuracy metrics for our model on the initial and new test set.

Label	F1 (Initial test set)	F1 (New test set)
Application	0.58	**0.68**
Characterization	0.81	**0.82**
Composition	0.94	0.93
Condition	0.80	**0.83**
Feature	0.80	0.78
Processing	0.82	0.80
Property	0.82	0.82
Structure	0.84	**0.85**
Total	0.87	0.86

used as adverbial), which confuses our model during classification.

4. Application example

In this section, ADR is employed to recognize and screen descriptors related to the activation energy. Specifically, CGDR extracts 106896 coarse-grained descriptors and their corresponding sentences from 1808 literature. Here, the prediction of activation energy is taken as an example for obtaining associated descriptors, and 408 high-quality performance-driven descriptors are screened from the knowledge base with FGDR (screening processes are outlined in detail in Supporting Information S.4), which are shown in Fig. 10(a). Further, the prediction of activation energy is investigated with the screened descriptors. The detailed process is as follows:

Firstly, two sample datasets of activation energy prediction are built by different experts. Based on simple molecular, structural parameters, and electronic, 31 descriptors (details are shown in Supporting Information S.5) are selected from the previous screened results by one expert, and then a sample dataset ($Dataset_{31}$) is constructed. Moreover, 45 descriptors, containing parameters of property, composition, structure, and external condition (details are shown in Supporting Information S.6), are selected by another, and the corresponding sample dataset ($Dataset_{45}$) is constructed. Fig. 10(b) shows part of the descriptors for activation energy prediction in this study. As shown, there are also some candidate descriptors in screened descriptors that may be relevant to the prediction of activation energy but have not been studied yet. For example, "bond strength" may be useful, since the ion transport capacity may be related to the bonding strength between M–O and X–O. For predicting activation energy, 85 CIFs describing NASICON materials are collected from the Inorganic Crystal Structure Database (ICSD). In the process of building these two sample datasets, some are directly read from CIFs, while others are calculated. Detailed information on samples can be found in Supporting Information S.7.

Secondly, six ML models (LASSO, GPR, Ridge, SVR, KNN, and RF) are employed to predict activation energy. Meanwhile, 10-fold cross-validation is utilized to split the two datasets containing 85 NASICON samples. To evaluate these models, Root Mean Square Error (*RMSE*), Mean Absolute Percent Error (*MAPE*), and R-square (R^2) are introduced into the experiments, of which details can be found in Supporting Information S.8.

In the end, the results of the six candidate ML models are shown in Table 4. As shown, GPR and RF, trained on $Dataset_{45}$, gain better

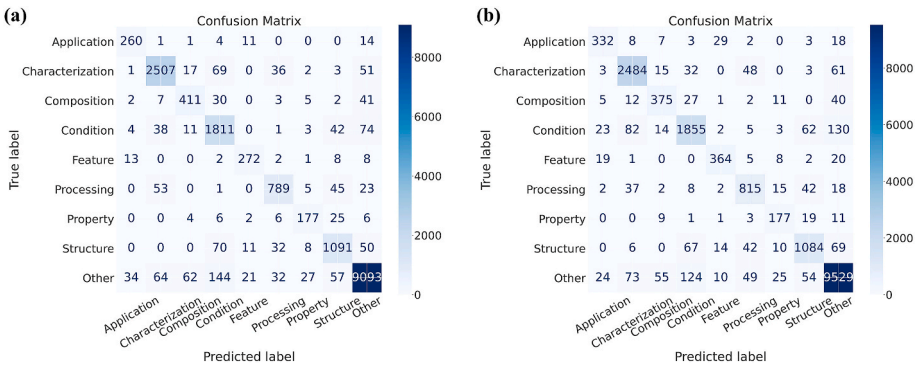

Fig. 9. (a) The confusion matrix on the initial test set. (b) The confusion matrix on the new test set.

