Green Chemistry

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Green Chem.*, 2025, **27**, 2046

Received 31st October 2024,
Accepted 16th January 2025

DOI: 10.1039/d4gc05515b

rsc.li/greenchem

# Machine learning prediction of physical properties of lignin derived porous carbon *via* catalytic pyrolysis†

Zihao Xie, Yue Cao* and Zhicheng Luo *

Lignin-derived porous carbon produced through catalytic pyrolysis is crucial for energy storage, adsorption, and catalysis. However, predicting specific surface area (SSA), total pore volume (TPV), and microporosity (MP) remains challenging due to the variability in lignin properties, chemical activators, and pyrolysis conditions, compounded by limited data availability. In this study, we applied a hybrid machine learning framework incorporating a pre-trained interpolation model and a final regressor to impute missing features, improving prediction accuracy and generalizability. This approach yielded high predictive accuracy with $R^2$ values of 0.82 (SSA), 0.86 (TPV), and 0.81 (MP) on a dataset of 112 samples, encompassing variations across six chemical activators (KOH, $ZnCl_2$, $H_3PO_4$, $K_2CO_3$, NaOH, and $Na_2CO_3$). Feature importance analysis highlighted the significant influence of KOH on SSA and TPV, and $H_3PO_4$ on MP. This research provides a framework to precisely tailor the pore structure of lignin-derived porous carbon *via* catalytic pyrolysis, enabling advancements in applications across diverse fields.

**Green foundation**

1. This study develops a machine learning model to predict the physical properties of lignin-derived porous carbon (LDPC) produced by catalytic pyrolysis, reducing the need for resource-heavy experimental trials and promoting a more sustainable approach to material design.
2. The model achieves high predictive accuracy ($R^2$ values of 0.82 for specific surface area, 0.86 for total pore volume, and 0.81 for microporosity), significantly minimizing experimental waste, energy consumption, and the use of hazardous chemicals in the synthesis of LDPC.
3. Future research will integrate additional structural and process variables into the model to further optimize the catalytic pyrolysis of lignin, enabling even greater reductions in energy usage, material waste, and improving the overall sustainability of LDPC production.

## Introduction

The global energy crisis and climate change concerns underscore the urgency of transitioning to renewable resources. Lignin, abundant in biomass, is increasingly recognized as a promising precursor for porous carbon due to its high carbon content and thermal stability.[1–5] Lignin-derived porous carbons (LDPCs) find applications in diverse fields including energy storage and environmental remediation.[6–8] The effectiveness of LDPC in these applications hinges on their pore structure, driving extensive experimental efforts to tailor the specific surface areas (SSA), total pore volumes (TPV), and microporosity (MP).[9,10]

Various synthesis strategies, such as templating and activation methods, are employed to manipulate LDPC pore structures, with catalytic pyrolysis being particularly prevalent.[11,12] This method leverages chemical interactions between activators and lignin during carbonization to induce pore formation. Different chemical activators exhibit unique etching mechanisms, making catalytic pyrolysis a promising avenue for tailoring LDPC pore structures.[13,14] However, quantitatively linking synthesis strategies to pore structures remains challenging without rigorous experimental validation (Fig. 1A).

While experimental methods dominate pore structure characterization, non-experimental approaches such as machine learning offer a promising alternative. Machine learn-

*MOE Key laboratory of Energy Thermal Conversion & Control, School of Energy and Environment, Southeast University, Nanjing 210096, China.*
*E-mail: zluo@seu.edu.cn*

† Electronic supplementary information (ESI) available. See DOI: **https://doi.org/10.1039/d4gc05515b**

