Journal of
Materials Chemistry A

PAPER

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d3ta02780e

Deep learning of electrochemical CO_2 conversion literature reveals research trends and directions†

Jiwoo Choi, ‡[ab] Kihoon Bang, ‡[a] Suji Jang,[a] Jaewoong Choi,[a] Juanita Ordonez,[c] David Buttler,[c] Anna Hiszpanski,[c] T. Yong-Jin Han,[c] Seok Su Sohn,[b] Byungju Lee,[a] Kwang-Ryeol Lee,[a] Sang Soo Han*[a] and Donghun Kim *[a]

Received 11th May 2023
Accepted 18th July 2023

DOI: 10.1039/d3ta02780e

rsc.li/materials-a

Large-scale and openly available material science databases are mainly composed of computer simulation results rather than experimental data. Some examples include the Materials Project, Open Quantum Materials Database, and Open Catalyst 2022. Unfortunately, building large-scale experimental databases remains challenging due to the difficulties in consolidating locally distributed datasets. In this work, focusing on the catalysis literature of CO_2 reduction reactions (CO_2RRs), we present a machine learning (ML)-based protocol for selecting highly relevant papers and extracting important experimental data. First, we report a document embedding method (Doc2Vec) for collecting papers of greatest relevance to the specific target domain, which yielded 3154 CO_2RR-related papers from six publishers. Next, we developed named entity recognition (NER) models to extract twelve entities related to material names (catalyst, electrolyte, *etc.*) and catalytic performance (Faradaic efficiency, current density, *etc.*). Among several tested models, the MatBERT-based approach achieved the highest accuracy, with an average F1-score of 90.4% and an F1-score of 95.2% in a boundary relaxation evaluation scheme. The accurate and accelerated NER-based data extraction from a large volume of catalysis literature enables temporal trend analyses of the CO_2RR catalysts, products, and performances, revealing the potentially effective material space in CO_2RRs. While this work demonstrates the effectiveness of our ML-based text mining methods for specifically CO_2RR literature, the methods and approach are applicable to and may be used to accelerate the development of other catalytic chemical reactions.

Introduction

A large amount of material science data has been accumulated worldwide.[1] As a result, machine learning has recently emerged as a powerful tool to solve materials science problems.[2] However, the large-scale and openly available materials science databases primarily include results from computer simulations rather than experimental data. Some exemplary databases include the Materials Project,[3] Open Quantum Materials Database (OQMD),[4] Novel Materials Discovery (NOMAD),[5] and Open Catalyst 2022 (OC22).[3,6] Typically, experiments are performed locally in each laboratory, and these widely distributed results are extremely difficult to collect in a single database. In this regard, the scientific literature can be a great source of focus since the literature contains a large amount of high-quality experimental data as a result of the peer review processes. Extracting key information from the literature is considered an excellent approach to construct large-scale experimental material databases and requires diverse natural language processing techniques.

Manually text mining literature is very time-consuming, and thus, automated processes for text mining, such as rule-based or machine learning-based approaches, are highly desirable. In particular, named entity recognition (NER)[7] is the most frequently used method for recognizing and extracting entities such as words or phrases in papers by classifying them according to predefined labels. The extracted entities from each paper could be stored as specific data, such as material's name and performance value. In contrasts, when using conventional analysis tools from literature databases including Web of Science[8]or Scopus,[9] we can only analyze and extract the predefined keywords or topics, resulting in the loss of detailed information within the literature. To extract the detailed information from literature, NER plays a critical role in automated text mining. NER has recently been used in materials science[10–15] and biomedical research[16–18] to extract the properties of materials or chemicals from texts. For example, Weston *et al.*[12] successfully

[a] Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea. E-mail: donghun@kist.re.kr; sangsoo@kist.re.kr

[b] Department of Materials Science and Engineering Korea University, Seoul 02841, Republic of Korea

[c] Materials Science Division, Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3ta02780e

‡ These authors contributed equally.

extracted general materials science entities, such as material names, phase names, and synthesis or characterization methods, by applying a long short-term memory (LSTM)-based NER model[19] to abstracts of 3.3 million materials science papers. Similarly, Park *et al.*[11] explored the literature on metal–organic frameworks (MOFs) and extracted entities such as MOF names and synthetic routes and conditions from 30 k papers. More recently, transformer-based models, such as bidirectional encoder representations from transformers (BERT), have been actively used in natural language processing (NLP) tasks due to their enhanced accuracies.[20–22] Notably, Trewartha *et al.*[21] developed MatBERT and compared the entity extraction performance of several BERT models, including BERT (ref. 20) and SciBERT,[23] and found that the models' performances are highly affected by the training domain.

Despite previous efforts, performing text mining for a specific target domain[24] like catalysis remains difficult. One difficulty is selectively collecting papers of greatest relevance to the specific target domain. Previous reports considered general materials science papers, and in those cases, collecting papers is straightforward since selection processes are not needed. However, if a user wants to perform text mining on, for example, the domain of electrochemical CO_2 reduction reactions (CO_2RRs), an automated and efficient selection process is necessary. Another difficulty is the development and application of NER models. Material names (catalysts, products) are multi-word entities involving sometimes even more than ten words, which is much larger than other types of labels such as performance values. For instance, in an article by Wang *et al.*,[25] a complex catalyst was used consisting of Ni nanoparticles (NPs) encapsulated in nitrogen-doped carbon nanotubes (CNTs) and assembled on the surface of graphene; this catalyst is abbreviated in its original text as "N/NiNPs@CNT/G". In this case, the evaluation of NER is challenging since multi-words entities are not only more difficult to identify but also frequently lack consensus on the entity's exact beginning and end. Moreover, entity annotations need to be performed very carefully to achieve high NER performance.[24] The catalyst names appearing in CO_2RR papers are not standardized and often include supports or structures in the catalyst name, as exemplified above. Thus, defining clear annotation rules, like which words should be annotated as catalyst materials, is important. For instance, in an article by Zhu *et al.*,[26] the authors compare the performance of copper catalysts having varied structures. Usually in NER tasks, catalysts are identified by solely their chemical composition (*i.e.*, "copper"), but in this case, annotating both the chemical and structural words as the catalyst name is critical (*i.e.*, "copper hollow fiber", "copper foam", "copper foil"). Since entity characteristics differ in various target application domains, pretrained NER models from prior studies are often not applicable out-of-the-box to a new domain. These problems call for the development of NER models that are specific to different target domains, which in our case is electrochemical CO_2RR.

In this work, machine learning (ML)-based text mining methods are presented and demonstrated to be effective for the electrochemical CO_2RR target domain. First, we report a method for selectively collecting papers pertinent to this specific domain and excluding irrelevant papers using a combination of Doc2Vec (ref. 27) and latent Dirichlet allocation (LDA)[28] algorithms. Next, NER models based on bidirectional LSTM (Bi-LSTM)[19] and BERT (ref. 20) algorithms are retrained using 500 manually annotated CO_2RR papers to extract key entities related to catalysis, such as material names and catalytic performance data. The MatBERT-based model shows the best performance, exhibiting an F1-score of 90.4% that reached up to 95.2% in a boundary relaxation[29] evaluation scheme owing to the accuracy enhancements[30,31] for multiword entities, such as catalyst and electrolyte names. The catalysis-specific NER model enables accelerated data extraction from a large volume of literature (in our case, 3154 CO_2RR papers), thereby allowing yearly trend analyses of the CO_2RR catalysts, products, and performance. This analysis reveals the recent element usage trends, including the boost of post-transition metal elements of Bi and In for formic acid production. Additionally, it also proposes the potentially effective material space in CO_2RRs, such as transition metal elements of Cr, Mn and Mo for single atom catalysts. To the best of our knowledge, this work is the first study reporting ML-based text mining for catalysis literature, and the proposed approach will be useful for other catalytic reactions in addition to CO_2RRs.

Results and discussion

Workflow of CO_2RR literature processing

Fig. 1 presents the four sequential steps in the CO_2RR literature processing workflow developed in this paper, including (1) paper collection and preprocessing, (2) exclusion of irrelevant papers, (3) data extraction using named entity recognition (NER), and (4) analysis of research trends and directions. In step (1), a total of 4838 CO_2RR-related papers were collected with the permissions based on keyword queries from the websites of six publishers. These papers were parsed into plain text and tokenized using ChemDataExtractor[32] for the subsequent text mining processes. Since the keyword query-based search in step (1) still yielded some "noise papers" irrelevant to CO_2RR, in step (2), these noise papers were filtered out using our combined approach of Doc2Vec and latent Dirichlet allocation (LDA), finally leaving the most relevant 3153 papers. In step (3), entity labeling and NER were performed. Twelve types of entities, including material names and electrochemical properties, were manually labeled. Examples of these twelve types of entities are provided in Table S1.† Using the labeled dataset, different NER models, including Bi-LSTM and BERT-based models, were developed. Finally, in step (4), the best NER model (MatBERT-based[21] approach) was applied to all 3153 papers to generate a large database. This database was analyzed to assess useful and recent research trends and directions in the CO_2RR field. More detailed results from each step are explained below.

Preparation of electrochemical CO_2RR papers

Our first approach to acquire the CO_2RR papers involves a keyword query-based search to collect papers from journal

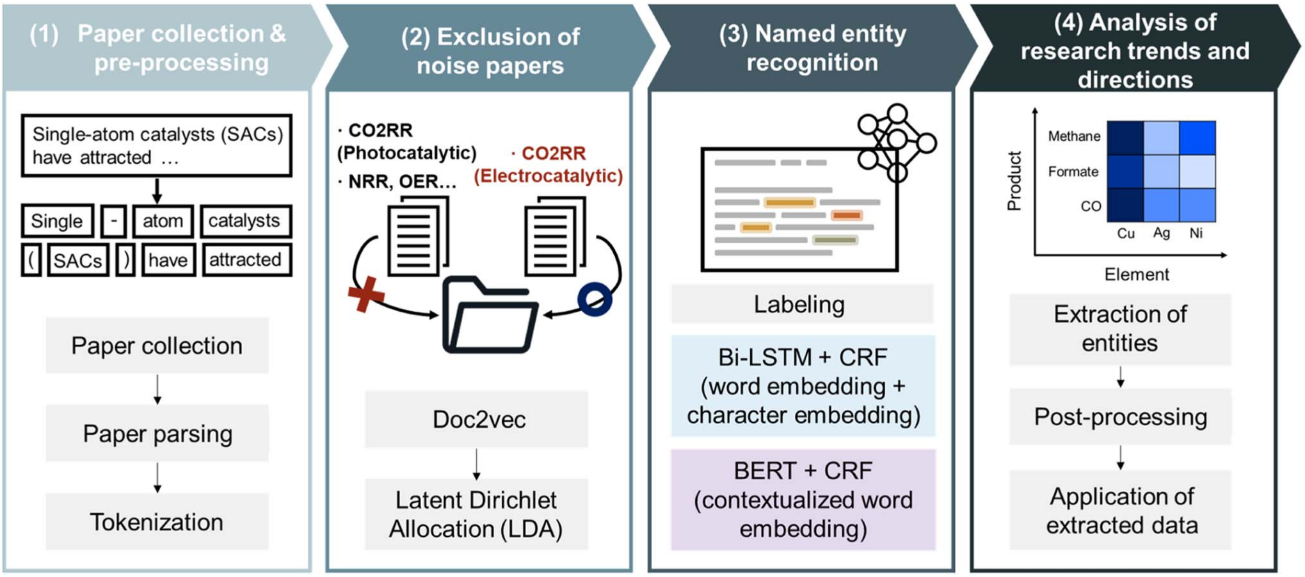

(1) Paper collection & pre-processing
Single-atom catalysts (SACs) have attracted …
Single
-
atom
catalysts
(
SACs
)
have
attracted
Paper collection
Paper parsing
Tokenization
(2) Exclusion of noise papers
· CO2RR (Photocatalytic)
· NRR, OER…
· CO2RR (Electrocatalytic)
Doc2vec
Latent Dirichlet Allocation (LDA)
(3) Named entity recognition
Labeling
Bi-LSTM + CRF (word embedding + character embedding)
BERT + CRF (contextualized word embedding)
(4) Analysis of research trends and directions
Methane
Formate
CO
Product
Cu
Ag
Ni
Element
Extraction of entities
Post-processing
Application of extracted data

Fig. 1 Workflow of CO_2RR literature processing. The following steps are crucial in the workflow: (1) papers are collected and then parsed in text format, and the text is tokenized for the training dataset. (2) Doc2Vec and LDA algorithms are applied to remove irrelevant papers. (3) Twelve types of entities are manually labeled, and the labeled training dataset is used to train two types of NER models. (4) The material names and catalyst performance are extracted based on the collected papers, and research trends and directions are analyzed.

websites (Fig. 2a). Journals were considered from the following six publishers: Springer Nature, Elsevier, Wiley, American Association for the Advancement of Science (AAAS), American Chemical Society (ACS), and Royal Society of Chemistry (RSC). Only research articles were collected; any papers categorized as reviews, perspectives, and news were excluded. To select papers relevant to CO_2 electroreduction catalysis, we considered many versions of keyword queries in the Boolean search and found the best search to be "CO2" AND "reduction" AND "reaction" AND "catalyst" AND "faradaic" AND ("electrochemical" OR "electroreduction" OR "electrocatalytic" OR "electro"). This search process resulted in a total of 4838 papers.

