Catalysis Science & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Catal. Sci. Technol.*, 2025, **15**, 1217

Machine learning and text mining approaches to design selective catalyst reduction synthesis routes†

Shuyuan Li,[a] Chenyu Huang,[a] Yunjiang Zhang,[a] Jing Li[b] and Shaorui Sun *[a]

Received 29th September 2024,
Accepted 8th January 2025

DOI: 10.1039/d4cy01159g

rsc.li/catalysis

The development of selective catalytic reduction (SCR) catalysts is often hindered by the complexity of experimental processes and the time-consuming trial-and-error approaches. Machine learning offers a promising solution by enabling more efficient and data-driven catalyst design. In this study, information was automatically extracted from the SCR-related scientific literature, including catalyst synthesis and catalyst properties, using rule-based techniques. These extracted data were then structured through feature engineering to build a machine learning-ready dataset. Models such as extreme gradient boosting regression (XGBR) and random forest (RF) were employed to predict catalyst performance and identify key factors influencing selectivity and conversion rates. To optimize synthesis routes, the designed synthesizable space was combined with the machine learning models to optimize key parameters and predict synthesis routes for SCR catalysts. Finally, synthesis information for SCR catalysts with high-performance was recommended. This work demonstrates the potential of using machine learning to accelerate SCR catalyst development, providing a scalable method for designing more efficient catalysts.

1. Introduction

Selective catalytic reduction (SCR) is a widely used process for reducing nitrogen oxide (NO_x) emissions from combustion engines, power plants, and the other industrial processes.[1,2] Catalysis technology is a crucial element in environmental technology and engineering, promoting sustainable environmental protection and efficient resource utilization by enhancing reaction efficiency and minimizing energy consumption. Due to increasing environmental regulations and the urgent need to mitigate air pollution, the design of efficient and robust SCR catalysts has become a focal point in the field of catalysis.[3] The conversion–selectivity trade-off is very commonly observed in catalytic systems, and constructing rapid screening methods for desired high-performance catalysts for reactions is a tremendous challenge.[4] In the traditional catalyst development process, researchers usually rely on repeated tests and experience accumulation in the laboratory to screen and optimize catalyst formulations and synthesis routes.[5] Catalyst development often involves a large number of complex variables, such as the selection of raw materials, the setting of reaction conditions (temperature, reaction time, *etc.*), as well as the microstructure and physicochemical properties of the catalyst.[6,7] Due to the complex correlation between these variables, determining the optimal catalyst formulation and synthesis conditions often requires multiple iterations of experiments, each of which may involve time-consuming preparation, running and analysis.[8] This makes the entire development process very slow and resource intensive.

With the development of data science, machine learning (ML) has become an essential tool in catalyst design and optimization, offering new ways to handle complex catalytic systems.[9–11] Compared with traditional experiment-driven approaches, ML methods provide unprecedented predictive capabilities for the properties of new materials through the analysis and modeling of large-scale datasets. Researchers have explored a variety of ML techniques, such as traditional ML models,[12–15] ensemble models,[14–18] and deep neural networks,[19–21] to predict catalyst performance and discover novel materials. These approaches have been applied to a wide range of catalytic processes, including the selective catalytic reduction,[22] hydrogen evolution reaction,[23] oxygen reduction reaction, and others.[24] Dong *et al.*[4] identified acidity descriptors and redox descriptors and applied a cluster analysis method to rapidly screen high-performance multi-element oxide SCR catalysts. Roy *et al.*[25] used an ML approach to aid in the exploration of high entropy alloy-

[a] *Department of Chemical Engineering and Technology, College of Materials Science and Engineering, Beijing University of Technology, Beijing, 100124, China. E-mail: sunsr@bjut.edu.cn*
[b] *School of Mathematics and Physics, Nanjing Institute of Technology, Nanjing 211167, China*

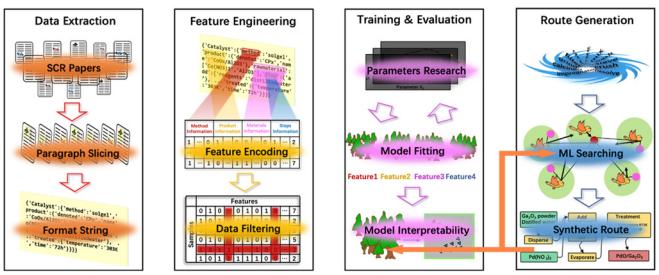
† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4cy01159g

based catalysts for selective reduction of carbon dioxide to methanol. Kim *et al.*[26] explored the important factors that can affect the selective SCR system in diesel engines using an ML approach. Okeleye *et al.*[27] focused on a data-driven approach to developing a model for predicting NO_x conversion efficiency across the SCR using three ML methods. Bae *et al.*[28] constructed an ML model utilizing a decision tree and evaluated the causal relationship between features and the NO_x removal efficiency of zeolite-based SCR catalysts at low temperatures, with the help of the constructed descriptors or some parameters that are strongly correlated with catalyst performance; an extensive search of the catalyst chemical space can be realized based on limited data. Constructing reliable and suitable descriptors is a fundamental step in quickly screening catalysts by this paradigm. However, there are still no constructed synthetic information descriptor relationships that can be used to describe the performance of SCR catalysts.

To realize AI-automated design of catalytic materials with increasing complexity, a large amount of high-quality data from high-throughput experiments, computational modeling, and literature is essential. High-throughput experiments provide vast datasets of experimental results under varied conditions, enabling rapid exploration of catalyst composition and performance. Computational modeling complements this by offering insights into reaction mechanisms, and theoretical predictions of catalytic behavior. However, literature data, derived from decades of published research, represent a valuable but underutilized source of information. Extracting data from studies allows researchers to leverage existing experimental and theoretical findings, significantly broadening the available dataset without requiring additional costly and time-consuming experiments.

A significant challenge in leveraging literature data for text mining in catalysis and chemistry lies in data standardization. Literature data are inherently unstructured, reported in diverse formats. Literature data are often presented in diverse formats, units, and terminologies, which introduces inconsistencies that hinder the integration and utilization of these data. Studies may report the same performance metric using different units or terms, and critical details about experimental conditions might be described ambiguously or omitted altogether. Addressing these issues requires robust preprocessing pipelines, normalization of data units, and, in some cases, manual validation to ensure the accuracy of extracted information. By overcoming the challenges of data standardization and integration, it becomes possible to unlock the full potential of machine learning for the AI-driven design of catalytic materials. To address these issues, our methodology incorporates several strategies. First, a preprocessing pipeline was developed to normalize extracted data, including the conversion of units and the alignment of terminologies. Second, regular expression-based methods were used to identify and standardize key performance metrics and reaction conditions. Third, when ambiguities arose in the text, manual validation was performed to ensure data quality and consistency.