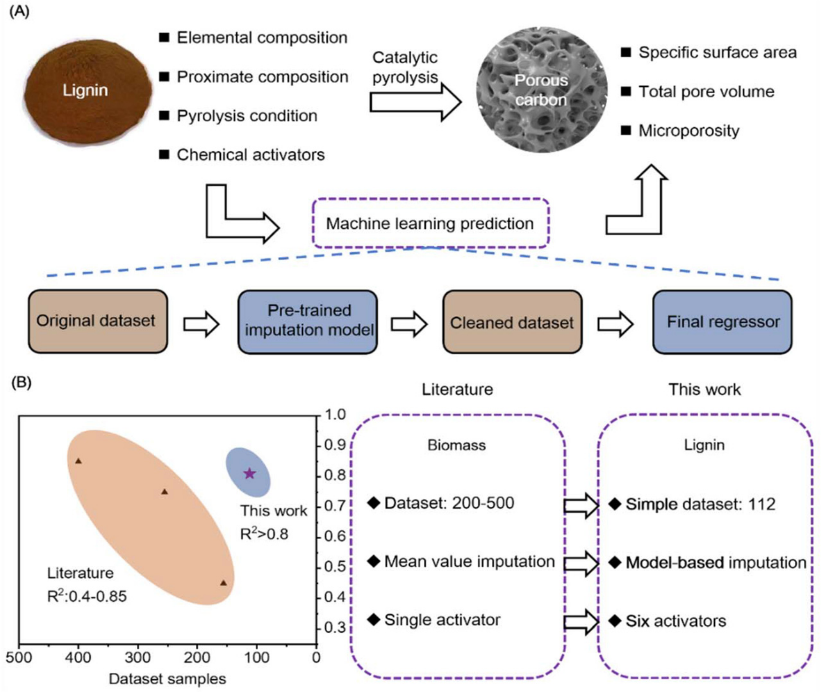

**Fig. 1** Outline of this work. (A) Machine learning process. (B) Comparisons with other literature.

ing has shown success in predicting biochar properties using models like support vector machines, random forests, and gradient boosting regression.[15–23] These models utilize input features such as biomass composition, pyrolysis conditions, and different chemical activators to predict carbon yields and surface areas with high accuracy using samples more than 200. By using biomass components and pyrolysis conditions, the yield and specific surface area of biochar prepared by direct pyrolysis were predicted by Leng *et al.* and Zhu *et al.* Using over 200 samples, they obtained $R^2$ above 0.9 in a relatively simple direct pyrolysis process.[24,25] Zou *et al.* predicted the preparation process of biochar using one-step and two-step activation methods. They included six activators in the one-step activation method and obtained $R^2$ above 0.7 under conditions of 216 samples and 15 features.[26] Wang *et al.* used a single type of activator as input in the prediction process. However, due to the influence of a large number of missing values, they needed to collect over 200 sample points on each activator in order for the model to exhibit an $R^2$ of 0.9 or higher. There is a serious dependence on samples in the prediction of biochar.[27] Despite the potential of machine learning in biochar studies, its application to predict LDPC pore structures *via* catalytic pyrolysis remains underexplored due to limited sample availability.

This study addresses the challenge of limited samples and quality in predictive modeling for LDPC, as shown in Fig. 1B. By employing a hybrid machine learning framework incorporating a pre-trained interpolation model and a final regressor, the study enhances prediction accuracy using a dataset of 112 samples across six chemical activators. The selected Gradient Boosting Regression model with Random Forest interpolation demonstrates robust performance, achieving $R^2$ values of 0.82 for SSA, 0.86 for TPV, and 0.81 for MP. Moreover, interpretability analysis reveals the pivotal role of activators like KOH in influencing SSA and TPV, and $H_3PO_4$ in affecting MP. These findings contribute a viable approach to precisely regulating LDPC pore structures *via* catalytic pyrolysis, thereby advancing their applications in various field.