Unfortunately, the keyword query-based search does not provide fully satisfactory screening results. A substantial number of irrelevant or "noise papers" were found, examples of which include photochemical CO_2RR papers, oxygen reduction reaction papers, and nitrogen reduction reaction papers. To filter out these irrelevant papers, we report a combined approach involving Doc2Vec and LDA methods. Doc2Vec is an algorithm that converts a document into a vector (the document version of Word2Vec[33,34]). We manually selected 20 seed papers that best represent our target topic of electrochemical CO_2RR. Then, 200-dimensional document embedding using Doc2Vec are used a cosine similarity calculation and we compute the difference between our seed papers and the remaining papers in our collection. Fig. 2b visualizes the document embedding results for all 4838 papers including the 20 seed papers which were reduced to two dimensions using principal component analysis (PCA) algorithms, and the maximum cosine similarity between an arbitrary paper and 20 seed papers was used as a screening metric. We chose to exclude the papers that differed the most from the seed papers.

To determine whether the Doc2Vec strategy was effective for removing irrelevant papers, topic modeling through latent Dirichlet allocation (LDA) was applied. LDA assigns the probabilities of several topics for each paper based on the words describing the topic. An appropriate number of topics must be selected, which is a hyperparameter of LDA. Based on the complexity[28] and coherence[35] of the LDA approach, the optimal number of topics was determined to be 11. Table S2† provides the detailed contents of the LDA topics. Several topics were identified as completely irrelevant to CO_2RR and thus grouped as noise. We further classified the eleven topics into four categories, which were entitled: CO_2RR, density functional theory (DFT), analysis, and noise. In general, catalyst papers often contain a characterization part that analyzes the structure and properties of a catalyst and a part that identifies the mechanism of the catalytic reaction using DFT simulations. Thus, we reflected these attributes in the final four categories. Fig. 2c shows the ratio of noise papers as a function of the Doc2Vec similarity. For 500 papers with the highest Doc2Vec similarity (>0.55), only 16% were identified as noise papers. On the other hand, for 500 papers with low Doc2Vec similarity (<0.35), approximately 70% were identified as noise papers. This comparison reveals the effectiveness of the Doc2Vec approach in filtering out irrelevant papers. The threshold for the Doc2Vec similarity was set as 0.4 in this work. The resulting process removed an additional 1683 papers, finally leaving 3153 papers in our corpus. The combined approach of Doc2Vec and LDA is useful for screening papers that are outside of the research target topic. This method requires the input of only tens of seed papers from a user, and thus can universally be applied to other research domains beyond CO_2RR with ease.

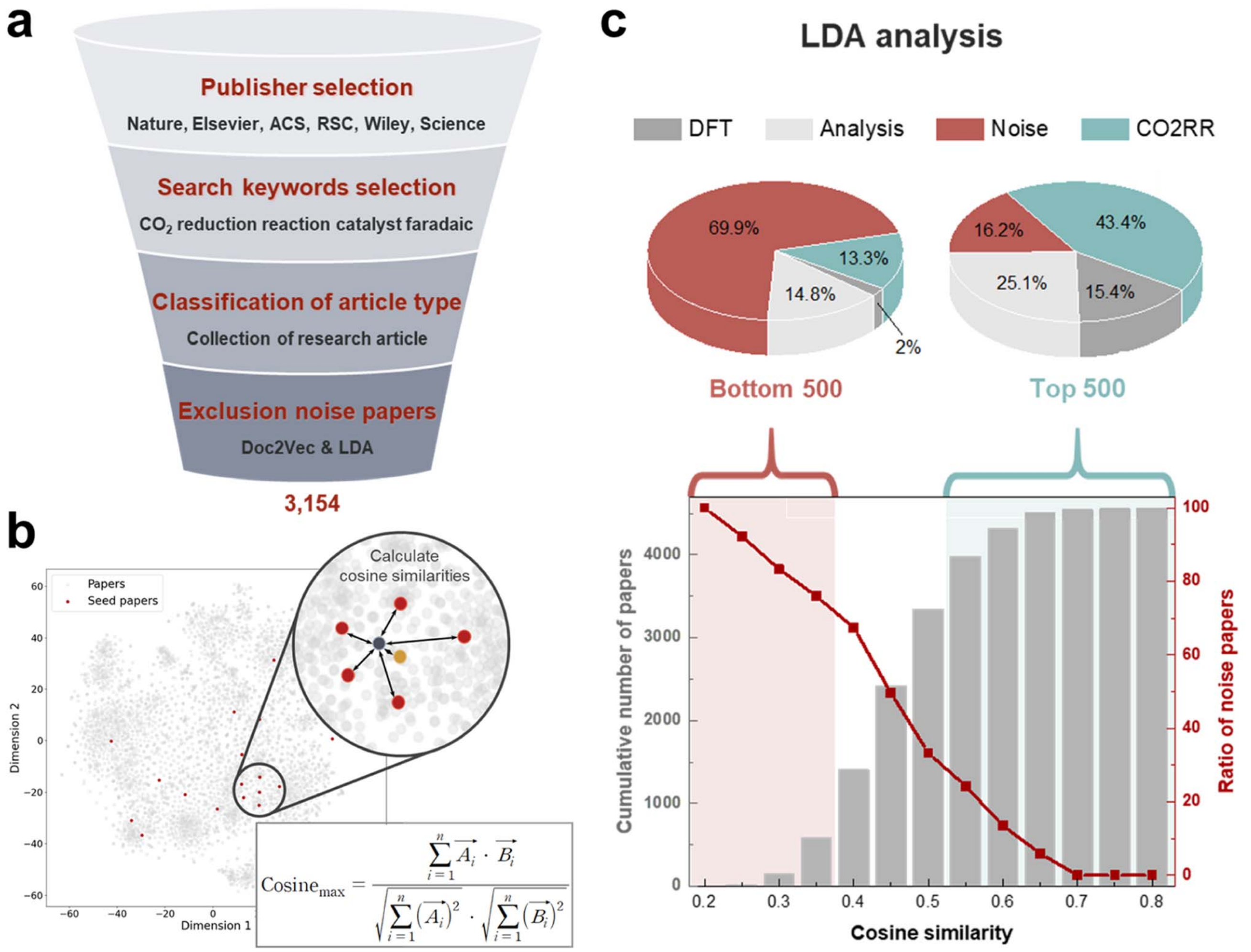


Fig. 2 Preparation of electrochemical CO2RR papers. (a) Scheme to collect only experimental CO_2RR research articles, which is composed of four steps (red text). (b) Paper embeddings made by Doc2Vec and visualized with principal component analysis (PCA). The similarity calculation between random papers and seed papers using the cosine similarity method is highlighted. (c) The cumulative number of papers and ratio of noise papers according to the cosine similarity are shown as bar and line graphs, respectively. A cosine similarity of 0.4 is chosen as a criterion for filtering out noise papers. The 11 topics obtained as a result of LDA were classified into 4 categories (DFT, analysis, noise, and CO_2RR). The pie charts show the results of applying the LDA analysis to the top 500 and bottom 500 papers in the corpus.

Data extraction using NER

We applied the NER approach to only the abstracts of papers rather than the full document text because we assume that the abstract contains the most important information in the paper and the limited scope of text simplifies the NER task (Table S3 in the ESI†). As shown in Fig. S1,† analyses of 100 randomly selected papers confirmed that the information about catalyst names and product names in the main bodies was also present as much as 98% of the abstracts. Also, we found that the information regarding performance (faradaic efficiency, current density, onset potential, overpotential, stability hour, and turnover frequency) was present more than 76% of the abstracts. These analyses support that the abstracts contain the most important information in papers, and justify the approach of prioritizing abstracts in the current study. For the NER task, we selected the following twelve entities that contain key information for CO_2RR: catalyst, product, electrolyte, reference electrode, current density, faradaic efficiency, stability hour, turnover frequency, overpotential, onset potential, potential, and concentration. Note that the first four entities contain material information, whereas the remaining eight entities are related to the electrochemical catalytic performance. Entities of material names are composed of chemical elements, while entities of electrochemical measurements are mainly composed of numbers and units. Several examples of each entity are provided in Table S4.† These entities were manually annotated based on only the abstracts of the top 500 papers with high Doc2Vec similarity (Fig. 2c). The inside-outside-beginning-end-single (IOBES) format, which is known to be effective in addressing multiword entities,[36,37] was adopted for the annotation process. Several annotation formats, including IOB, IOBE, and IOBES, were considered, and IOBES generally produced the best NER performance (Table S4 in the ESI†).

Two types of NER models were developed, Bi-LSTM and BERT-based models, and their network architectures are shown in Fig. 3. The LSTM model was designed to solve the long-term dependency problem of recurrent neural networks (RNNs), which show poor performance for long sequences.[38] Word embedding and character-level representations[39] were both applied in the Bi-LSTM models. Word embedding was performed by training the Word2Vec algorithm[34] based on our own corpus of catalysis papers. The character-level representations were obtained from the output of the character-level Bi-LSTM training based on the same corpus. The character-level representations reflect words such as plurals or molecular formulae. These two embedding vectors were concatenated and used as inputs to the Bi-LSTM model.