The ML techniques were combined with the text mining approach to systematically extract and analyze synthesis parameters from the scientific literature on SCR catalysts. By leveraging a large corpus of scientific papers, a comprehensive database of synthesis conditions, catalysts, and performance data was built. ML models were employed to predict catalyst performance and identify key factors influencing selectivity and conversion rates. To optimize synthesis routes, the designed synthesizable space was combined with the ML models to optimize key parameters and predict synthesis routes for SCR catalysts. Finally, synthesis information for SCR catalysts with high-performance was recommended. The integration of text mining and machine learning for catalyst design represents a paradigm shift in the field of materials science, where large-scale data extraction from studies can lead to more efficient and informed catalyst development.



2. Methods

2.1 Data extraction

The acquisition of a comprehensive corpus is essential for facilitating in-depth analysis of SCR synthesis pathways. To achieve this, a Python-based web-scraping program was developed, which systematically queried the ScienceDirect homepage with the search term "selective catalytic reduction" and recorded the responses from each page. As a result, a total of 12 136 articles were obtained, primarily in XML format. This format is advantageous as it allows for easier extraction of structured data, reducing errors associated with other formats and enhancing the efficiency of subsequent analyses.

The structure of XML files follows a hierarchical tree model, where each element is nested within a node. To extract meaningful information from this format, the process begins by identifying the root node, which contains all the major headings of the article. Using keyword-based searches such as "abstract" and "conclusion," the corresponding sections of the document were located by navigating through the element tree. Once these sections were identified, the content associated with these headings was extracted for further analysis. For sections related to catalyst synthesis, a more targeted approach was employed. Regular expressions

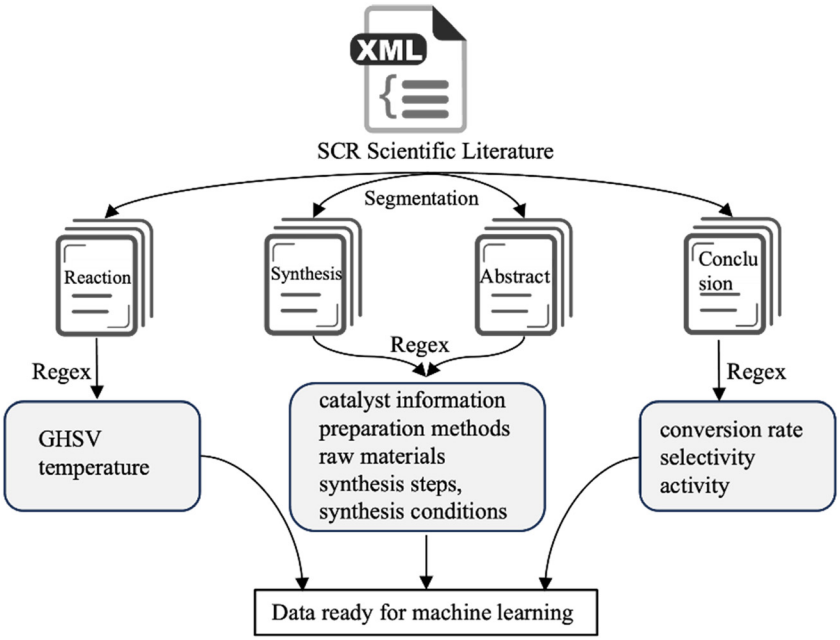


Fig. 1 Literature segmentation and information extraction components.

(Regex) were used to match headings and nodes containing terms such as "synthesis" and "preparation". Regex is provided in ESI† 1. This allows for the accurate extraction of paragraphs directly related to the synthesis process. Through this iterative approach, each article was segmented into five distinct parts: "abstract", "synthesis", "reaction", "conclusion" and "other". The "other" section is discarded, as the focus of the analysis is on the first four sections, where the most relevant information is concentrated. This process is shown in Fig. 1. This method not only improves the precision of data extraction but also optimizes the overall workflow by narrowing down the content to the most critical segments. A regular expression-based approach was then employed to analyze the titles of these articles, narrowing the dataset to 2305 articles specifically describing the catalyst preparation process. Performance data, including conversion rates and selectivity, were subsequently extracted from the abstracts and conclusions of these 2305 articles. Finally, it was found that only 446 articles explicitly reported performance data in their abstracts or conclusions.

In the text mining process, the focus was on extracting the synthesis, reaction and performance data of SCR catalysts. A rule-based approach, utilizing Regex, was employed due to its efficiency, strong specificity, and high accuracy in identifying domain-specific information.

For the extraction of synthesis details, preparation methods, raw materials, synthesis steps, and synthesis conditions were systematically identified from the synthesis sections of the literature. The methodology employed has been thoroughly detailed in our previous work.[29] The extraction process of catalyst preparation methods was carried out in two main stages. First, general catalyst preparation phrases, such as "prepared by", "synthesized using", or "synthesized *via*", were identified using the pattern "(prepared|synthesized) (by|using|*via*)(.*?)(technique|method|using|of|\. |\,)\b" to capture common descriptions of catalyst synthesis methods. These phrases were then processed to extract the specific preparation methods described in the corresponding sentences. Duplicates were removed by converting the list of methods to a set, ensuring only unique descriptions were retained. Second, a more targeted search was performed to identify specific and well-known catalyst preparation techniques, such as "impregnation", "sol–gel", and "chemical vapor deposition", using a predefined list of methods in ESI† 2. This rule-based approach allowed for systematic extraction of synthesis methods from unstructured text. Additionally, the rule-based approach can be easily extended to accommodate various other cases by modifying or adding patterns to the regular expressions, allowing the extraction process to be adapted to different types of catalyst preparation methods that may be described in the literature. Raw materials are extracted from the catalyst preparation paragraphs using the ChemDataExtractor,[30] which employs a chemistry-aware natural language processing pipeline with chemical named entity recognition to identify and extract materials from the text.

For the extraction of operation steps, a commonly used step dictionary was created, and a dictionary-matching approach was employed to identify relevant steps. The dictionary created for extracting operation steps is visible in the data availability. The focus was on identifying key operation steps without considering their sequence, as the primary purpose was to capture interesting actions described in the process. For the extraction of synthetic conditions, the Regex was developed to extract key synthetic conditions, such

as drying, calcination, aging, temperature, and time, from the catalyst preparation paragraphs in the literature. Firstly, sentences related to synthetic conditions were identified in the catalyst preparation paragraphs using specific keywords. Drying sentences were identified by searching for keywords such as "dried" and "drying" within a 50-character (ch) window, capturing relevant information about the drying process. Calcination sentences were extracted by searching for terms like "calcine" and its variants within a 100-ch range. Similarly, aging sentences were identified by locating the keyword "age" within a 20-ch window, allowing for the extraction of relevant aging information. Once these sentences were extracted, temperature and time data were then extracted by focusing on the characteristic patterns of temperature and time mentioned in the sentences with the pattern of "\b\d+(?:\.\d+)?\s*[°°°]?\s*[Cc]\b|\b\d+(?:\.\d+)?\s*[Kk]\b|room temp" and "\d+.?h|overnight|\d{2,3} ?min|\w{1,7}hours". Temperature-related terms such as specific values with units (*e.g.*, °C, K, or "room temp") and time-related terms such as hours, minutes, or phrases like "overnight" were targeted.