## Experimental methods

### Data collection

To collect the dataset, we conducted a comprehensive review of the literature using keywords such as lignin-derived carbon (LDPC), catalytic pyrolysis, and chemical activation. We searched well-known databases, including Web of Science and Google Scholar, to gather samples for machine learning. The search covered publications from the past ten years to ensure the inclusion of the most recent and relevant data (Table S1†). Experimental measurement errors in the studies may influence the predictions made by the machine learning model. However, since the available studies do not provide error bars or detailed information on measurement uncertainties, these errors were not considered. The collected data included input features such as lignin characteristics, chemical activators, and pyrolysis condition (Table S2†). Specifically, the lignin characteristics included proximate composition and elemental composition, namely volatile matter (VM), ash (Ash), fixed carbon (FC), and elements C, H, O, N and S. Linkage ratios are important structural features that are related to the elemental composition and properties of LDPC. However, due to the limited structural data available in the literature, these features were not included in this study. The chemical activators encompassed the type of activating agent (Agent) and the impregnation ratio (A/S). The pyrolysis conditions included pyrolysis temperature (Temp), retention time (RT), and heating rate (HR).

To ensure consistency in the dataset units, VM, Ash, FC, C, H, O, N, and S were all measured in %. SSA was measured in $m^2\ g^{-1}$, TPV in $cm^3\ g^{-1}$, Temp in °C, RT in hours, and HR in $°C\ min^{-1}$. Agent includes the following six type: KOH, $ZnCl_2$, $H_3PO_4$, $K_2CO_3$, NaOH, $Na_2CO_3$. In total, we collected 112 samples (*i.e.*, 112 porous carbon) from 32 published papers to predict the specific surface area (SSA), total pore volume (TPV), and microporosity (MP) of LDPC prepared *via* catalytic pyrolysis (Table S2†).

### Dataset normalization and analysis

The dataset was normalized before training the models to ensure the appropriate scale of numerical values of input features. The normalization of parameters was performed according to eqn (1). Additionally, the type of chemical activator significantly influenced prediction outcomes. To enhance model performance and generalizability, we employed one-hot encoding for the types of activating agents. Pearson correlation coefficient (PCC) was used to estimate the initial correlation between input features and the prediction target, calculated using eqn (2) to determine the correlation coefficient $\rho_{xy}$ between two features.[28]

$$x_i^* = \frac{x_i - \mu}{s} \tag{1}$$

where $x_i$ is the value of variable $i$; $x_i^*$ is the normalized value of origin $x_i$; $\mu$ is the mean value of $x_i$; and $s$ represents the standard deviation of $x_i$.

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - y)^2}} \tag{2}$$

where $\rho_{xy}$ is the value of PCC between any two variables; $\bar{x}$ and are the mean of one variable $x$ and the other variable $y$. The range of $\rho_{xy}$ is from −1 to 1, where 0 means no linear correlation, and a negative or positive value means negative or positive correlation.

### Hybrid machine learning framework

The collected samples' features are not always complete, in this work, features such as VM, Ash, FC, SSA and MP have many missing values. Due to the relationship between features and prediction labels, when there are missing values in features, the pre-trained interpolation model can be established to interpolate the missing values through other features. In this article, samples without missing features were divided into a new dataset, and the pre-trained interpolation model was trained in this dataset using the missing features as the prediction target. The pre-trained interpolation model was then used to fill other missing samples' features in the original dataset.[29] After the interpolation, a cleaned dataset can be obtained to fitted by the final regressor, which was used to predict the SSA, TPV, and MP of the LDPC.

We employed two widely recognized machine learning algorithms to build our hybrid machine learning framework, Random Forest (RF) for the pre-trained interpolation model and Gradient Boosting Regression (GBR) for the final regressor. These algorithms are favored for their effectiveness in handling nonlinear problems and their ability to handle imbalanced datasets with minimal hyperparameter tuning. RF model utilizes bagging theory to construct multiple decision trees independently. Each tree is trained on a random subset of features at each node, and the final prediction is the average of predictions from all trees in the forest.[30] GBR model, on the other hand, follows boosting theory by sequentially building decision trees. Each subsequent tree fits the residuals of the previous tree, placing higher emphasis on instances where earlier predictions were inaccurate. The final prediction is a weighted sum of predictions from each tree.[26]