On the other hand, BERT is a transformer-based model[20] that is known to understand the entire context considerably better than Bi-LSTM models due to self-attention.[40] Self-attention mechanisms obtain contextual information by calculating the correlations between all words in the input text. It is very important to note that in contrast to the embeddings used in Bi-LSTM models, contextualized word embeddings are used in the BERT model, which means that the same words can be embedded differently depending on the context. Several general and domain-specific BERT models have been reported, including BERT_base,[20] SciBERT,[23] MatSciBERT,[41] MatBERT,[21] BioBERT,[42] and FinBERT,[43] although a catalysis-specific version has not yet been devised. Since none of these previous models can be directly implemented to address catalysis papers, we considered four pretrained BERT models (BERT_base, SciBERT, MatSciBERT, and MatBERT) and fine-tuned them using our

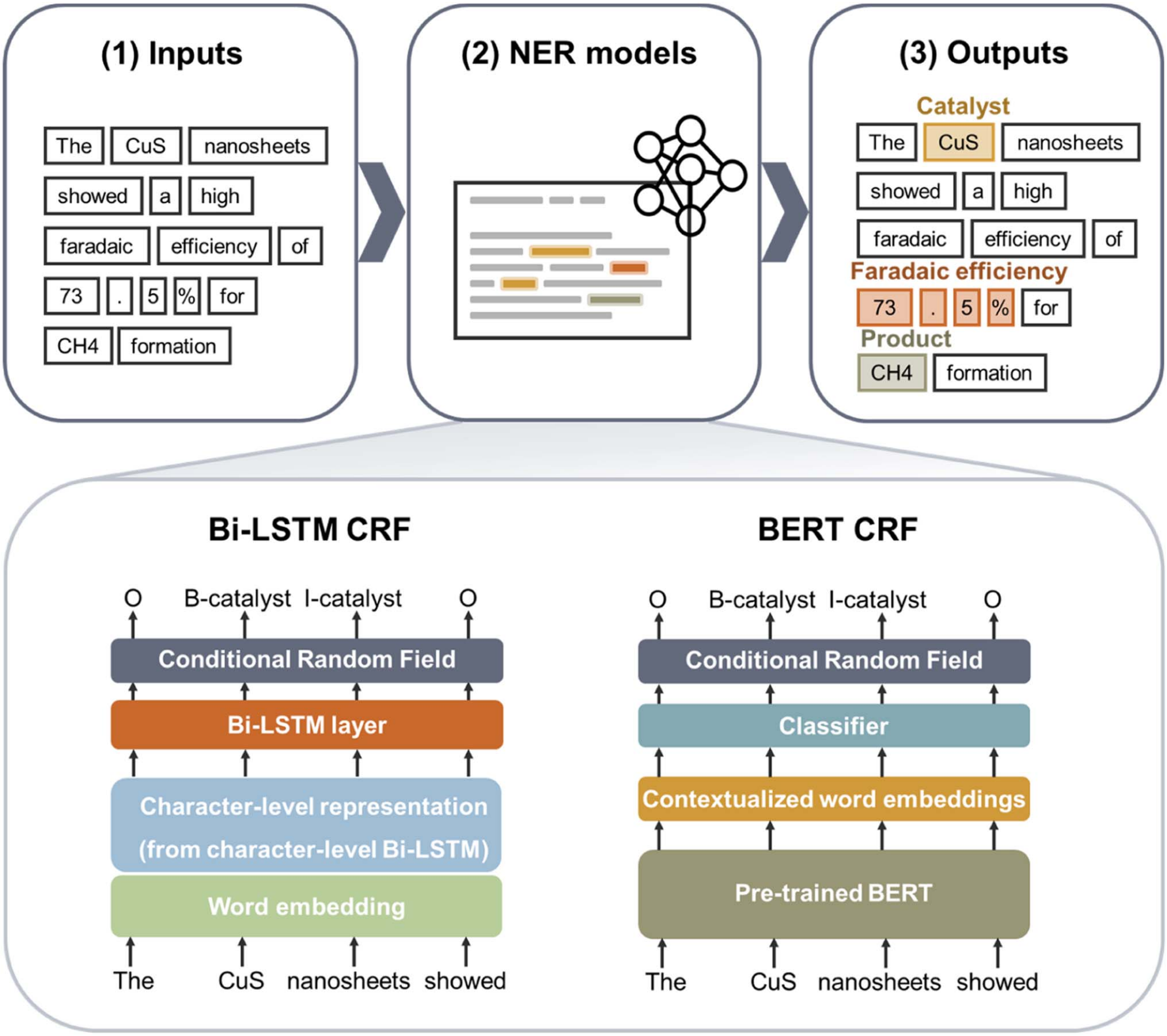


Fig. 3 Structures of the NER models. NER consists of three steps as follows. (1) The text is tokenized, and the token sequences are used as inputs to the model. (2) Two types of NER models were used. In the Bi-LSTM CRF model, tokens are converted into word embeddings and character-level representations and passed to the Bi-LSTM and CRF layers. In the BERT CRF model, tokens are embedded with contextualized word embeddings. Both models return sequences of entities in IOBES format. (3) As a result, the words in the papers are recognized as entities such as the catalysts, faradaic efficiency, and products.

own CO_2RR corpus. Note that the Bi-LSTM and BERT models both have a conditional random field (CRF) layer[44] to fix the sequence problems associated with entities in IOBES format. For example, the CRF layer learns the constraints that I and E tags cannot appear in the first word of a sentence and that entity names must be consistent with the B–I–I–E pattern (for example, B-catalyst must not be followed by I-product).

The NER performance of the Bi-LSTM and BERT-based models is presented in Fig. 4. First, for the Bi-LSTM models, two types of input embeddings were investigated, including word embeddings alone and the concatenation of word and character-level embeddings. Overall, the model with the concatenated embedding performs very similarly to (or only slightly better than) the model with the word embeddings alone, which indicates that character-level features are not critical. The F1-scores for the entities of catalyst, electrolyte, and onset potential are less than 80%, which is lower than those of the other entities. The low performance of the electrolyte and onset potential entities may be due to the relatively smaller number of training datasets, as shown in Fig. S2.† On the other hand, the low performance of the catalyst entity is not because of the training dataset size but because catalyst names are typically multiword entities. An example of a catalyst name is "nitrogen-doped carbon sheets",[45] which is composed of four words. It is more difficult to predict multiword entities accurately because mismatches at the left or right ends results in false NER predictions. Alternatively, we provide the results of another evaluation method, namely, boundary relaxation at the left/right ends,[46] as shown in Fig. 4a. In this evaluation scheme, if the model's prediction for an entity is matched at either the left or right end, it is considered a correct prediction. Therefore, by definition, the NER performance is improved in the boundary relaxation evaluation scheme, as confirmed by the F1-score of 92.7% for all entities in Fig. 4a. The performance enhancement is notably large for the catalyst entity because catalyst names are typically multiword entities.

Next, for the BERT models, four types of pretrained BERT models (BERT_base, MatSciBERT, SciBERT, and MatBERT) were fine-tuned based on our catalysis corpus. Note that each BERT model was trained on a different corpus. BERT_base was trained based on a general corpus from Wikipedia (∼2500 M words) and a book corpus (∼800 M). SciBERT was trained based on the scientific literature (1.14 M papers, with 18% from the computer science domain and the remaining 82% from the biomedical domain). MatSciBERT was trained based on the same corpus used in SciBERT training as well as additional materials science papers (∼150 k). Finally, MatBERT was trained based on the materials science literature (∼2 M). The performance of all NER models, as evaluated through 10-fold cross-validation, is presented below. The fine-tuned MatBERT model produced the best NER performance, with an F1-score of 90.4% on average, and the BERT_base model produced the lowest performance (86.8%). Detailed information on the 10-fold cross-validation results of the MatBERT model is provided in Table S5.† This difference can be well understood based on the constitution of the training corpus. Interestingly, the F1-score of MatSciBERT (trained based on some materials science literature) was lower than that of SciBERT (trained based on no materials science literature). This result likely occurred because MatSciBERT was trained based on uncased vocabulary without differentiating between uppercase and lowercase letters, which is not beneficial to performing the NER task with our corpus. Similar to the Bi-LSTM studies, an evaluation based on boundary relaxation at the left/right ends was adopted, and in this scheme, MatBERT produced an overall F1-score of 95.2%. The superior performance of the BERT models over the Bi-LSTM models indicates that the self-attention mechanism in the BERT models is critical to understanding the context of our corpus.

The NER performance was further analyzed based on detailed NER examples, as shown in Fig. 5. Fig. 5a shows two MatBERT prediction examples (examples 1 and 2) where boundary relaxations were important. The text in parentheses is the ground truth, while the text in the background color is the predicted result based on the NER model. In the first example, "BiOI nanoplate precursor-derived" is an adjective that describes "Bi nanosheets".[47] Here, the ground truth for the catalyst entity is "Bi nanosheets"; however, the MatBERT model prediction is "BiOI nanoplate precursor-derived Bi nanosheets", which also includes the adjective terms. Similarly, in the second example, the ground truth for the catalyst entity is "Cu hollow fiber", but the model predicts a longer word set of "Cu hollow fiber with an interconnected pore structure", which also includes its structural description.[48] For these cases involving multiword entities, the evaluation method based on the boundary relaxation at the left/right ends is more reasonable since the predicted results include all the necessary information to fully understand the meaning.

Fig. 5b shows two NER examples (examples 3 and 4) to explain why the BERT models outperform the Bi-LSTM models. In example 3, the term "35 min" represents the electrodeposition time.[49] The Bi-LSTM model incorrectly predicts this term as the entity representing the stability hour, whereas the MatBERT model correctly ignores the term. This difference is likely because MatBERT can capture the context within the text better than the Bi-LSTM models and was capable of differentiating stability hours (one of our 12 entities), and other time entities were not confused. Since minute is a word representing a time, the Bi-LSTM model is confused by this term. In example 4, the term ultrathin NCS represents a support material, not a catalyst material.[50] The Bi-LSTM model incorrectly predicts this as a catalyst entity, whereas MatBERT correctly ignores this term and does not label it as any of the twelve entities. These examples suggest that MatBERT can understand the contextualized meaning of words better than the Bi-LSTM models. Finally, Fig. 5c shows two specific examples (Examples 5 and 6) to explain why MatBERT outperforms the other BERT models. The MatBERT and BERT_base models were trained based on material science literature and general corpora, including Wikipedia and book corpora. In example 5, the term RHE denotes the reversible hydrogen electrode, which is a commonly used term in electrochemistry. MatBERT correctly tokenizes RHE to RHE without further letter decompositions, whereas BERT_base incorrectly tokenizes the word into three tokens of R, ##H, and ##E (## means a subword of