For reaction conditions, gas hourly space velocity (GHSV) and reaction temperature as the primary reaction conditions were extracted. Different GHSVs can alter the residence time of reactants on the catalyst surface, affecting conversion efficiency. Similarly, different reaction temperatures can influence the overall reaction performance. The selection of these two parameters was based on their significant impact on SCR reaction performance in practical applications. The process of extracting GHSV involves several key steps. First, the relevant paragraph was split into individual sentences. The keywords "space velocity" and "GHSV" were used to identify relevant sentences that may contain GHSV data. A Regex pattern ((\d{1,4} ?,?\d{3},?\d?\d?\d?)\s*(h−1|h^{-1}|s^{-1}|m/s|mL/(h·g)|mL·g−1·h−1|/h|hr−1)) was applied to these sentences to search for numerical values followed by units commonly associated with GHSV (such as h^{-1}, s^{-1}, mL h^{-1} g^{-1}, and others). For each sentence, the presence of the keywords was checked. If a keyword was found, the pattern was applied to extract the corresponding value and unit. If no matching value was found, "no values" was assigned. For extracting the reaction temperatures (RT), the first step was also to segment the paragraphs into sentences. Then, two different patterns were used to search for temperature-related information in the sentences. The first pattern searched for numerical temperature values followed by units such as °C or K. This pattern "\b(\d+\.?\d*)\s*(−|to)?\s*(\d+\.?\d*)?\s*(°C|K)\b(?!\/[a-zA-Z]+)" can capture both single temperatures (*e.g.*, "300 °C") and ranges (*e.g.*, "100–200 °C"). The second pattern "\b(ambient|room temperature)\b" was used to capture non-numerical temperature references like "ambient" or "room temperature". For each sentence in the paragraph, the script first applies the first pattern to find numeric temperatures with units (°C or K). If no match was found with the first pattern, the script checked for "ambient" or "room temperature" using the second pattern. The matched temperature, if found, was stored in a dictionary along with the DOI. If no temperature was found, "none" was recorded.

The performance data, including conversion rate and selectivity, were specifically extracted from both the abstract and conclusion sections. These performance indicators were carefully integrated with the synthesis data to provide a more complete representation of the catalysts' properties. As shown in Fig. 2, the information on selectivity and conversion rates was automatically extracted from the abstract and conclusion sections of the scientific literature. First, sentences containing the relevant keywords, such as "conversion rate" and "selectivity" were located. Then, using the pattern "\d \d.?%|0\.\d{2,4}", the corresponding values and units are extracted from within 20 characters surrounding the keywords. This structured data served as a robust foundation for subsequent ML tasks, where these parameters will be used to build predictive models.

2.2 Data processing

Structured data were extracted using regular expressions, but errors and incomplete information may arise due to variations in the text. To address this, several strategies have been implemented to handle parsing errors and incomplete information.

Whenever a Regex match fails or no relevant data is found, a default value (*e.g.*, "none" or "no values") was assigned to the corresponding field. This ensures that the extraction process does not terminate abruptly. After the initial extraction, data processing was performed to check the extracted information. The data processing process involved consistency checks, handling of range values, filling missing values, and addressing outliers. Numerical values were

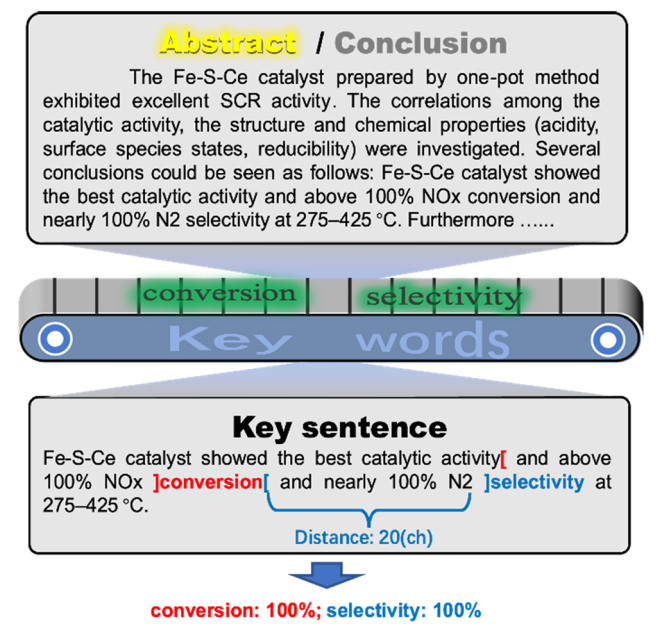


Fig. 2 Conversion rates and selective information extraction methods.

checked to ensure they fall within expected ranges (*e.g.*, ensuring that temperature values are within realistic bounds) and units were checked to ensure they match the expected format. The temperature values were converted to Kelvin (K), time values to hours (h), and space velocity values to h^{-1} to ensure consistency and uniformity across the dataset.

For values presented as ranges, such as "100–200", the mean value of the range was calculated (*e.g.*, the average of 100 and 200). This approach allows the range to be represented by a single value, facilitating further analysis. In cases of incomplete information, discrete values were encoded as 0, while missing continuous features were filled using mean imputation. Outliers were detected using boxplot analysis, which identifies values that deviate significantly from the rest of the dataset. Once outliers were detected, they were replaced with the mean of the relevant feature to prevent them from affecting the modeling.

2.3 Feature engineering

Machine learning models primarily operate on numerical data and cannot directly process unstructured textual information. To address this, we first encoded the synthesis information of catalysts into a format that is comprehensible to ML models. For example, embedding models used in natural language processing can convert textual data into numerical vectors. However, these embeddings typically exhibit high dimensionality and are designed for encoding lengthy texts rather than words.

Features were extracted and encoded from five types of information: catalyst reaction conditions, synthesis methods, raw materials, synthesis steps, and synthesis conditions. One-hot encoding was applied to categorical variables such as synthesis methods and synthesis steps, as these represent finite sets. Based on the catalyst synthesis information extracted from 446 relevant articles, 15 common synthesis methods, including "solid", "sol–gel", "impregnation", and "hydrothermal", were identified. Additionally, 14 typical operational steps, such as "dry", "calcined", and "aged", were encoded. Numerical features, such as operating temperature and time, and reaction conditions can be directly input into the model. For the raw materials, two methods were explored: one-hot encoding and ASCII (American Standard Code for Information Interchange) encoding. One-hot encoding was used to represent each raw material as a separate vector, ensuring that each material was treated independently, without implying any ordinal relationship. ASCII encoding converts each character of the raw material into its corresponding ASCII value, generating numerical representations. The encoding process using ASCII is illustrated in Fig. 3. The names of the raw materials were converted to uppercase, and each letter was encoded individually to form the corresponding ASCII codes. A uniform length of 30 characters for all encoded raw materials was established to reduce the search space when exploring new materials. This length of 30 also captures the majority of raw materials effectively.