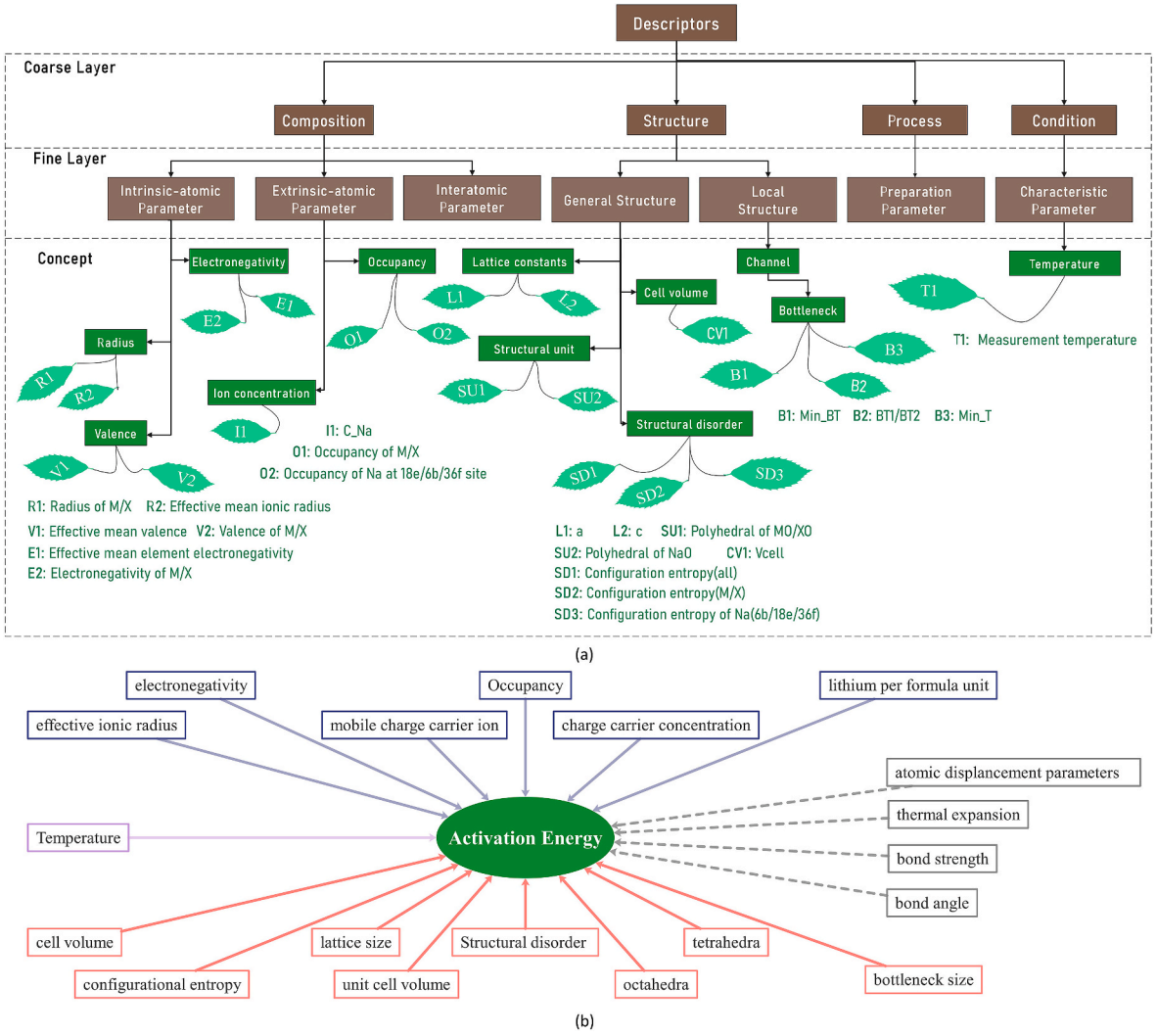

Fig. 10. (a) The visualization of partial candidate descriptors screened by FGDR. (b) Partial descriptors are selected based on (a) for activation energy prediction in this paper, of which dotted lines indicate potential ones still to be developed.

performances beyond LASSO, GPR, and SVR trained on $Dataset_{31}$, of which *RMSE*, *MAPE*, and R^2 are improved by 0.06, 0.03, and 0.16, respectively. This is because the $Dataset_{45}$ has more comprehensive descriptors than $Dataset_{31}$, such as "configuration entropy", "bottleneck" and so on. Moreover, the R^2 of activity energy prediction results on both datasets are generally more than 80% with all ML models, which is because our screened result is based on high-quality descriptors. That is the objective and effective descriptors ADR provides.

5. Conclusion

Structure-activity relationships are of great importance for optimizing properties of materials and discovering novel materials, in which descriptors as the key features are often supplied by domain experts. Materials science literature contains domain knowledge about

Table 4
The comparison of experimental results on $Dataset_{31}$ and $Dataset_{45}$ for different ML models.