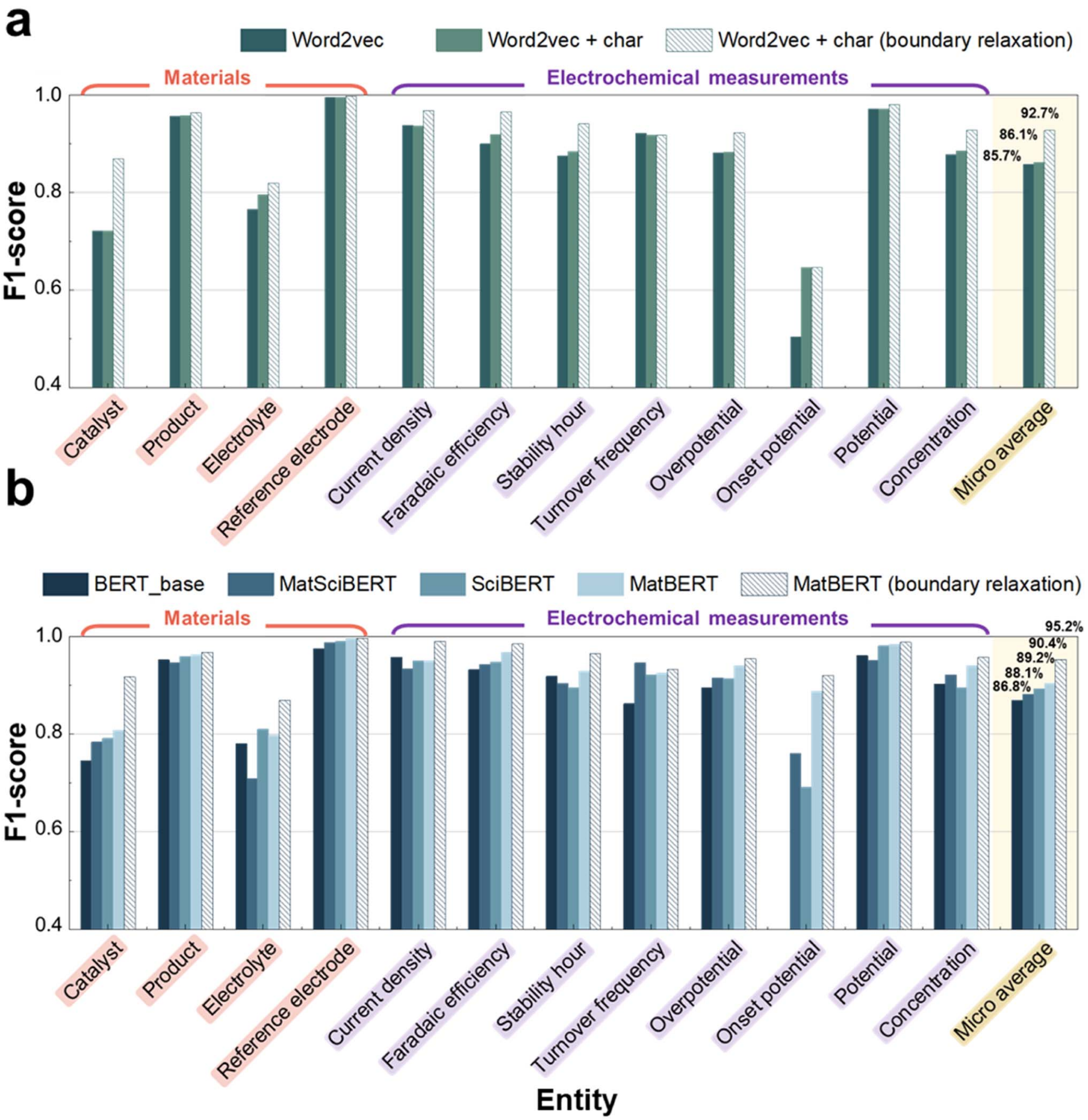
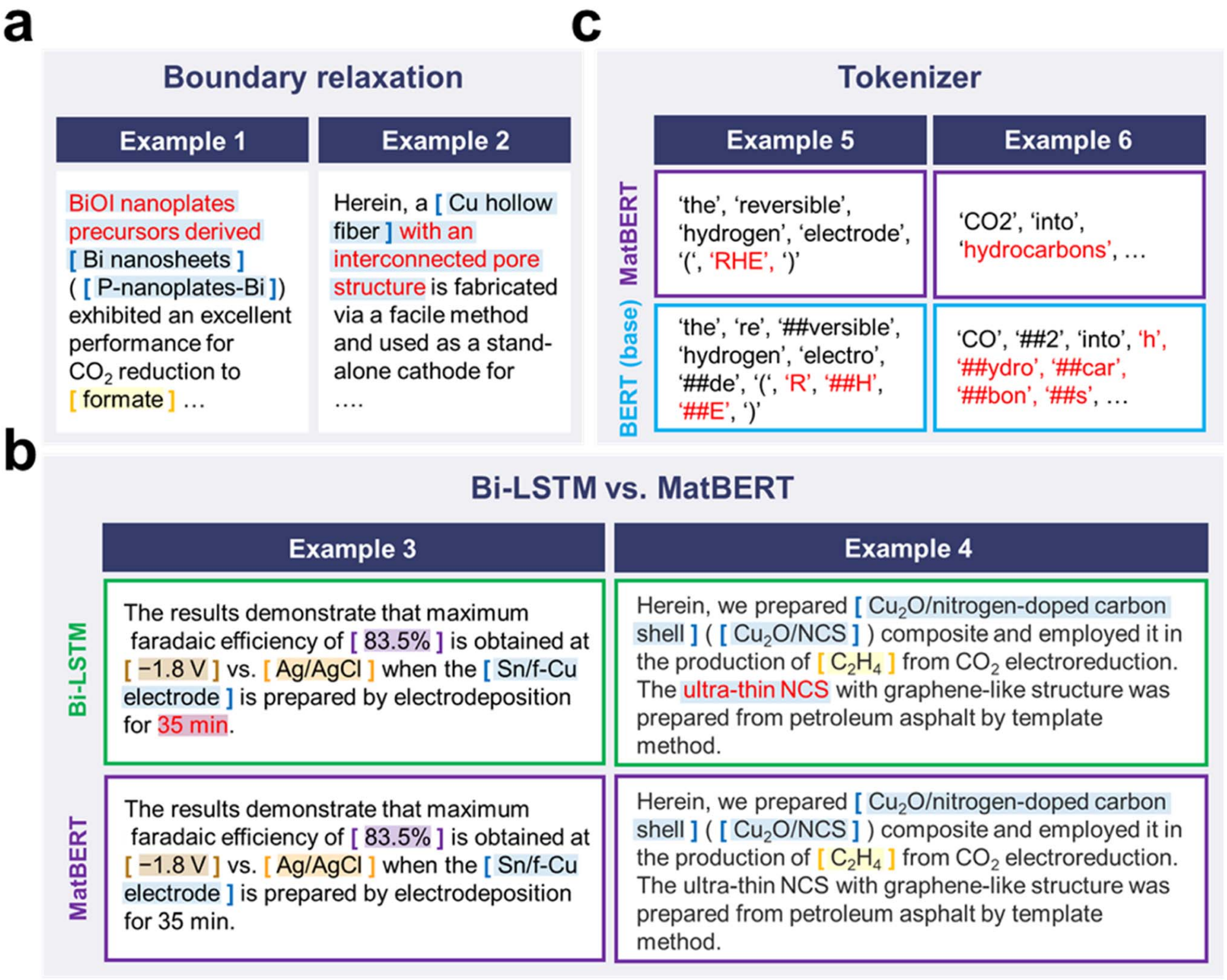


Fig. 4 NER performance of Bi-LSTM and BERT-based models. The graph shows the F1-scores for each entity using the (a) Bi-LSTM CRF model and (b) BERT CRF model. The hatched bars in the graph are the boundary relaxation results. The entities highlighted in pink represent material names, and the entities highlighted in purple represent electrochemical measurements. The micro average marked in yellow is the global average F1-score, which addresses imbalanced entities.

the preceding tokens). This is because the term RHE does not exist in the vocabulary set used for BERT_base training. Similarly, in example 6, a single term of hydrocarbon was tokenized into five tokens of h, ##ydro, ##car, ##bon, ##s by the BERT_base tokenizer. While the tokenization of BERT_base is not incorrect, it makes reconstructing the meaning of these common term harder and more prone to having their meaning diluted by common subwords. These tokenizer results explain why MatBERT, which was trained based on the most relevant literature, achieves the best NER performance.

NER-based analysis of research trends

The trained MatBERT model with the best NER F1-score was used to extract the entities from the abstracts of all 3154 CO_2RR papers. In order to validate the extracted data, we annotated 100 randomly chosen papers that were not used for model training, and then



Example 1: BiOI nanoplates precursors derived [Bi nanosheets] ([P-nanoplates-Bi]) exhibited an excellent performance for CO$_2$ reduction to [formate] …

Example 2: Herein, a [Cu hollow fiber] with an interconnected pore structure is fabricated via a facile method and used as a stand-alone cathode for ….

Example 5, MatBERT: 'the', 'reversible', 'hydrogen', 'electrode', '(', 'RHE', ')'

Example 6, MatBERT: 'CO2', 'into', 'hydrocarbons', …

Example 5, BERT (base): 'the', 're', '##versible', 'hydrogen', 'electro', '##de', '(', 'R', '##H', '##E', ')'

Example 6, BERT (base): 'CO', '##2', 'into', 'h', '##ydro', '##car', '##bon', '##s', …

Example 3, Bi-LSTM: The results demonstrate that maximum faradaic efficiency of [83.5%] is obtained at [−1.8 V] vs. [Ag/AgCl] when the [Sn/f-Cu electrode] is prepared by electrodeposition for 35 min.

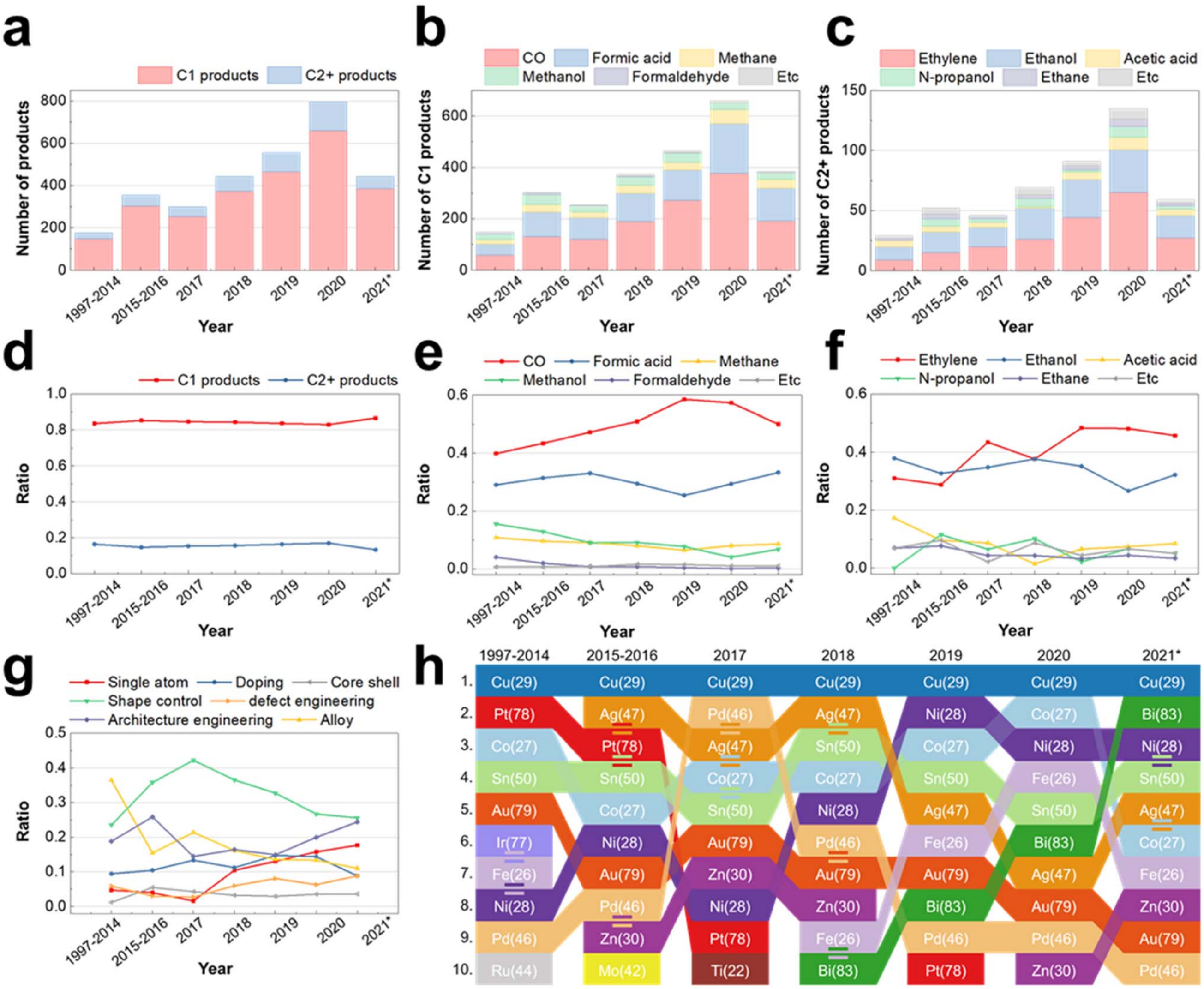
Example 4, Bi-LSTM: Herein, we prepared [Cu$_2$O/nitrogen-doped carbon shell] ([Cu$_2$O/NCS]) composite and employed it in the production of [C$_2$H$_4$] from CO$_2$ electroreduction. The ultra-thin NCS with graphene-like structure was prepared from petroleum asphalt by template method.