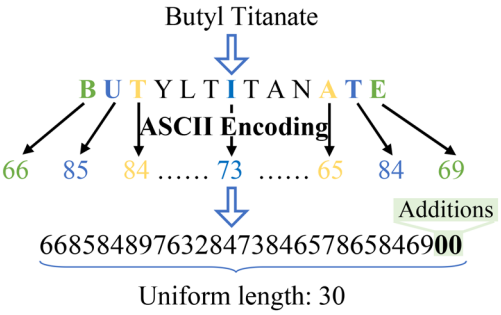


Fig. 3 ASCII encoding method for raw materials in catalyst synthesis processes.

Finally, these encoded components were combined to form the features of the entire dataset. Two machine learning datasets were established: one focused on conversion rates, consisting of 446 samples, and the other on selectivity, including 100 samples.

2.4 Machine learning models

The extreme gradient boosting (XGBoost) regressor from the xgboost library (xgboost.XGBRegressor()) and the random forest (RF) regressor from scikit-learn[31] (sklearn.ensemble.RandomForestRegressor) were employed as the primary machine learning models. The XGBoost regressor (XGBR) is a powerful gradient boosting that excels in regression tasks, particularly with structured data. It is known for its efficiency, scalability, and ability to handle large datasets, making it suitable for complex problems with non-linear relationships. Its built-in regularization techniques help prevent overfitting, further enhancing its performance. The random forest regressor is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of these trees for regression tasks. This model is robust to overfitting and provides a stable performance across various datasets. Its ability to capture interactions and non-linearities among features contributes to its effectiveness in predictive modeling.

Every machine learning algorithm has their individual parameters and hyperparameters. Parameters are those components of the model that are learned during the training process, and these could not be tuned manually. However, hyperparameters are those components which can be tuned before the training process, and they increase the prediction accuracy of a machine learning model. In this work, hyperparameter tuning was performed on the selected algorithms for both datasets of the SCR catalyst conversion rate and selectivity, with the assistance of the GridSearchCV method. All the details of the hyperparameters for each data set can be found in the ESI† Table S1.

2.5 Evaluation

When evaluating the performance of regression models (models that predict continuous outcomes such as catalyst performance based on synthesis parameters), several key metrics are used to quantify prediction accuracy. The most common metrics, including mean squared error (MSE) and *R*-squared (R^2), were used. Here's how each metric is calculated.

The *R*-squared (R^2) metric, also known as the coefficient of determination, measures the proportion of variance in the dependent variable (y) that is predictable from the independent variables (X). It gives an overall indication of how well the model's predictions fit the actual data:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where y_i is the actual value for the i-th data point, $\hat{y}_i$ is the predicted value for the i-th data point, $\bar{y}$ is the mean of the actual values, and n is the number of observations. R^2 ranges from 0 to 1. A value of 1 indicates that the model perfectly explains all the variability in the data, while a value of 0 indicates that the model explains none of the variability (the model predictions are no better than simply predicting the mean of the data). A negative R^2 can occur if the model performs worse than the mean prediction.

Mean squared error (MSE) is a commonly used metric to evaluate the performance of regression models. It measures the average squared difference between the actual values and the predicted values. The formula for MSE is:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

3. Results and discussion

3.1 Feature analysis

Two datasets were established and evaluated for ML models based on the methods described, specifically for predicting the conversion rate and selectivity of SCR catalysts. Both datasets shared the same features. Normalization was applied to the features to ensure that they are on a comparable scale. Additionally, an analysis of the correlation between features was conducted to identify potential relationships that could influence model predictions.

The heatmap, shown in Fig. 4, represents the correlation matrix of the features used in our ML models. The display step size for the features on the axes in the figure is two. Correlations are measured using Pearson's correlation coefficient, which ranges from −1 to 1. A coefficient of 1 indicates a perfect positive correlation, 0 indicates no correlation, and −1 indicates a perfect negative correlation. Higher correlations are observed among raw material features, which are expected due to inherent relationships between the materials. These correlations are natural, as raw materials often have compositional similarities. The majority of the features exhibit correlations close to 0, suggesting low interdependence. This is advantageous for model training, as it minimizes the risk of multicollinearity.

3.2 Establishment and validation of models

Model training is essentially the process of finding suitable parameters for the mathematical formulation of the

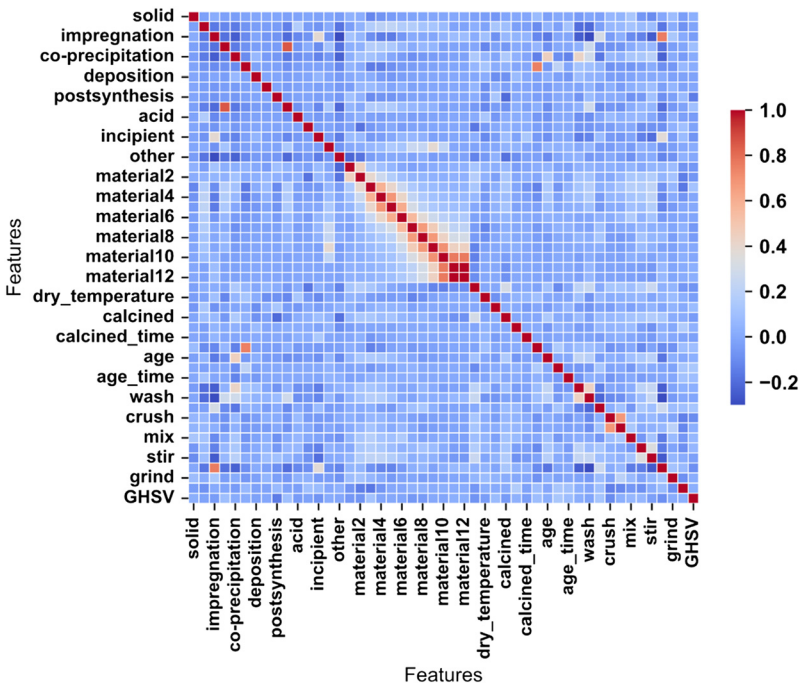


Fig. 4 Correlation matrix of all features.

corresponding model so that the gap between the result of the model's evaluation of the specified dataset and the real situation is minimized. The entire dataset was divided into 80% for the training set and 20% for the test set. Grid search and cross-validation were performed on the 80% training set to obtain the model with the best performance (highest R^2) and its hyperparameters. The optimal hyperparameters of the model are shown in Table S1.† The best-performing model can be employed to predict the conversion and selectivity of SCR catalysts.