Model	$Dataset_{31}$			$Dataset_{45}$		
	RMSE	*MAPE*	R^2	*RMSE*	*MAPE*	R^2
LASSO	0.09	0.06	0.86	0.06	0.04	**0.94**
GPR	0.09	0.06	0.86	0.05	0.04	**0.96**
Ridge	0.09	0.06	0.86	0.05	0.04	**0.95**
SVR	0.10	0.07	0.84	0.07	0.06	**0.92**
KNN	0.11	0.07	0.80	0.09	0.06	**0.87**
RF	0.10	0.06	0.83	0.05	0.04	**0.96**

numerous descriptors, which are laborious and time-consuming for obtaining. Therefore, we develop an automatic descriptors recognizer (ADR), which is composed of a conditional data augmentation model incorporating materials domain knowledge (cDA-DK), coarse- and fine-grained subrecognizers (CGDR and FGDR, respectively), to automatically select descriptors used for predicting materials property from materials literature. Firstly, based on small size hand-annotated dataset, more high-quality domain data can be generated with the proposed cDA-DK model, which can train the NER model to be more generalizable. Then, the CGDR based on a NER model can automatically recognize descriptor entities of diverse types from materials literature. A training and test run on 55 relevant articles yielded an F1 score of 87%. The experimental results show the NER model can fully capture the contextual semantic features of materials text to classify words or phrases, and then be used for descriptors automatic recognition. Finally, the FGDR can screen high-quality performance-driven descriptors from the results of the previous step.

To validate the proposed method, activation energy is used as the target property to screen descriptors. 106896 descriptor entities are recognized from 1808 NASICON literature sources by CGDR, and then 408 descriptors screened through FGDR. With these screened descriptors, two datasets are constructed as the input to ML models for activation energy prediction. These models achieve good predicted result, which demonstrates the effectiveness of the ADR, which can be applied to any other materials properties and extended to extract structure-property relations.

CRediT authorship contribution statement

Yue Liu: Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Supervision. **Xianyuan Ge:** Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing. **Zhengwei Yang:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Shiyu Sun:** Validation, Writing – review & editing. **Dahui Liu:** Formal analysis, Resources, Data curation, Writing – original draft. **Maxim Avdeev:** Writing – original draft. **Siqi Shi:** Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing, Supervision.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2021YFB3802100), the National Natural Science Foundation of China (Grant No. 52073169), and the State Key Program of National Natural Science Foundation of China (Grant No. 61936001). 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. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpowsour.2022.231946.

References

[1] B.J. Shields, J. Stevens, J. Li, M. Parasram, A.G. Doyle, Bayesian reaction optimization as a tool for chemical synthesis, Nature 590 (7844) (2021) 89–96.
[2] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, JJ. Phys. Commun. Materiomics 3 (3) (2017) 159–177.
[3] Y. Liu, A. Bg, A. Xz, E. Yld, E. Ssd, Machine learning assisted materials design and discovery for rechargeable batteries, Energy Storage Mater. 31 (2020) 434–450.
[4] A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data, IEEE Intell. Syst. 24 (2) (2009) 8–12.
[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35.
[6] R. Jalem, M. Nakayama, T. Kasuga, An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks, J. Mater. Chem. 2 (2013) 720–734.
[7] R. Jalem, K. Kanamori, I. Takeuchi, M. Nakayama, H. Yamasaki, T. Saito, Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application, Sci. Rep. 8 (1) (2018) 1–10.
[8] 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 (1) (2016) 306–320.
[9] Y. Xu, Y. Zong, K. Hippalgaonkar, Machine learning-assisted cross-domain prediction of ionic conductivity in sodium and lithium-based superionic conductors using facile descriptors, Journal of Physics Communications 4 (5) (2020), 055015.
[10] 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 (14) (2021).
[11] S. Zhu, C. He, N. Zhao, J. Sha, Data-driven analysis on thermal effects and temperature changes of lithium-ion battery, J. Power Sources 482 (2021), 228983.
[12] N.H. Paulson, J. Kubal, L. Ward, S. Saxena, W. Lu, S.J. Babinec, Feature engineering for machine learning enabled early prediction of battery lifetime, J. Power Sources 527 (2022), 231127.
[13] Y. Liu, X. Zou, Z. Yang, S. Shi, Machine learning embedded with materials domain knowledge, J. Chin. Ceram. Soc. 50 (3) (2022) 863–876.
[14] S. Shi, Z. Tu, X. Zou, S. Sun, Z. Yang, Y. Liu, Applying data-driven machine learning to studying electrochemical energy storage materials, Energy Storage Sci. Technol. 11 (3) (2022) 739–759.
[15] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigat. 30 (1) (2007) 3–26.
[16] E. Kim, K. Huang, S. Jegelka, E. Olivetti, Virtual screening of inorganic materials synthesis parameters with deep learning, npj Comput. Mater. 3 (1) (2017) 1–9.
[17] 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 (21) (2017) 9436–9444.
[18] E. Kim, et al., Machine-learned and codified synthesis parameters of oxide materials, Sci. Data 4 (1) (2017) 1–9.
[19] S. Mysore, et al., Automatically extracting action graphs from materials science synthesis procedures, arXiv preprint arXiv:1711.06872 (2017).
[20] T. Rocktäschel, M. Weidlich, U. Leser, ChemSpot: a hybrid system for chemical named entity recognition, Bioinformatics 28 (12) (2012) 1633–1640.
[21] R. Leaman, C.-H. Wei, Z. Lu, tmChem: a high performance approach for chemical named entity recognition and normalization, J. Cheminf. 7 (1) (2015) 1–10.
[22] M.C. Swain, J.M. Cole, ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature, J. Chem. Inf. Model. 56 (10) (2016) 1894–1904.
[23] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal, A. Valencia, Information retrieval and text mining technologies for chemistry, Chem. Rev. 117 (12) (2017) 7673–7761.
[24] X. Zhao, S. Lopez, S. Saikin, X. Hu, J. Greenberg, Text to insight: accelerating organic materials knowledge extraction via deep learning, Proc.Assoc.Info. Sci. Technol. 58 (1) (2021) 558–562.
[25] V. Tshitoyan, et al., Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (7763) (2019) 95–98.
[26] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
[27] L. Weston, et al., Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, J. Chem. Inf. Model. 59 (9) (2019) 3692–3702.