Example 3, MatBERT: The results demonstrate that maximum faradaic efficiency of [83.5%] is obtained at [−1.8 V] vs. [Ag/AgCl] when the [Sn/f-Cu electrode] is prepared by electrodeposition for 35 min.

Example 4, MatBERT: Herein, we prepared [Cu$_2$O/nitrogen-doped carbon shell] ([Cu$_2$O/NCS]) composite and employed it in the production of [C$_2$H$_4$] from CO$_2$ electroreduction. The ultra-thin NCS with graphene-like structure was prepared from petroleum asphalt by template method.

Fig. 5 Example of the results from the tokenizer and NER models. (a and b) The entity predictions of the NER models are shown. The text in parentheses is the ground truth, which means the target of the NER models, and the text with background color is the predicted label based on the NER models. Incorrect predictions by the NER models are highlighted in red. (c) Tokens tokenized through different BERT tokenizers are shown, and the red text emphasizes distinct tokens.

these annotation results were compared to the predicted results from the NER models, as shown in Table S6.† The average performance of our NER model across 100 papers was found to be approximately 91% which well demonstrates the scalability of our model and affirms the reliability of the extracted data. Using the extracted data, several yearly trend analyses were performed, as shown in Fig. 6. We specifically focused on trends in CO$_2$RR products, catalyst design strategies, and elements used in the catalysts. Fig. 6a–c show in various forms the number of CO$_2$RR products per year, and Fig. 6d–f show their relative ratios. Here, C1 products refer to products with a single carbon, such as CO, formic acid, methanol, and methane, while C2+ products refer to products with two or more carbons, such as ethylene and ethanol. Although research interest in the CO$_2$RR field has continued to increase, the ratios of C1 and C2+ products remained unchanged from 1997 to 2021, as shown in Fig. 6a and d. Although C2+ products are considered more valuable products in industry, the CO$_2$RR research community still puts more effort (>80%) into producing C1 products. This trend is probably because C1 products require less energy than C2+ products, making them easier to produce. Moreover, C1 products such as CO molecules are considered intermediate species for producing C2+ products.[51] In terms of electrical energy conversion efficiency, C1 products are more economical products because substantial energy costs are required for the reduction of carbon dioxide into high carbon products.[52] Fig. 6b and e present various C1 products in terms of their numbers and ratios; CO and formic acid, which are the simplest two-electron reduction products, appeared the most, with proportions of approximately 52% and 30%, respectively.[53] CO is likely the most studied product because it is an intermediate species that can be easily converted into other valuable C2+ products *via* additional reduction steps.[54] Fig. 6c and f show various C2+ products in a similar format; ethylene and ethanol, which are both 12-electron reactions, were the two most studied products over all periods. Over the last three years from 2019 to 2021, the ethylene ratio has increased, which is likely due to



	1997-2014	2015-2016	2017	2018	2019	2020	2021*
1.	Cu(29)	Cu(29)	Cu(29)	Cu(29)	Cu(29)	Cu(29)	Cu(29)
2.	Pt(78)	Ag(47)	Pd(46)	Ag(47)	Ni(28)	Co(27)	Bi(83)
3.	Co(27)	Pt(78)	Ag(47)	Sn(50)	Co(27)	Ni(28)	Ni(28)
4.	Sn(50)	Sn(50)	Co(27)	Co(27)	Sn(50)	Fe(26)	Sn(50)
5.	Au(79)	Co(27)	Sn(50)	Ni(28)	Ag(47)	Sn(50)	Ag(47)
6.	Ir(77)	Ni(28)	Au(79)	Pd(46)	Fe(26)	Bi(83)	Co(27)
7.	Fe(26)	Au(79)	Zn(30)	Au(79)	Au(79)	Ag(47)	Fe(26)
8.	Ni(28)	Pd(46)	Ni(28)	Zn(30)	Bi(83)	Au(79)	Zn(30)
9.	Pd(46)	Zn(30)	Pt(78)	Fe(26)	Pd(46)	Pd(46)	Au(79)
10.	Ru(44)	Mo(42)	Ti(22)	Bi(83)	Pt(78)	Zn(30)	Pd(46)

Fig. 6 NER-based yearly trend analysis of CO2RR products, catalyst design strategies, and elements in catalysts from 1997 to 2021. The yearly trends of (a–f) products and (g) strategies for catalyst synthesis are shown with the total counts. (h) The top 10 most used elements in order of ranking. The numbers in parentheses represent the atomic numbers of the elements. 2021* means that only papers published by June 2021 were used for analysis. Etc in the legends of (b) and (e) includes carbonate and bicarbonate, while Etc in (c) and (f) includes oxalate, acetaldehyde, and acetone.

ethylene having a lower energy barrier than ethanol, which means that ethylene is relatively easier to produce.[55]

Fig. 6g shows the trends in catalyst design strategy from 1997 to 2021. Seven strategies were selected, including core–shell, defect engineering, alloy, single atom, doping, architecture engineering, and shape control. These strategies were defined based on the words that appear frequently in the titles of the papers, as shown in Table S7.† Until 2017, the strategy of shape control dominated; however, its ratio gradually declined, whereas the ratios of the single-atom catalyst (SAC) and architecture engineering strategies clearly increased. In particular, SAC has recently drawn much attention because it offers the benefits of high catalytic activity, selectivity, and diverse elemental combinations.[56]

Dual-atom catalysts (DACs) have recently garnered much attention in the research and development of electroreduction processes. In comparison to SACs, they have not only a similar ability to utilize almost 100% of the catalyst atoms but also much enhanced electronic property, improved activity, and selectivity.[57] Upon examining the 3153 literature we collected, it was verified that there are 7 papers focusing on dual-atom catalysts. We searched Web of Science for dual-atom catalysts to collect the most recent developments in the field. There has been a considerable amount of research articles (25 papers) published in 2022 and 2023. Combining these recently acquired publications with the initial seven, we then applied our model to a total of 32 papers, and achieved approximately 88% F1-score (micro average over all entities) as summarized in Table S8.† We further analyzed the combinations of catalytic elements used in the dual-atom catalysts that were taken from the literature. As shown in Fig. S3,† the combinations of transition metals (Fe, Ni, Co, *etc.*) that are frequently utilized as single-atom catalysts were also most found in dual-atom catalysts.

Fig. 6h shows the trend of the most used elements (top 10 elements) in catalysts by year. Examining the overall ranking of the elements, it is evident that there has been a shift in trends, with a decreased reliance on noble metals and increased utilization of transition metals. As illustrated in Fig. S4,† transition metals such as Fe, Co, Ni, Cu, and Zn are associated with CO production and have similar or higher FE values compared to noble metals like Pd, Ag, and Au. Note that Cu was the most used in all years from 1997 to 2021. Cu is the only element offering appropriate binding energy to produce both C1 and C2+ products such as hydrocarbons and alcohols.[58] In contrast, the rank of Pt has decreased gradually every year. Pt is an expensive noble metal and has been reported to perform worse than Cu in terms of CO_2RR activities.[59] Interestingly, Bi has appeared in the top-10 ranking since 2018, and its frequency has increased rapidly. Bi is inexpensive, inert and environmentally friendly and, most importantly, is known to be highly active for formic acid production.[60]

An NER-based trend analysis of the CO_2RR performance was also carried out from 1997 to 2021, as shown in Fig. 7. Among several performance-related entities, including the faradaic efficiency (FE), current density, overpotential, turnover frequency, and stability hours,[61] the former two entities (FE and current density) were chosen for detailed trend analyses because these two entities were reported the most frequently in the abstracts of papers. Note that the FE (unit%) means the selectivity (unit%) in electrochemistry, and the current density (unit mA cm^{-2}) is the measured current value divided by the electrode area, which represents the reaction rate or catalytic activity.[62–65] In Fig. 7a–d, a clear difference was observed for FE trends between C1 products (CO and formic acid) and C2 products (ethylene and ethanol). Since 2016, the reported FE values for C1 products have mostly been over 80%, whereas for C2 products the FE values have typically been below 80%. These C2 products, including ethylene and ethanol, require complicated multistep reactions involving 12 electrons, and thus, selectively synthesizing these C2 products is more difficult than C1 products.

The trends of the reported current density values are presented in Fig. 7e–h. The current density values are lower for C2 products than C1 products. Similarly to the FE values, the complicated multistep reactions for the C2 product require a high overpotential and result in low activity compared to the relatively simpler reactions for the C1 products. Nevertheless, current density values have gradually increased over time. The improvements can be attributed to the development of both new catalyst materials and new electrochemical cell architectures. For example, the best performance of a CO-producing catalyst[62] was from the reaction taking place in solid oxide electrolysis cells (SOECs) at high temperatures. Similarly, the best performance of ethylene-producing catalysts[64] was achieved by introducing a new architecture to gas-flow electrochemical cells. However, the majority of the reported current density values were less than 200 mA cm^{-2}. Since the current density is required to be more than 200 mA cm^{-2} for CO_2RR catalysts to be utilized industrially,[66] there is still much room for improvement in terms of the current density.

NER-based guidance of research directions

NER offers the ability to conduct yearly trend analyses for diverse entities and may also offer novel insights regarding CO_2RR research directions. We present the explorable area map in Fig. 8a, which was generated by counting the catalytic elements and CO_2RR products that appear in the same papers. The area map shows which elements have been most actively used for specific products and indicates which parts were overlooked and are potentially promising. Some key information that can be determined based on the area map is summarized as follows. First, Cu is the most actively used catalytic element for producing both C1 and C2 products, and its dominance is particularly notable for C2 products, including ethylene and ethanol. Second, post-transition metals such as Sn and Bi have been widely used for the production of formic acid.[67] The yearly trends of metalloid and post-transition metal elements are provided in Fig. S5,† which shows a clear increase in Bi and In elements. Moreover, for formic acid productions, the percentage of Bi in the highest faradaic efficiency range (80–100%) is noticeably higher than other transition metals as shown in Fig. S4,† which supports the recent increasing trend of Bi. Finally, precious metals such as Ag, Pd, and Au are used as catalysts for C2 products, although they are not used in large numbers. Recent studies have shown that alloying precious metals with Cu is a viable approach to increase selectivity in generating C2 products.[68–70] Although Cu is known to have optimal CO binding energy for CO_2RRs, it suffers from low selectivity to specific products because of the numerous CO_2RR reaction pathways. This problem could be solved by alloying with other elements, such as precious metals. Similarly, other elements, such as transition metals and post-transition metals, could also be alloyed with Cu to increase the selectivity of C2 products.

Next, the relations among the elements, catalyst synthesis strategies, and CO_2RR products were analyzed using association rule mining (ARM)[71] and are visualized as graphs in Fig. 8b–g. ARM is a rule-based machine learning method for discovering interesting relationships between variables in large databases, which is often called market basket analysis. A detailed description of this algorithm is provided in the Methods section. ARM can be applied to reveal how often certain elements, products, and synthesis strategies are found simultaneously in papers. The nodes in the graph represent the types of entities, and the thickness of the edges represents their correlation strength.