The dataset was split into training (80%) and testing (20%) sets to fit and test the hybrid machine learning framework. Hyperparameters such as the number of decision trees and maximum tree depth were tuned within ranges of 70–200 and 5–15, respectively. During hyperparameter tuning, a five-fold cross-validation strategy was used to train both models. Specifically, the training dataset was randomly divided into five parts, with four-fifths used to train the model and one-fifth used for performance validation. This process was iterated five times to validate all five parts of the dataset. The average performance of the five validations was used to select the optimal hyperparameters. The root mean square error (RMSE) was introduced to identify the best hyperparameters, with the model yielding the smallest RMSE corresponding to the optimal hyperparameters.[31]

After obtaining the optimal hyperparameters, the framework was retrained using the training dataset (80% of the collected data). The correlation coefficient ($R^2$) and RMSE were used to evaluate the predictive performance of the optimal RF and GBR models, representing the degree of fit and the deviation between actual and predicted values (eqn (3) and (4)). A larger $R^2$ and smaller RMSE indicate more accurate predictions and better performance of the trained model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

where $y_i$ is the target value, $\hat{y}_i$ is the output value, and $\bar{y}$ is the mean value of all target value.

### Model interpretation

Despite their ability to capture complex, nonlinear relationships and achieve high predictive accuracy, interpreting machine learning models remains challenging. We employed several techniques to enhance the interpretability of our final regressor.

**Permutation importance analysis.** This method aids in understanding the contribution of each input feature to the model's predictions. It involves systematically shuffling the values of individual features, retraining the model, and then measuring the change in predictive performance. A higher permutation importance score indicates that the feature has a more significant impact on the target variable, while scores close to zero suggest minimal influence.

**Partial dependence plots.** To further elucidate the relationship between input features and the target variable, we utilized Partial dependence plots. These plots illustrate how changes in a single feature affect the predicted outcome, while keeping all other features constant. By visualizing these dependencies, PDPs provide insights into the direction and magnitude of a feature's influence on the model's predictions. This approach enhances interpretability by revealing trends and interactions that might not be immediately evident from raw data analysis or permutation importance alone.

By integrating permutation importance analysis and partial dependence plots, we gained a comprehensive understanding of how specific features, such as chemical activators and pyrolysis conditions, influence the SSA, TPV, and MP of LDPC synthesized *via* catalytic pyrolysis. These interpretive techniques not only validate model predictions but also provide actionable insights for optimizing LDPC synthesis and application in various fields.

## Results and discussion

### Original dataset analysis

The original dataset comprises various input features and output targets. The input features include elemental compositions, proximate compositions, and pyrolysis conditions, while the output targets are Surface Area (SSA), Total Pore Volume (TPV), and Micropore Volume (MP). We visualized the distributions of these features and targets using boxplots (Fig. S1†). Most distributions were comprised of inliers, although some features, particularly elemental compositions, contained outliers. Table S3† outlines the missing values across the dataset. Notably, Volatile Matter (VM), Ash, and Fixed Carbon (FC) had missing values totaling approximately 19.64%. SSA and MP had missing values of 9.82% and 34.82%, respectively.

### Pre-trained interpolation model on original dataset

To address the missing features, we trained a pre-trained interpolation Random Forest model using a subset of 41 samples with complete features. We evaluated the model's performance through $R^2$ metrics (Fig. S2 and Table S4†). The results showed reasonable training $R^2$ values of over 0.85 and test $R^2$ values exceeding 0.65 for VM, Ash, and FC. For SSA and MP, we observed high training $R^2$ values of 0.76 and 0.71, respectively. However, the test $R^2$ values for SSA and MP were lower, at 0.27 and 0.50, respectively. This disparity can be attributed to the limited sample size of 41. The strong training $R^2$ values indicate reliable fittings for these features. As a result, the interpolated features can be incorporated into the original dataset, leading to the creation of a cleaned dataset.