Fig. 5 illustrates the predictive performance of random forest (RF) and XGBoost regression (XGBR) models for conversion rate prediction using two different encoding methods for raw materials: one-hot encoding in Fig. 5(a) and (b) and ASCII encoding in Fig. 5(c) and (d). The features used for modeling include preparation methods, raw materials, synthesis steps, synthesis conditions, and reaction conditions. The scatter points illustrate the predicted *versus* actual values for the train set (cyan) and test set (magenta). The RF model, with one-hot encoding, achieves strong performance, with the test R^2 of 0.757. The RF model demonstrates strong and consistent performance across both encoding methods, with slightly better results when using one-hot encoding. In contrast, the XGBR model shows moderate performance with one-hot encoding, as seen in Fig. 5(b), but achieves improved generalization and more balanced predictions between the train and test sets with ASCII encoding, as shown in Fig. 5(d). These results highlight the impact of encoding methods on model performance.

Fig. 6 compares the predictive performance of RF models for selectivity prediction using two different encoding methods for raw materials: one-hot encoding in Fig. 6(a) and ASCII encoding in Fig. 6(b). The features used in both models include preparation methods, raw materials, synthesis steps, synthesis conditions, and reaction conditions. Using one-hot encoding, the RF model in Fig. 6(a) achieves a train R^2 of 0.845, demonstrating strong predictive performance. The predicted values for both train and test sets align closely with the actual values. Using ASCII encoding, the RF model performs better,

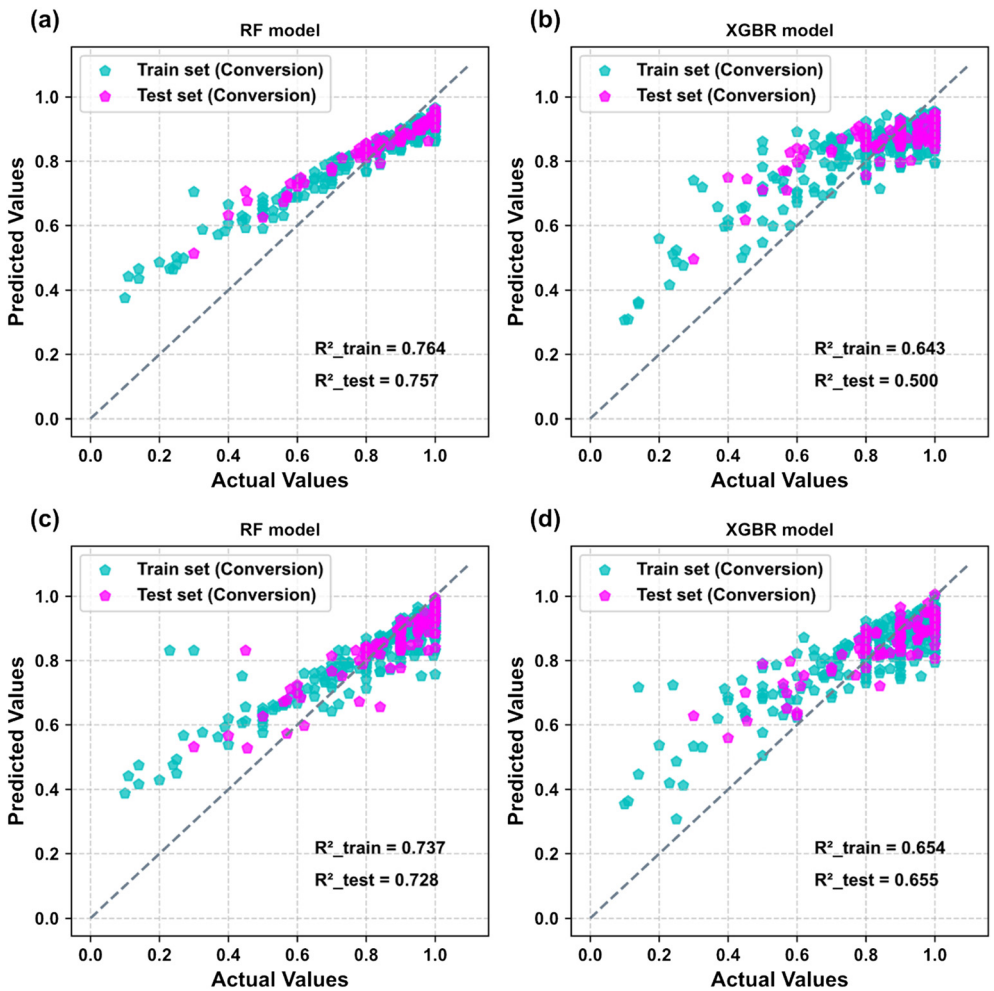


Fig. 5 Performance comparison of random forest (RF) and XGBoost regression (XGBR) models using different encoding methods for raw material features. (a) and (b) show the results when raw materials are encoded using the one-hot encoding method, while (c) and (d) show the results when raw materials are encoded using the ASCII encoding method. The remaining features are consistent.

achieving a train R^2 of 0.879 and a test R^2 of 0.865. This suggests improved generalization and higher accuracy compared to one-hot encoding. Both encoding methods exhibit high predictive performance, but ASCII encoding offers a slight advantage, likely due to its more compact representation of raw materials, which may enhance the model's ability to generalize. These results underscore the effectiveness of RF models in predicting selectivity and the impact of encoding methods on model performance. A detailed assessment is shown in Tables S2–S5.†

3.3 Model interpretability

To gain a better understanding of the importance of input features on the specific direction of model decisions, a feature density scatter plot in SHAP (SHapley Additive exPlanations)[32] was employed as a holistic approach to interpretation. Fig. 7 presents the SHAP summary plots for random forest models predicting conversion and selectivity using two different encoding methods for raw material features. Fig. 7(a) presents the model predicting conversion with one-hot encoded raw materials, while Fig. 7(b) shows the prediction of conversion with ASCII-encoded raw materials. Fig. 7(c) illustrates the model predicting selectivity with one-hot encoding, and Fig. 7(d) shows the prediction of selectivity with ASCII encoding. This scatter plot sorted the Shapley values of each feature into their corresponding position coordinates. As shown in Fig. 7, the *y*-axis represents the importance of the model's predictive features, while the *x*-axis indicates their effect on model predictions (red points signify large Shapley values, and blue points indicate small Shapley values). The Shapley values were combined with sample point colors to investigate the relationship between feature variation and decision direction. Shapley values greater than 0 are interpreted as having a positive impact, whereas those less than 0 indicate a negative impact.

For the prediction of conversion rates with one-hot encoding, key features contributing to conversion prediction include calcination temperature, GHSV, impregnation, and reaction temperature (RT). These features exhibit both positive and negative impacts on the model's predictions, as indicated by their SHAP values. For both calcination temperature and GHSV, low values have a positive effect. Using ASCII encoding, the materials dominate the feature importance, along with GHSV and other physical processes (*e.g.*, crushes and sieves). The impact of the material encoding is more pronounced compared to the one-hot encoded model.

For the prediction of conversion rates with one-hot encoding, the most important features for selectivity prediction include GHSV, calcination temperature, BEA zeolite, NH_4VO_3, and reaction temperature. Using ASCII encoding, the material features (*e.g.*, material 2, material 3, material 4, *etc.*) again dominate, followed by calcination-related features like calcination temperature and impregnation. The materials encoded as ASCII values show a strong impact on selectivity prediction.