Fig. 8b shows the entire graph of the ARM result, and Fig. 8c–g show the subgraphs that illustrate certain nodes and their connections for emphasis and detailed analysis. Fig. 8c presents a subgraph that highlights nodes that are connected to formic acid. Many post-transition metals, including Sn, Bi, and In, can be observed, which is consistent with the results in Fig. 8a. Fig. 8d shows a subgraph highlighting nodes that are connected to bismuth. We found that the shape control strategy appeared,

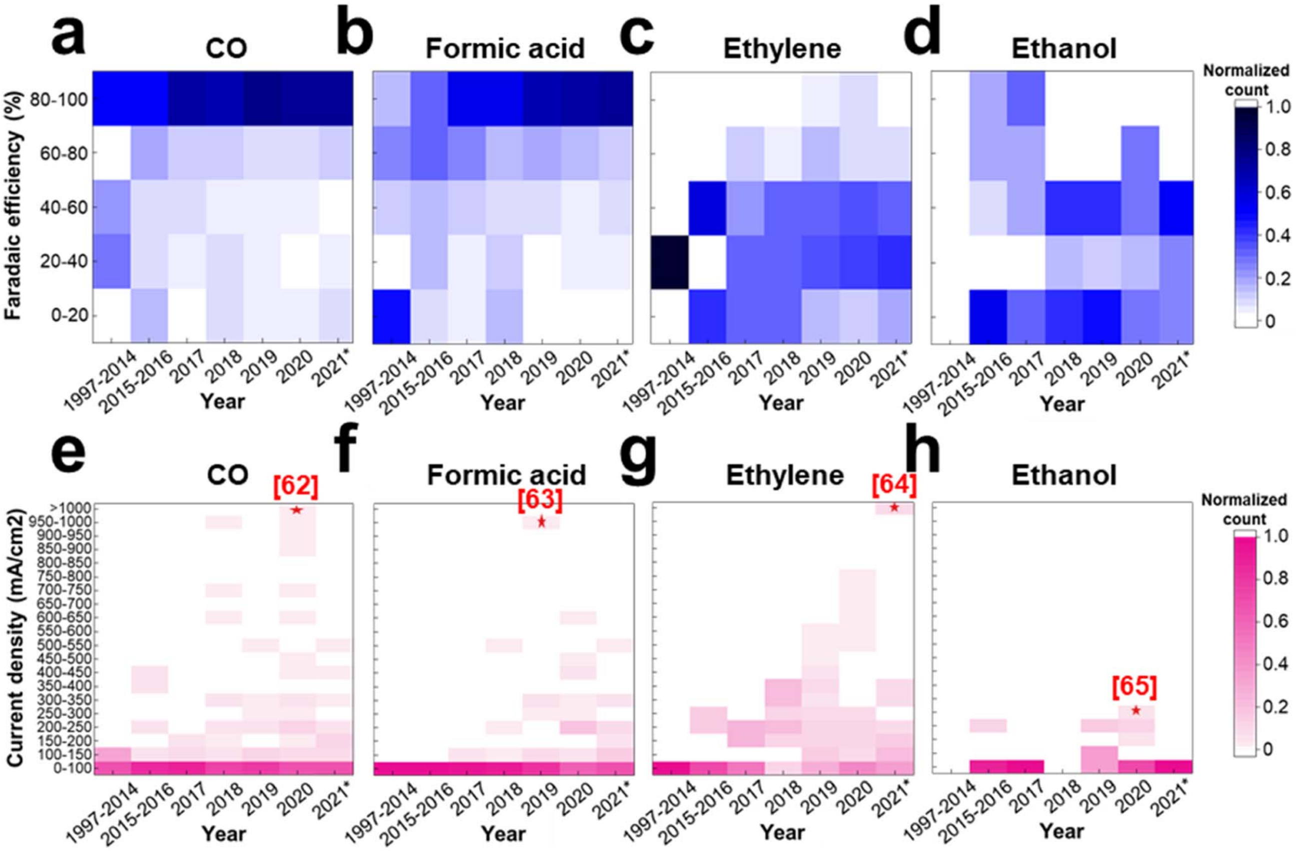


Fig. 7 NER-based yearly trend analysis of the CO2RR performance from 1997 to 2021. The heatmaps show the faradaic efficiency (blue) and current density (pink) by year. The color represents the normalized count, which is calculated by dividing the faradaic efficiency or current density in a specific range by the total faradaic efficiency or current density for that year. The red stars and numbers indicate the best performance of the current density for specific products and ref. 62–65 2021* means that only papers published by June 2021 were used for analysis.

indicating that Bi-based catalysts are frequently synthesized by shape control strategies for formic acid production. This kind of ternary relationship was not found in the area maps that relate only two entities but were noted in the ARM analysis.

Fig. 8e shows a subgraph highlighting nodes that are connected to ethylene. Interestingly, Cu and Ag appeared together, indicating that these two elements are synergistic for ethylene production.[72,73] Indeed, many Cu–Ag bimetallic catalysts were found, with Ag observed for desorption–resorption and diffusion processes for C–C coupling on the Cu surface to effectively synthesize ethylene species. In addition, in Fig. S6,† one can observe that Cu, Ag, and Co are the elements that perform well in ethanol productions. Further investigations of these relevant articles uncovered that the ethanol production involved various catalysts such as Cu–Ag,[65] AgCo alloy,[74] Ag nanoparticles deposited on 3D graphene-wrapped nitrogen-doped carbon foam,[75] and CoO-anchored N-doped carbon materials comprising mesoporous carbon (MC) and carbon nanotubes (CNT).[76] Based on these findings, we believe that the development of catalysts employing synthetic strategies like bimetallic catalysts combining Cu with other elements and the incorporation of Co and Ag into carbon matrices through doping hold promise for the creation of catalysts capable of producing C2 products such as ethanol and ethylene.

Fig. 8f presents a subgraph highlighting nodes that are connected to SAC, where transition metals, including Ni, Co and Fe, can be observed.[77] This result is consistent with the explorable area map relating the product and catalytic elements for the SAC strategy case, as provided in Fig. S7.† The SAC strategy is particularly active for CO production and is generally not functional for other products.[78] SAC is not linked to any of the C2 products because SAC cannot be used for C–C coupling due to its structural limitations. Furthermore, the distribution of 151 SAC papers (for CO productions) connected over used elements and FE value ranges, as shown in Fig. S8.† It was observed that transition metals such as Ni, Fe, Co, Cu, and Zn are commonly used as SACs. Also, we found that Ni and Fe are mostly connected to the highest FE range (90–100%). Cobalt (Co) overall underperforms than Ni or Fe, and this finding is probably due that Co SAC has low activation barriers for hydrogen evolution reaction (HER) and low selectivity to CO_2RR.[79] Interestingly, Ni is linked to the SAC node, although bulk Ni is not active for CO production because of severe CO poisoning issues due to its very strong binding energy.[61,80,81] However, Ni in SAC offers a reduced CO binding energy and can thereby increase the activity and selectivity for CO production.[80]

Fig. 8g shows subgraphs that highlight nodes that are connected to Mo, Mn, and Cr. In our dataset, no paper reported the

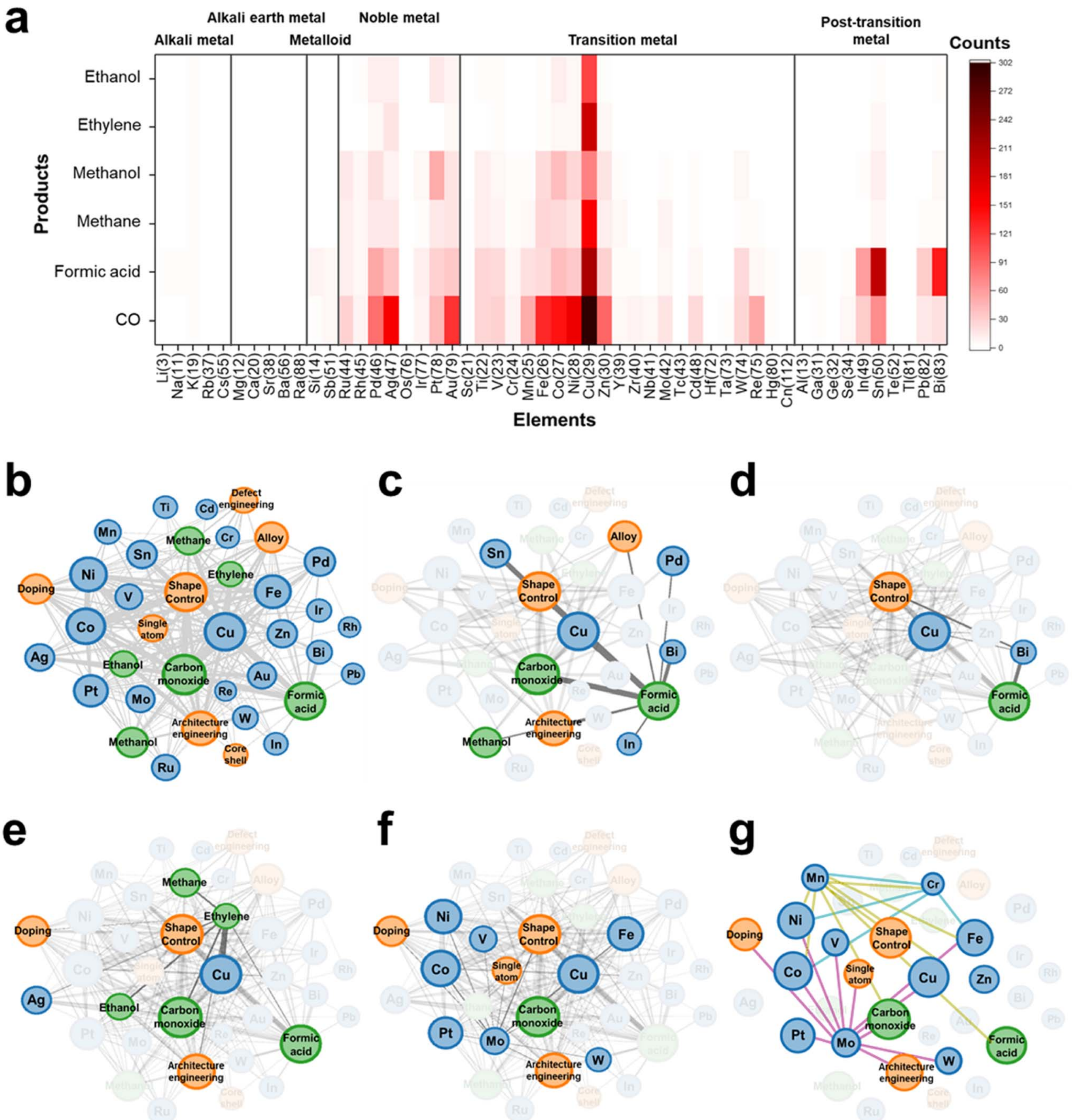


Fig. 8 NER-based guidance of research directions. (a) Explorable area heatmaps of products for each element in CO_2RR catalysts. The color represents the count of products related to the specific elements that appear in the same abstract. Groups of elements are divided with vertical lines. In the tick labels on the *x*-axis, the number in parentheses represents the atomic number. (b–g) Visualization of the ARM results. The nodes with blue, orange, and green colors represent the elements of the catalyst, the strategy to synthesize the catalyst, and the product, respectively. The thickness of an edge represents the association between two nodes. (b) Is the entire graph of the ARM result. (c) Is a subgraph highlighting nodes that are connected to formic acid. (d) Is a subgraph highlighting nodes that are connected to bismuth. (e) Is a subgraph highlighting nodes that are connected to ethylene. (f) Is a subgraph highlighting nodes that are connected to SAC. (g) Subgraphs highlighting nodes that are connected Mo, Mn, and Cr are merged. Nodes that are connected to Mo, Mn, and Cr are represented by pink, green, and light blue edges, respectively.

effectiveness of Mo, Mn, and Cr elements as SACs in CO_2RRs. However, these elements were strongly connected to Fe, Co, and Ni, which are the most active elements for SAC. In ARM, the strong connections indicate that the connected components are used together in the same study and may thus possess similar properties. For example, Cr and Mn were used as SACs on graphene for CO oxidation and had similar CO adsorption energies compared to Fe, Co, and Ni.,[82] which suggests the potential of

Cr, Mn, and Mo to be used as SACs for CO_2RRs. In fact, we found some recent papers (not included in our dataset) reporting Cr, Mn, and Mo elements as SACs,[83,84] which supports the usefulness of our NER models and the subsequent analyses.