### PCC analysis of cleaned dataset

Fig. 2 presents the Pearson Correlation Coefficient (PCC) values for pairwise comparisons among the variables in the cleaned dataset after interpolation. A strong positive correlation (PCC: 0.76) was observed between Surface Area (SSA) and Total Pore Volume (TPV), indicating that higher TPV often coincides with higher SSA. Within the SSA analysis, Volatile Matter (VM) and Fixed Carbon (FC) exhibited higher PCC values of 0.37 and 0.46, respectively, compared to other proximate components. The positive correlation between VM and SSA suggests that the abundant formation of small gas molecules during carbonization promotes the physical exfoliation of lignin-derived porous carbon (LDPC). Conversely, a negative correlation between VM and TPV (PCC: −0.36) was noted, indicating differing influences.

Chemical activators also play a significant role in affecting SSA, with KOH and $Na_2CO_3$ showing notable correlations (PCC: 0.39 and −0.36, respectively). The quantity of activators further influenced SSA, with a PCC of 0.48. Interestingly, despite their relationship with SSA, different activators impact TPV differently. $Na_2CO_3$ exhibited a higher correlation with TPV (PCC: 0.31) compared to other activating agents, suggesting unique activation mechanisms.

Regarding MP, a correlation with TPV was found (PCC: −0.45), though it was less pronounced than the correlation with SSA. Elemental and proximate compositions—specifically hydrogen (H), oxygen (O), and FC—were identified as key influencers for MP in LDPC, with correlation coefficients of 0.30, 0.24, and 0.56, respectively. Among the activating agents, $H_3PO_4$ showed a significant negative correlation with MP (PCC: −0.51), while KOH displayed a moderate positive correlation (PCC: 0.29), highlighting their potential for precise control over MP in LDPC. Notably, the impregnation ratio of activator to sample (A/S) demonstrated a lower correlation with MP, suggesting that the activation mechanisms of these agents are more critical than their relative quantities.

Fig. S3† illustrates the PCC values for pairwise comparisons in the original dataset. The PCC distributions between the cleaned dataset and the original dataset were similar, indicating that the main contributing features for SSA, TPV, and MP remained consistent. This unaltered correlation of features confirms the successful interpolation of the original dataset using the pre-trained interpolation model.

### Final regressor on cleaned dataset

We fitted the cleaned dataset using the final regressor, the Random Forest-based Gradient Boosting Regressor (RF-based