One-hot encoding provides more emphasis on physical and chemical process-related features (*e.g.*, calcination temperature, GHSV) across both conversion and selectivity predictions. ASCII encoding shifts the importance towards material features, with raw material encodings playing a dominant role. These results highlight the influence of different feature encoding methods on the interpretability and focus of random forest models when predicting conversion and selectivity.

3.4 Spatial diversity

The spatial distributions reflect the reasonability of the data, which influences the model construction process. In the spatial diversity, the *t*-SNE analysis employed only

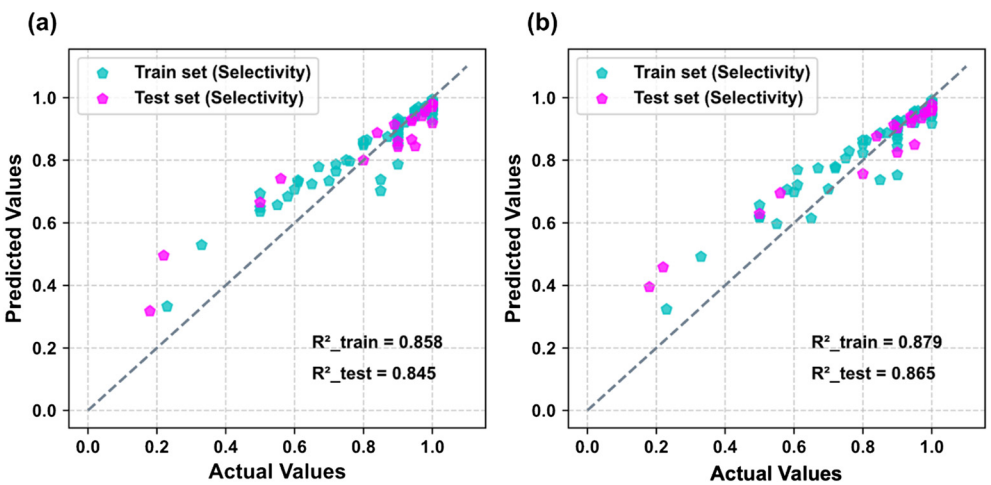


Fig. 6 Performance comparison of RF models using different encoding methods for raw material features. (a) shows the results when raw materials are encoded using the one-hot encoding method, while (b) shows the results when raw materials are encoded using the ASCII encoding method. The remaining features are consistent.

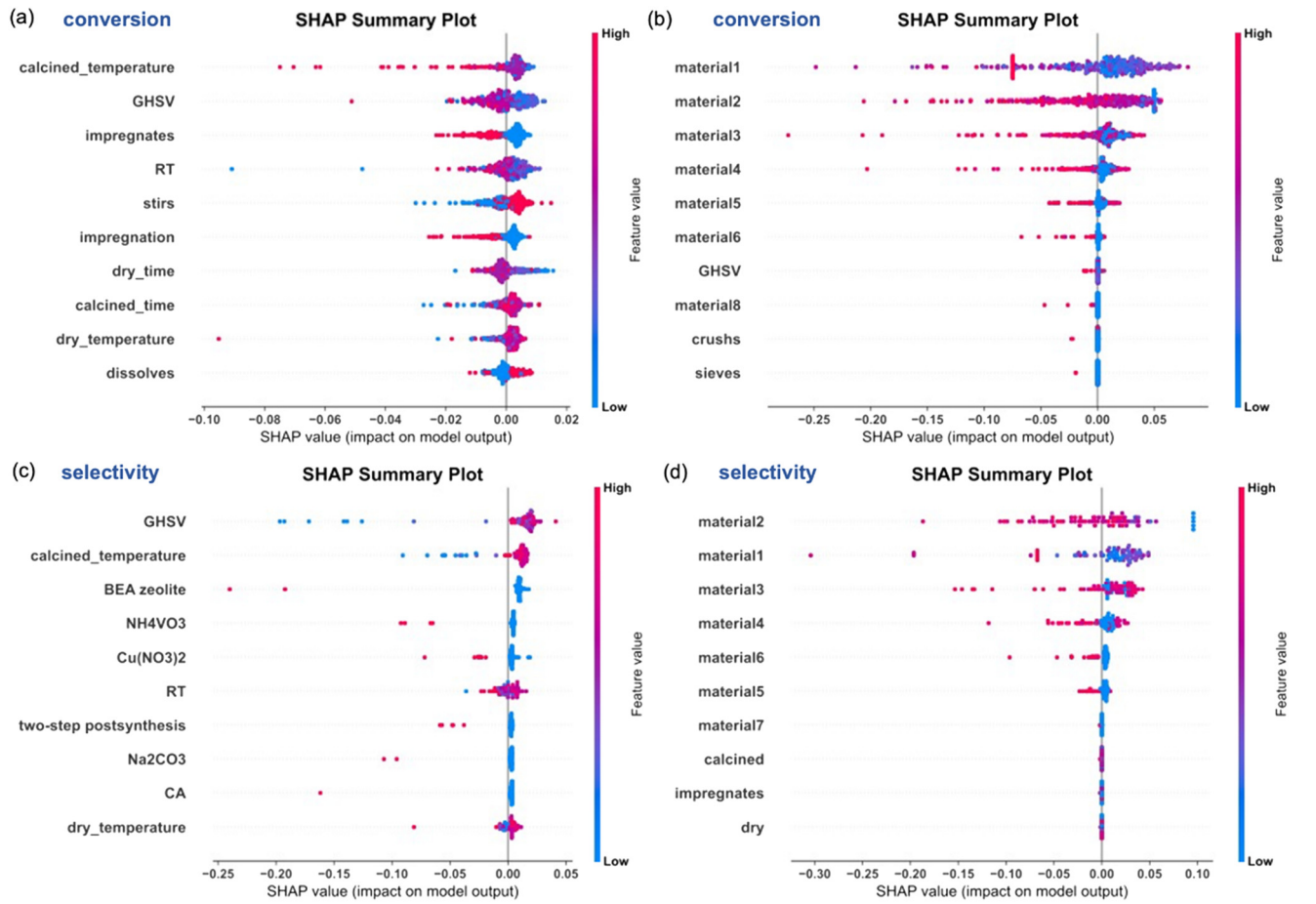


Fig. 7 SHAP summary plots for RF models predicting conversion and selectivity using two different encoding methods for raw material features. (a) presents the model predicting conversion with one-hot encoded raw materials, while (b) shows the prediction of conversion with ASCII-encoded raw materials. (c) illustrates the model predicting selectivity with one-hot encoding, and (d) shows the prediction of selectivity with ASCII encoding.

machine learning inputs, such as materials and synthesis features, to evaluate the distribution and diversity of the features. Machine learning target variables, including conversion and selectivity parameters, were excluded from the analysis. We evaluated the rationality of the data by encoding synthesis information of catalysts for the