Methods

Doc2Vec and LDA training

We trained Doc2Vec and LDA for document clustering before developing the NER models. Since the words in the corpus are used for learning, text preprocessing was used to reduce the corpus complexity so that only meaningful words were used for training. Doc2Vec was trained using both the abstracts and main texts of 4838 papers. These 4838 papers were obtained from the keyword query-based screening process with the keywords "CO2" AND "reduction" AND "reaction" AND "catalyst" AND "faradaic" AND ("electrochemical" OR "electroreduction" OR "electrocatalytic" OR "electro"). The papers used for training were tokenized using ChemDataExtractor.[32] The chemical finding function in ChemDataExtractor was used to preprocess all tokens except chemicals to lowercase. The size of the document vector was set to 200.

LDA was trained using only the abstracts of 4838 papers. Preprocessing similar to that performed for Doc2Vec was applied, and lemmatization was performed to find the lemma of the word inflected in various forms in a sentence using the NLTK library.[85] In addition, stop words, words with one letter and words with an occurrence frequency of 20 or less were removed. After the number of topics (k) is determined, the LDA algorithm finds k topics based on all documents and determines which topics a random document contains. To develop a proper LDA model, it is vital to set an appropriate number of topics. To determine the appropriate number of topics, both perplexity and coherence were considered. Perplexity determines how accurately the probabilistic model predicts the outcome, while coherence determines how semantically consistent the topics are. In our case, the number of topics (k) was chosen as 11 because of the small perplexity and high coherence values associated with this number of topics.

NER model development

For the NER models, both the Bi-LSTM CRF model and BERT CRF model were used. First, we trained the Bi-LSTM model using the flair library.[86] In the Bi-LSTM model, the embedding layer consists of a 100-dimensional word embedding layer and a 50-dimensional character-level word embedding layer. In the Bi-LSTM model, the maximum number of epochs was 100, the learning rate was 0.2, and the batch size was 8. Word embeddings were pretrained with both the abstracts and main texts of 4838 papers using Word2Vec. On the other hand, for BERT models, we retrained the existing BERT models (MatBERT, MatSciBERT, SciBERT, and BERT base) using our own corpus obtained from the catalysis literature. The tokens in the BERT models were generated using each model's tokenizer. Each pretrained model was retrained with 500 CO_2RR papers that were selected according to the Doc2Vec-based ranking, as shown in Fig. 2c. The optimizer was AdamW, and training was carried out while changing the learning rate using the warmup scheduler. The batch size was selected as 32 based on hyperparameter fitting, as shown in Table S9.† The Bi-LSTM and BERT models both have CRF layers to address sequence problems associated with entities in IOBES format by learning the constraints and rules, including that the tag of the first word in a label starts with B (Beginning) in IOBES format.

The annotated abstracts were divided into training, validation, and test sets prior to classifier layer training. The test set was used to evaluate the model's final performance based on data that were not used for training, while the validation set was used to optimize the model's hyperparameters. The training, validation, and test sets were divided according to the ratio 8 : 1 : 1. Tenfold cross validation was used to obtain more generalized performance compared to training with only one division.

Postprocessing after NER-based data extraction

We created a dictionary of products and chemical elements to process synonyms among the words extracted using the NER model. For example, copper and Cu are the same element, and both were converted to Cu. Additionally, CH_4 and methane are the same chemical, and both were converted to methane. Among the extracted entities, there are various units for the same entity. For example, the current density values can be expressed in mA cm^{-2}, A cm^{-2}, and μA cm^{-2} depending on the authors' needs. In this case, the current density values were all converted to mA cm^{-2} through the Python code.

Association rule mining

We used the Python library mxltend to implement ARM. Apriori is a popular algorithm for ARM that extracts frequent item sets from given data. The apriori function requires data in a one-hot encoded data frame format, and thus, we used TransactionEncoder to prepare an appropriate data format. Frequent item sets were formed with min_support values greater than or equal to 0.005. Support means the number of appearances of a specific item divided by the total number of instances. min_support means the minimum support for the item combination to be returned. In other words, if the support of an item combination is 0.005 or less, it is not formed as a frequent item set. When creating a frequent item set, there is no limitation on the maximum number of item combinations, and all possible item combinations under the given condition are returned. Association rules are created based on frequent item sets, which are the *a priori* results. The default value confidence was used as an evaluation index, and min_threshold was set to 0.005 to generate only association rules with confidence levels of 0.005 or higher.

Conclusions

This work presents a text mining protocol for catalysis literature, taking CO_2RR-related literature as an example. Performing text mining based on a large volume of literature in a specific target domain requires very careful and specialized annotations and model development, unlike the same tasks in general areas. We first developed a method based on Doc2Vec and LDA to selectively screen CO_2RR-related papers while reducing the number of irrelevant papers. This method led to the collection of 3154 CO_2RR papers, which were used for further analysis. This method requires the input of only tens of seed papers prepared by a user and may therefore be effective for use in other domains. Next, we developed and reported a MatBERT-based NER model, which was applied to extract twelve key entities, including catalyst names and catalytic performance information in the CO_2RR papers. This model achieved an extraction F1-score of 90.4% (up to 95.2% in the boundary relaxation evaluation scheme). The NER-based accelerated data extraction scheme from a large volume of literature enables several interesting analyses, such as the yearly trend analyses of CO_2RR catalysts, products, and performance. This analysis highlights the recent element usage trends, such as the boost of post-transition metal elements of Bi and In for formic acid production. In addition, it also reveals the potentially effective material space in CO_2RRs, including transition metal elements of Cr, Mn and Mo for SAC. This work is, to our understanding, the first attempt to apply text mining to a large volume of catalysis literature and will serve as a great reference for similar future studies in the catalysis field.

Finally, we note the limitations of our study as well as potential future studies, especially for large scale. First, this study explored only abstracts of the papers as opposed to the full bodies of the papers. For example, the catalytic performance may be greatly influenced by the type of cell; however, our study was not able to obtain this information because such specific information is typically found in the main body rather than in the abstract. Thus, a performance analysis considering these detailed parameters was unfortunately not conducted. Expanding our study to the entire body of the paper remains an important future research direction. The second limitation of this work is the absence of relation extraction between entities. NER enables only entity extraction and cannot be used to determine the relationships among the entities. For example, the catalyst names and catalytic performance were not systematically linked. The lack of proper entity relations may limit data utilization. Thus, it is important to extract entities and determine their relations using NLP techniques such as relation extraction.[87,88] Third, the data extractions from tables and figures are also important, as they often contain key information, such as performance data. This information cannot be obtained using current NER methods. The tabular data from the tables of a paper could be handled using data analysis libraries including Pandas;[89] however, it is important to standardize the tabular data because the formats of the table are much different across papers and journals. For figures, numerous vision techniques are rapidly developing which are applicable to a graph[90] and microscopy images,[91,92] which were proven to be very effective to extract the data from the figures in papers. Lastly, we observed that the NER performance is lower for catalyst name entities than for other entities, probably because researchers describe catalyst names in various forms in different papers. The MatBERT model pretrained with materials science papers worked fairly well after fine-tuning based on catalysis literature; however, the accuracy can be further increased by developing catalysis-based BERT models pretrained with millions of catalysis papers, which should be considered in a future study. Despite these limitations, the large amount of data extracted from the catalyst literature allowed us to evaluate valuable information, such as research trends and directions in the CO_2RR field, which will hopefully inspire similar efforts in other catalysis fields.

Data and code availability

The text-mining code and related data are available at **https://github.com/KIST-CSRC/CO2RR_NER** or can be obtained from corresponding authors upon request.

Author contributions

D. K. and S. S. H. conceived the idea and supervised the project. Jiwoo Choi and K. B. prepared the training dataset from literature and performed most of deep learning studies. S. J. contributed to Bi-LSTM model development. Jaewoong Choi and B. L. contributed to ARM analysis. J. O., D. B., A. H., T. Y.-J. H., S. S. S., and K.-R. L. contributed to result analyses. All authors contributed to manuscript writing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Center for Materials Research Data (NCMRD) through the National Research Foundation of Korea funded by the Ministry of Science and ICT (NRF-2021M3A7C2089739) and KIST institutional projects (2E32531 and 2E32533). This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program Under Project No. 22-ERD-027. LLNL-JRNL-846352.

References

1 O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti and G. Ceder, *Iscience*, 2021, **24**, 102155.
2 T. Zhou, Z. Song and K. Sundmacher, *Engineering*, 2019, **5**, 1017–1026.
3 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner and G. Ceder, *APL Mater.*, 2013, **1**, 011002.
4 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, *npj Comput. Mater.*, 2015, **1**, 1–15.
5 C. Draxl and M. Scheffler, *MRS Bull.*, 2018, **43**, 676–682.

6 R. Tran, J. Lan, M. Shuaibi, S. Goyal, B. M. Wood, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi and N. Shoghi, *arXiv*, 2022, preprint, arXiv:2206.08917, DOI: **10.48550/arXiv.2206.08917**.

7 D. Nadeau and S. Sekine, *Lingvisticae Investigationes*, 2007, **30**, 3–26.

8 Clarivate, *Web of Science*, **https://clarivate.com/webofsciencegroup/**, accessed on 2019.

9 J. F. Burnham, *Biomed. Digit Libr.*, 2006, **3**, 1–8.

10 E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum and S. Jegelka, *J. Chem. Inf. Model.*, 2020, **60**, 1194–1201.

11 H. Park, Y. Kang, W. Choe and J. Kim, *J. Chem. Inf. Model.*, 2022, **62**, 1190–1198.

12 L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder and A. Jain, *J. Chem. Inf. Model.*, 2019, **59**, 3692–3702.

13 A. M. Hiszpanski, B. Gallagher, K. Chellappan, P. Li, S. Liu, H. Kim, J. Han, B. Kailkhura, D. J. Buttler and T. Y.-J. Han, *J. Chem. Inf. Model.*, 2020, **60**, 2876–2887.

14 T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari and G. Ceder, *Chem. Mater.*, 2020, **32**, 7861–7873.

15 H. Huo, Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan and G. Ceder, *npj Comput. Mater.*, 2019, **5**, 1–7.

16 D. Kim, J. Lee, C. H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung and J. Kang, *IEEE Access*, 2019, **7**, 73729–73740.

17 S. Raza and B. Schwartz, presented in part at the *Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022, Proceedings of Machine Learning Research*, 2022.

18 A. Śniegula, A. Poniszewska-Marańda and Ł. Chomątek, *Procedia Comput. Sci.*, 2019, **160**, 260–265.

19 Z. Huang, W. Xu and K. Yu, *arXiv*, 2015, preprint, arXiv:1508.01991, DOI: **10.48550/arXiv.1508.01991**.

20 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *arXiv*, 2018, preprint, arXiv:1810.04805, DOI: **10.48550/arXiv.1810.04805**.

21 A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder and A. Jain, *Patterns*, 2022, **3**, 100488.

22 A. Friedrich, H. Adel, F. Tomazic, J. Hingerl, R. Benteau, A. Maruscyk and L. Lange, *arXiv*, 2020, preprint, arXiv:2006.03039, DOI: **10.48550/arXiv.2006.03039**.