[28] T. He, et al., Similarity of precursors in solid-state synthesis as text-mined from scientific literature, Chem. Mater. 32 (18) (2020) 7861–7873.
[29] S.M. Yimam, A.A. Ayele, G. Venkatesh, I. Gashaw, C. Biemann, Introducing various semantic models for Amharic: experimentation and evaluation with multiple tasks and datasets, Future Internet 13 (11) (2021) 275–293.
[30] I. Segura-Bedmar, P. Martínez Fernández, M. Herrero Zazo, Semeval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (Ddiextraction 2013), Association for Computational Linguistics, 2013.
[31] S. Eltyeb, N. Salim, Chemical named entities recognition: a review on approaches and applications, J. Cheminf. 6 (1) (2014) 1–12.
[32] Z. Nie, et al., Automating materials exploration with a semantic knowledge graph for Li-ion battery cathodes, Adv. Funct. Mater. (2022), 2201437.
[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional Transformers for language understanding," minneapolis, Minnesota, jun 2019: association for computational linguistics, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume vol. 1 (Long and Short Papers), pp. 4171-4186, doi: 10.18653/v1/N19-1423.
[34] J. Lee, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[35] J.T. Shen, et al., Mathbert: a pre-trained language model for general nlp tasks in mathematics education, arXiv preprint arXiv:2106.07340 (2021).
[36] T. Gupta, M. Zaki, N. Krishnan, MatSciBERT: a materials domain language model for text mining and information extraction, npj Comput. Mater. 8 (1) (2022) 1–11.
[37] X. Jiao, et al., TinyBERT: distilling BERT for natural language understanding, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov 2020: Association for Computational Linguistics, in Findings of the Association for Computational Linguistics: EMNLP, 2020, pp. 4163–4174, https://doi.org/10.18653/v1/2020.findings-emnlp.372.
[38] J. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," Hong Kong, China, nov 2019: association for computational linguistics, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6382-6388, doi: 10.18653/v1/D19-1670.
[39] X. Wu, S. Lv, L. Zang, J. Han, S. Hu, Conditional bert contextual augmentation, in: International Conference on Computational Science, Springer, 2019, pp. 84–95.
[40] X. Dai, H. Adel, An analysis of simple data augmentation for named entity recognition, arXiv preprint arXiv:2010.11683 (2020).
[41] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP," online, oct 2020: association for computational linguistics, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 119-126, doi: 10.18653/v1/2020.emnlp-demos.16.
[42] Y. Liu, et al., Roberta: a robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[43] H. Yan, B. Deng, X. Li, X. Qiu, TENER: adapting transformer encoder for named entity recognition, arXiv preprint arXiv:1911.04474 (2019).
[44] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[45] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).