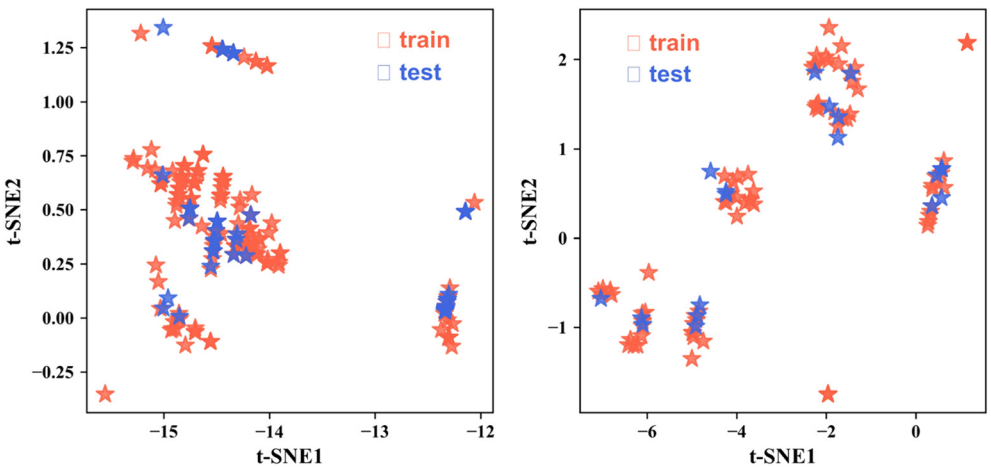


Fig. 8 Diversity distribution of modeled data. From left to right, the two-dimensional spatial distribution of the SCR catalyst conversion and selectivity data sets, respectively.

training and test sets. The *t*-distributed stochastic neighbor embedding (*t*-SNE) was applied to visually represent their chemical spatial distribution. The *t*-SNE[33] is a tool to visualize high-dimensional data by giving each data point a location in a two or three-dimensional map. As displayed in Fig. 8, the *t*-SNE diagram clearly demonstrates the wide chemical spatial distribution of the training and test data. The quality of the data is an important issue to consider before constructing machine learning models. Therefore, the high diversity in the training and test sets proves that our data have excellent robustness. Our analysis indicates that the features used to build this model were reasonable and differed in their chemical structures.

3.5 Synthesis route prediction

First, based on the machine learning training dataset, a designable space for material synthesis information was established. The scope of synthesis methods includes techniques such as [['solid', 'sol–gel', 'impregnation', 'hydrothermal', 'co-precipitation', 'ion-exchange', 'deposition', 'template', 'two-step postsynthesis', 'thermal', 'acid', 'calcination', 'incipient', 'one-step', 'other']]. The search space for synthesis steps and conditions includes the following parameters: ['dry', 'dry_temperature', 'dry_time', 'calcined', 'calcined_temperature', 'calcined_time', 'exchange', 'age', 'age_temperature', 'age_time', 'filter', 'wash', 'evaporative', 'crush', 'sieve', 'mix', 'dissolve', 'stir', 'impregnate', 'grind']. For synthesis conditions, the ranges for temperature and time are set between the minimum and maximum values found in the literature used to construct the dataset. Specifically, the upper and lower temperature limits for the dry process are 473 K and 298 K, respectively, with a maximum duration of 48 hours. For the calcination process, the temperature limits are set between 1298 K and 523 K, with a maximum duration of 25 hours. For the age process, the temperature range is set between 1023 K and 298 K, with a maximum duration of 120 hours. The designable space for raw materials includes all raw materials present in the dataset.

A random search method is employed to explore combinations of synthesis methods, raw materials, synthesis steps, and conditions within this designable space. These combinations are then input into the pre-trained machine learning models to predict catalyst performance, with the key performance indicators being conversion rate and selectivity. The trained models are used to predict the conversion rates and selectivity for 100 000 different combinations generated within the designable space. From these predictions, the top-performing combinations, according to both conversion and selectivity, are selected as the optimal synthetic routes for SCR catalyst development. Synthetic information in the space was searched and recommended by the machine learning model, presented in ESI† 3. Fig. 9 illustrates one of the potential synthetic messages recommended by machine learning models.

4. Conclusions

This study illustrates the application of machine learning and text mining techniques to predict the performance of SCR catalysts. By extracting relevant synthesis and

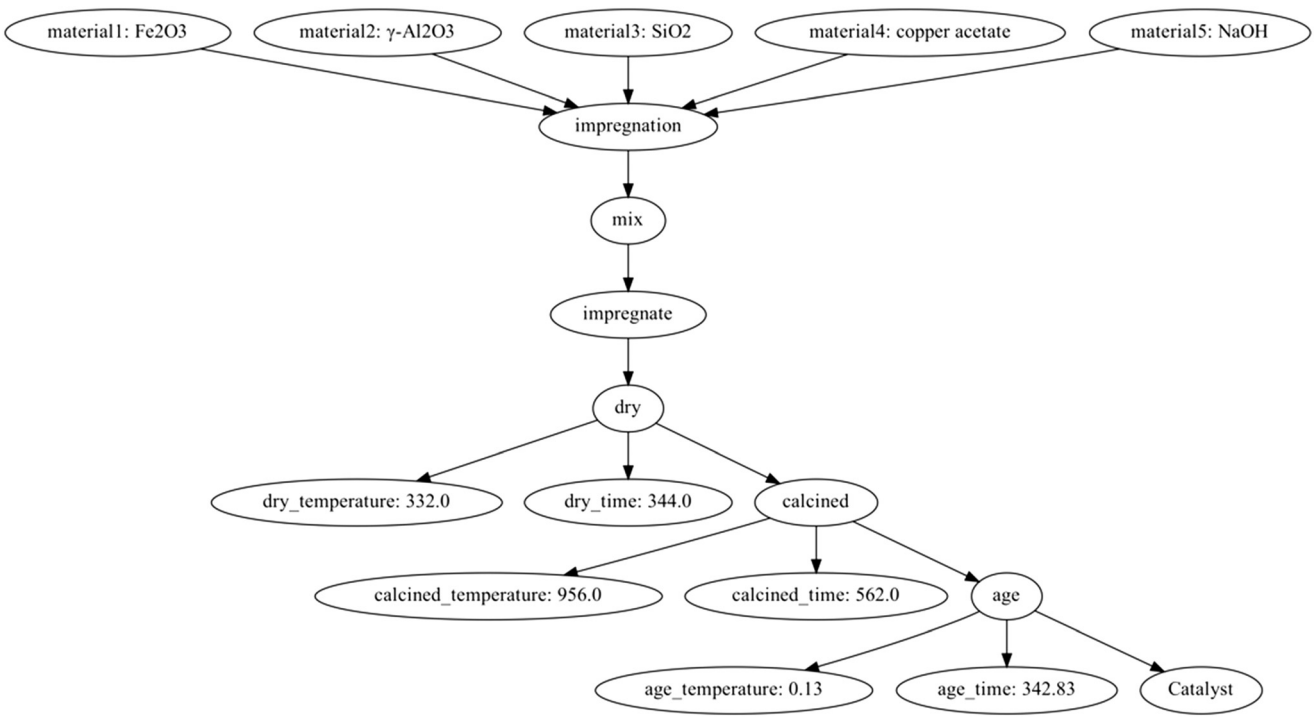


Fig. 9 One of the potential synthetic messages recommended by machine learning models.