23 I. Beltagy, K. Lo and A. Cohan, *arXiv*, 2019, preprint, arXiv:1903.10676, DOI: **10.48550/arXiv.1903.10676**.

24 W. J. Wilbur, A. Rzhetsky and H. Shatkay, *BMC Bioinf.*, 2006, **7**, 1–10.

25 T. Wang, J. Yang, J. Chen, Q. He, Z. Li, L. Lei, J. Lu, M. K. Leung, B. Yang and Y. Hou, *Chin. Chem. Lett.*, 2020, **31**, 1438–1442.

26 C. Zhu, G. Shen, W. Chen, X. Dong, G. Li, Y. Song, W. Wei and Y. Sun, *J. Power Sources*, 2021, **495**, 229814.

27 Q. Le and T. Mikolov, presented in part at the *Proceedings of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research*, 2014.

28 D. M. Blei, A. Y. Ng and M. I. Jordan, *J. Mach. Learn. Res.*, 2003, **3**, 993–1022.

29 R. T.-H. Tsai, S.-H. Wu, W.-C. Chou, Y.-C. Lin, D. He, J. Hsiang, T.-Y. Sung and W.-L. Hsu, *BMC Bioinf.*, 2006, **7**, 1–8.

30 T. T. Dinh, T. P. Vo-Chanh, C. Nguyen, V. Q. Huynh, N. Vo and H. D. Nguyen, *BMC Bioinf.*, 2022, **23**, 1–21.

31 X. Li, H. Liu, F. Kury, C. Yuan, A. Butler, Y. Sun, A. Ostropolets, H. Xu and C. Weng, *AMIA Summits on Translational Science Proceedings*, 2021, vol. 2021, p. 394.

32 M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.

33 T. Mikolov, K. Chen, G. Corrado and J. Dean, *arXiv*, 2013, preprint, arXiv:1301.3781, DOI: **10.48550/arXiv.1301.3781**.

34 T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, *Adv. Neural Inf. Process. Syst.*, 2013, **26**.

35 D. Mimno, H. Wallach, E. Talley, M. Leenders and A. McCallum, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, 2011, pp. 262–272.

36 B. Tang, H. Cao, Y. Wu, M. Jiang and H. Xu, presented in part at the *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, Maui, Hawaii, USA, 2012.

37 B. Tang, H. Cao, Y. Wu, M. Jiang and H. Xu, *BMC Med. Inf. Decis. Making*, 2013, **13**, S1.

38 S. Hochreiter and J. Schmidhuber, *Neural Comput.*, 1997, **9**, 1735–1780.

39 G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, *arXiv*, 2016, preprint, arXiv:1603.01360, DOI: **10.48550/arXiv.1603.01360**.

40 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Adv. Neural Inf. Process. Syst.*, 2017, **30**.

41 T. Gupta, M. Zaki and N. Krishnan, *npj Comput. Mater.*, 2022, **8**, 1–11.

42 J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, *Bioinformatics*, 2020, **36**, 1234–1240.

43 D. Araci, *arXiv*, 2019, preprint, arXiv:1908.10063, DOI: **10.48550/arXiv.1908.10063**.

44 H. M. Wallach, *Technical Reports (CIS)*, 2004, p. 22.

45 T. Gao, T. Xie, N. Han, S. Wang, K. Sun, C. Hu, Z. Chang, Y. Pang, Y. Zhang and L. Luo, *ACS Appl. Energy Mater.*, 2019, **2**, 3151–3159.

46 R. T.-H. Tsai, S.-H. Wu, W.-C. Chou, Y.-C. Lin, D. He, J. Hsiang, T.-Y. Sung and W.-L. Hsu, *BMC Bioinf.*, 2006, **7**, 1–8.

47 P. Liu, H. Liu, S. Zhang, J. Wang and C. Wang, *J. Colloid Interface Sci.*, 2021, **602**, 740–747.

48 D. Liu, Y. Hu, E. Shoko, H. Yu, T. T. Isimjan and X. Yang, *Electrochim. Acta*, 2021, **365**, 137343.

49 Y. Wang, J. Zhou, W. Lv, H. Fang and W. Wang, *Appl. Surf. Sci.*, 2016, **362**, 394–398.

50 H. Ning, X. Wang, W. Wang, Q. Mao, Z. Yang, Q. Zhao, Y. Song and M. Wu, *Carbon*, 2019, **146**, 218–223.

51 D. Li, H. Zhang, H. Xiang, S. Rasul, J.-M. Fontmorin, P. Izadi, A. Roldan, R. Taylor, Y. Feng and L. Banerji, *Sustainable Energy Fuels*, 2021, **5**, 5893–5914.

52 O. S. Bushuyev, P. De Luna, C. T. Dinh, L. Tao, G. Saur, J. van de Lagemaat, S. O. Kelley and E. H. Sargent, *Joule*, 2018, **2**, 825–832.

53 B. Zha, C. Li and J. Li, *J. Catal.*, 2020, **382**, 69–76.

54 L. Wang, S. A. Nitopi, E. Bertheussen, M. Orazov, C. G. Morales-Guio, X. Liu, D. C. Higgins, K. Chan, J. K. Nørskov and C. Hahn, *ACS Catal.*, 2018, **8**, 7445–7454.

55 J. Wang, H. Yang, Q. Liu, Q. Liu, X. Li, X. Lv, T. Cheng and H. B. Wu, *ACS Energy Lett.*, 2021, **6**, 437–444.

56 Y. Chen, S. Ji, C. Chen, Q. Peng, D. Wang and Y. Li, *Joule*, 2018, **2**, 1242–1264.

57 M. Jafarzadeh and K. Daasbjerg, *ACS Appl. Energy Mater.*, 2023, **6**(13), 6851–6882.

58 L. Zaza, K. Rossi and R. Buonsanti, *ACS Energy Lett.*, 2022, **7**, 1284–1291.

59 W. Niu, J. Wu, C. Chen, Y. You, Y. Zhu, L. Lu, P. Cheng and S. Zhang, *Appl. Phys. Lett.*, 2022, **120**, 143904.

60 P. Deng, H. Wang, R. Qi, J. Zhu, S. Chen, F. Yang, L. Zhou, K. Qi, H. Liu and B. Y. Xia, *ACS Catal.*, 2020, **10**, 743–750.

61 X. Zhang, S.-X. Guo, K. A. Gandionco, A. M. Bond and J. Zhang, *Mater. Today Adv.*, 2020, **7**, 100074.

62 S. Wang, H. Jiang, Y. Gu, B. Yin, S. Chen, M. Shen, Y. Zheng, L. Ge, H. Chen and L. Guo, *Electrochim. Acta*, 2020, **337**, 135794.

63 A. Löwe, C. Rieg, T. Hierlemann, N. Salas, D. Kopljar, N. Wagner and E. Klemm, *ChemElectroChem*, 2019, **6**, 4497–4506.

64 F. P. García de Arquer, C.-T. Dinh, A. Ozden, J. Wicks, C. McCallum, A. R. Kirmani, D.-H. Nam, C. Gabardo, A. Seifitokaldani and X. Wang, *Science*, 2020, **367**, 661–666.

65 Y. C. Li, Z. Wang, T. Yuan, D.-H. Nam, M. Luo, J. Wicks, B. Chen, J. Li, F. Li and F. P. G. De Arquer, *J. Am. Chem. Soc.*, 2019, **141**, 8584–8591.

66 T. Fan, W. Ma, M. Xie, H. Liu, J. Zhang, S. Yang, P. Huang, Y. Dong, Z. Chen and X. Yi, *Cell Rep. Phys. Sci.*, 2021, **2**, 100353.

67 D. Wang, C. Liu, Y. Zhang, Y. Wang, Z. Wang, D. Ding, Y. Cui, X. Zhu, C. Pan and Y. Lou, *Small*, 2021, **17**, 2100602.

68 S. Lee, G. Park and J. Lee, *ACS Catal.*, 2017, **7**, 8594–8604.

69 H.-P. Yang, S. Qin, Y.-N. Yue, L. Liu, H. Wang and J.-X. Lu, *Catal. Sci. Technol.*, 2016, **6**, 6490–6494.

70 F. Jia, X. Yu and L. Zhang, *J. Power Sources*, 2014, **252**, 85–89.

71 S. Kotsiantis and D. Kanellopoulos, *GESTS International Transactions on Computer Science and Engineering*, 2006, vol. 32, pp. 71–82.

72 Y. E. Jeon, Y. N. Ko, J. Kim, H. Choi, W. Lee, Y. E. Kim, D. Lee, H. Y. Kim and K. T. Park, *J. Ind. Eng. Chem.*, 2022, **116**, 191–198.

73 L. R. L. Ting, O. Pique, S. Y. Lim, M. Tanhaei, F. Calle-Vallejo and B. S. Yeo, *ACS Catal.*, 2020, **10**, 4059–4069.

74 Q. Zhang, S. Tao, J. Du, A. He, Y. Yang and C. Tao, *J. Mater. Chem. A*, 2020, **8**, 8410–8420.

75 K. Lv, Y. Fan, Y. Zhu, Y. Yuan, J. Wang and Q. Zhang, *J. Mater. Chem. A*, 2018, **6**, 5025–5031.

76 J. Du, S. Li, S. Liu, Y. Xin, B. Chen, H. Liu and B. Han, *Chem. Sci.*, 2020, **11**, 5098–5104.

77 Y. Zhu, X. Yang, C. Peng, C. Priest, Y. Mei and G. Wu, *Small*, 2021, **17**, 2005148.

78 J. Zhang, W. Cai, F. X. Hu, H. Yang and B. Liu, *Chem. Sci.*, 2021, **12**, 6800–6819.

79 W. Ju, A. Bagger, G.-P. Hao, A. S. Varela, I. Sinev, V. Bon, B. Roldan Cuenya, S. Kaskel, J. Rossmeisl and P. Strasser, *Nat. Commun.*, 2017, **8**, 944.

80 C. W. Lee, C. Kim and B. K. Min, *Nano Convergence*, 2019, **6**, 1–11.

81 Y. Hori, H. Wakebe, T. Tsukamoto and O. Koga, *Electrochim. Acta*, 1994, **39**, 1833–1839.

82 L. Xu, L.-M. Yang and E. Ganz, *Theor. Chem. Acc.*, 2018, **137**, 1–13.

83 F. Pan, W. Deng, C. Justiniano and Y. Li, *Appl. Catal., B*, 2018, **226**, 463–472.

84 P. Huang, M. Cheng, H. Zhang, M. Zuo, C. Xiao and Y. Xie, *Nano Energy*, 2019, **61**, 428–434.

85 E. Loper and S. Bird, *arXiv*, 2002, preprint, arXiv:cs/0205028, DOI: **10.48550/arXiv.cs/0205028**.

86 A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter and R. Vollgraf, 2019.

87 N. Konstantinova, 2014.

88 Z. Nasar, S. W. Jaffry and M. K. Malik, *ACM Computing Surveys (CSUR)*, 2021, **54**, 1–39.

89 W. McKinney, *Python for High Performance and Scientific Computing*, 2011, vol. 14, pp. 1–9.

90 F. Bajić and J. Job, *Journal of Imaging*, 2022, **8**, 136.

91 L. Yao, Z. Ou, B. Luo, C. Xu and Q. Chen, *ACS Cent. Sci.*, 2020, **6**, 1421–1430.

92 L. von Chamier, R. F. Laine, J. Jukkala, C. Spahn, D. Krentzel, E. Nehme, M. Lerche, S. Hernández-Pérez, P. K. Mattila and E. Karinou, *Nat. Commun.*, 2021, **12**, 2276.