performance data from the scientific literature and constructing a machine learning dataset through feature engineering, models such as XGBR and random forest were employed to predict catalyst selectivity and conversion rates. Both models demonstrated strong predictive capabilities, identifying the key factors influencing catalyst performance. The results underscore the potential of machine learning to enhance the efficiency of catalyst development by providing reliable performance predictions. This approach reduces the reliance on traditional trial-and-error methods and offers a data-driven framework for advancing catalyst research and design, providing a scalable method for designing more efficient catalysts.

Data availability

The complete list of DOIs for articles, and the code used for data processing and analysis have been uploaded to our repository, available at **https://github.com/Shaoruisun/ML_TM_SCR_Synthesis/**.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the National Key Research & Development Program of China (grant no. 2021YFA1201000) and the National Science Foundation of China (62104101).

References

1 A. Rasool and M. A. Dar, *Catal. Sci. Technol.*, 2024, **14**, 5687–5698.
2 Z. Deng, M. A. Padalino, J. E. L. Jan, S. Park, M. W. Danneman and J. N. Johnston, *J. Am. Chem. Soc.*, 2024, **146**, 1269–1275.
3 Y. Zhang, B. Guan, C. Zheng, J. Zhou, T. Su, J. Guo, J. Chen, Y. Chen, J. Zhang, H. Dang, Y. Yuan, C. Xu, B. Xu, W. Zeng, Y. He, Z. Wei and Z. Huang, *J. Cleaner Prod.*, 2024, **434**, 139920.
4 Y. Dong, M. Ran, X. Zhang, S. Lin, H. Zhao, Y. Yang, S. Liu, C. Zheng and X. Gao, *ACS ES&T Eng.*, 2024, **4**, 1312–1320.
5 S. Liu, J. Gao, W. Xu, Y. Ji, T. Zhu, G. Xu, Z. Zhong and F. Su, *Chem. Eng. J.*, 2024, **486**, 150285.
6 C. Zhang, G. Xu, Y. Zhang, C. Chang, M. Jiang, L. Ruan, M. Xiao, Z. Yan, Y. Yu and H. He, *Appl. Catal., B*, 2024, **348**, 123820.
7 H. Li, L. Schill, R. Fehrmann and A. Riisager, *Inorg. Chem. Front.*, 2023, **10**, 727–755.
8 A. Trunschke, *Catal. Sci. Technol.*, 2022, **12**, 3650–3669.
9 S. Li, Y. Zhang, Y. Hu, B. Wang, S. Sun, X. Yang and H. He, *J. Mater.*, 2021, **7**, 1029–1038.
10 Z. Fang, S. Li, Y. Zhang, Y. Wang, K. Meng, C. Huang and S. Sun, *J. Phys. Chem. Lett.*, 2024, **15**, 281–289.
11 M. Xing, Y. Zhang, S. Li, H. He and S. Sun, *J. Phys. Chem. C*, 2022, **126**, 17025–17035.
12 H. Mai, T. C. Le, D. Chen, D. A. Winkler and R. A. Caruso, *Chem. Rev.*, 2022, **122**, 13478–13515.
13 G. Tanimu, J. Olawale Ajadi, Y. Yahaya, H. Alasiri and N. A. Adegoke, *ChemCatChem*, 2023, **15**, e202300598.
14 M. Wang and H. Zhu, *ACS Catal.*, 2021, **11**, 3930–3937.
15 D. A. Rosser, B. R. Farris and K. C. Leonard, *Digital Discovery*, 2024, **3**, 667–673.
16 W. Guo, A. Shafizadeh, H. Shahbeik, S. Rafiee, S. Motamedi, S. A. Ghafarian Nia, M. H. Nadian, F. Li, J. Pan, M. Tabatabaei and M. Aghbashlo, *Journal of Energy Storage*, 2024, **89**, 111688.
17 H. Sun, Y. Li, L. Gao, M. Chang, X. Jin, B. Li, Q. Xu, W. Liu, M. Zhou and X. Sun, *J. Energy Chem.*, 2023, **81**, 349–357.
18 M. Suvarna, P. Preikschas and J. Pérez-Ramírez, *ACS Catal.*, 2023, **81**, 349–357.
19 K. Meng, C. Huang, Y. Wang, Y. Zhang, S. Li, Z. Fang, H. Wang, S. Wei and S. Sun, *J. Chem. Inf. Model.*, 2023, **63**, 6043–6052.
20 C. Liang, B. Wang, S. Hao, G. Chen, P.-A. Heng and X. Zou, *Adv. Funct. Mater.*, 2024, **34**, 2404392.
21 K. P. Treder, C. Huang, C. G. Bell, T. J. A. Slater, M. E. Schuster, D. Özkaya, J. S. Kim and A. I. Kirkland, *npj Comput. Mater.*, 2023, **9**, 1–12.
22 Y. Chen, J. Feng, X. Wang, C. Zhang, D. Ke, H. Zhu, S. Wang, H. Suo and C. Liu, *Environ. Sci. Technol.*, 2023, **57**, 18080–18090.
23 C. Martínez-Alonso, V. Vassilev-Galindo, B. M. Comer, F. Abild-Pedersen and K. T. Winther, *Catal. Sci. Technol.*, 2024, **14**, 3784–3799.
24 Y. Liu, Z. Liang, J. Huang, B. Zhong, X. Yang and T. Wang, *Catal. Sci. Technol.*, 2023, **13**, 6281–6290.
25 D. Roy, S. C. Mandal and B. Pathak, *J. Phys. Chem. Lett.*, 2022, **13**, 5991–6002.
26 S. Kim, Y. Park, S. Yoo, O. Lim and B. F. Samosir, *Sustainability*, 2023, **15**, 7077.
27 S. A. Okeleye, A. Thiruvengadam, M. G. Perhinschi and D. Carder, *Energy*, 2024, **290**, 130117.
28 S. Bae, H. Lee, J. Shin, H. S. Kim, Y. Kim, D. H. Kim and J. M. Lee, *Chem. Mater.*, 2022, **34**, 7761–7773.
29 S. Li, Y. Zhang, Z. Fang, K. Meng, R. Tian, H. He and S. Sun, *J. Chem. Inf. Model.*, 2023, **63**, 6249–6260.
30 M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
31 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos and D. Cournapeau, *Machine Learning in Python*.
32 S. M. Lundberg and S.-I. Lee, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.
33 L. van der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.