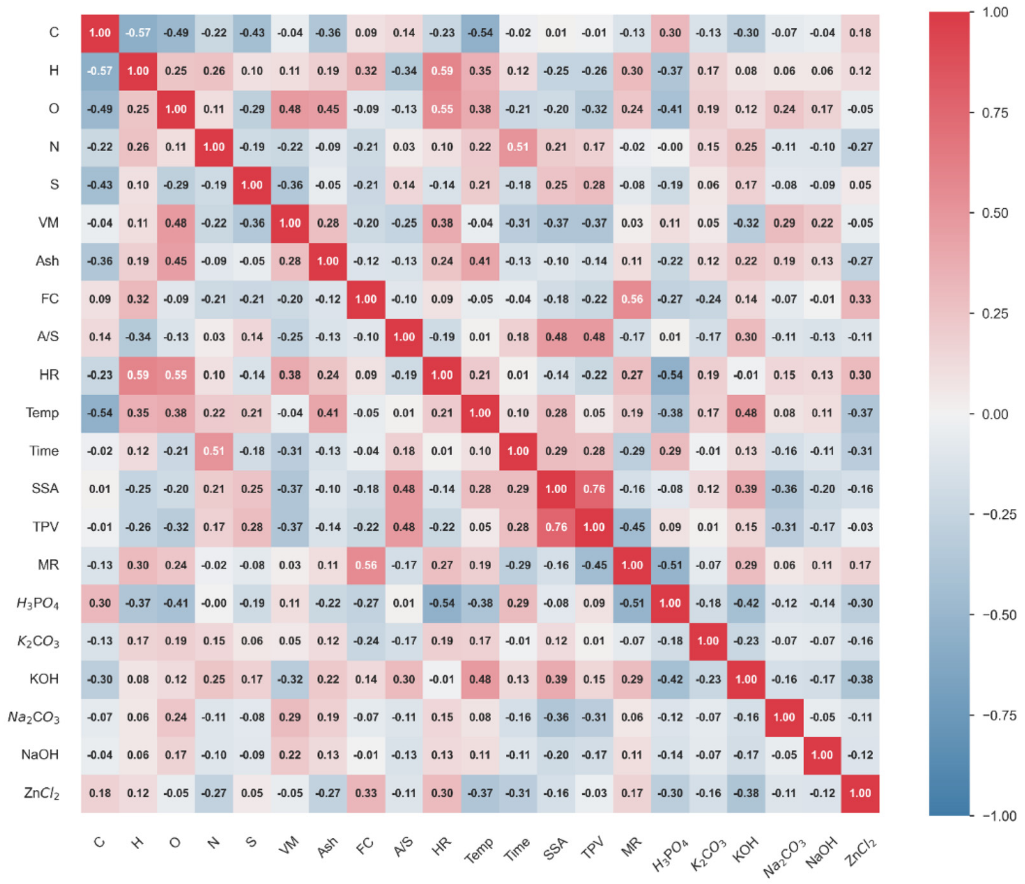

**Fig. 2** Pearson correlation matrix among any two features of the cleaned dataset.

GBR). To evaluate its performance, we also trained traditional Gradient Boosting Regressor (GBR) and Random Forest (RF) models on the original dataset to predict the same targets: SSA, TPV, and Micropore MP. The prediction accuracy was assessed using two metrics: $R^2$ and RMSE for both the training and testing datasets. Table 1 summarizes the RMSE and $R^2$ values for these models across the various targets. For SSA, which had 10.82% missing values in the original dataset, the RF-based GBR model performed similarly to the other models on the training set. However, on the test set, it demonstrated superior generalization, achieving an $R^2$ of 0.82 and an RMSE of 238.93 $m^2$ $g^{-1}$. In contrast, both RF and GBR showed lower generalization abilities, with smaller $R^2$ values and larger RMSE when using mean interpolation. These results indicate that RF-based interpolation significantly enhanced the reliability of the test dataset.

**Table 1** Training and testing performance of RF-based GBR, GBR, and RF models

| | | $R^2$ | | RMSE | | |
|---|---|---|---|---|---|---|
| Target | Model | Train | Test | Train | Test | Unit |
| SSA | RF-based GBR | 0.93 | 0.82 | 362.30 | 238.93 | $m^2$ $g^{-1}$ |
| | GBR | 0.91 | 0.74 | 342.89 | 280.46 | |
| | RF | 0.92 | 0.65 | 345.36 | 370.44 | |
| TPV | RF-based GBR | 0.94 | 0.86 | 0.26 | 0.17 | $cm^3$ $g^{-1}$ |
| | GBR | 0.89 | 0.73 | 0.25 | 0.24 | |
| | RF | 0.86 | 0.52 | 0.25 | 0.31 | |
| MP | RF-based GBR | 0.97 | 0.81 | 11.51 | 8.44 | % |
| | GBR | 0.85 | 0.38 | 14.48 | 18.54 | |
| | RF | 0.93 | 0.51 | 14.25 | 17.19 | |

TPV exhibited a similar trend to SSA. Pearson correlation analyses revealed significant effects of VM, Ash, and FC on TPV. The RF-based interpolation of these variables resulted in an $R^2$ of 0.86 and an RMSE of 0.17 $cm^3$ $g^{-1}$, demonstrating excellent generalization. For MP, which had the highest proportion of missing values at 44.82%, the RF-based interpolation also provided substantial benefits. The RF-based GBR model achieved an $R^2$ of 0.81 and an RMSE of 8.44%, outperforming both the RF and GBR models.

Fig. 3A–I show scatter plots comparing predicted values to actual values for SSA, TPV, and MP using the RF, GBR, and RF-based GBR models. The black dashed line (X = Y) indicates perfect alignment between predicted and actual values (Fig. 3D–F). For SSA, Fig. 3G and J present the $R^2$ and RMSE for the training and test sets using the RF-based GBR model. The model achieved an $R^2$ of 0.93 for the training set and 0.82 for the test set, demonstrating strong fitting and generalization

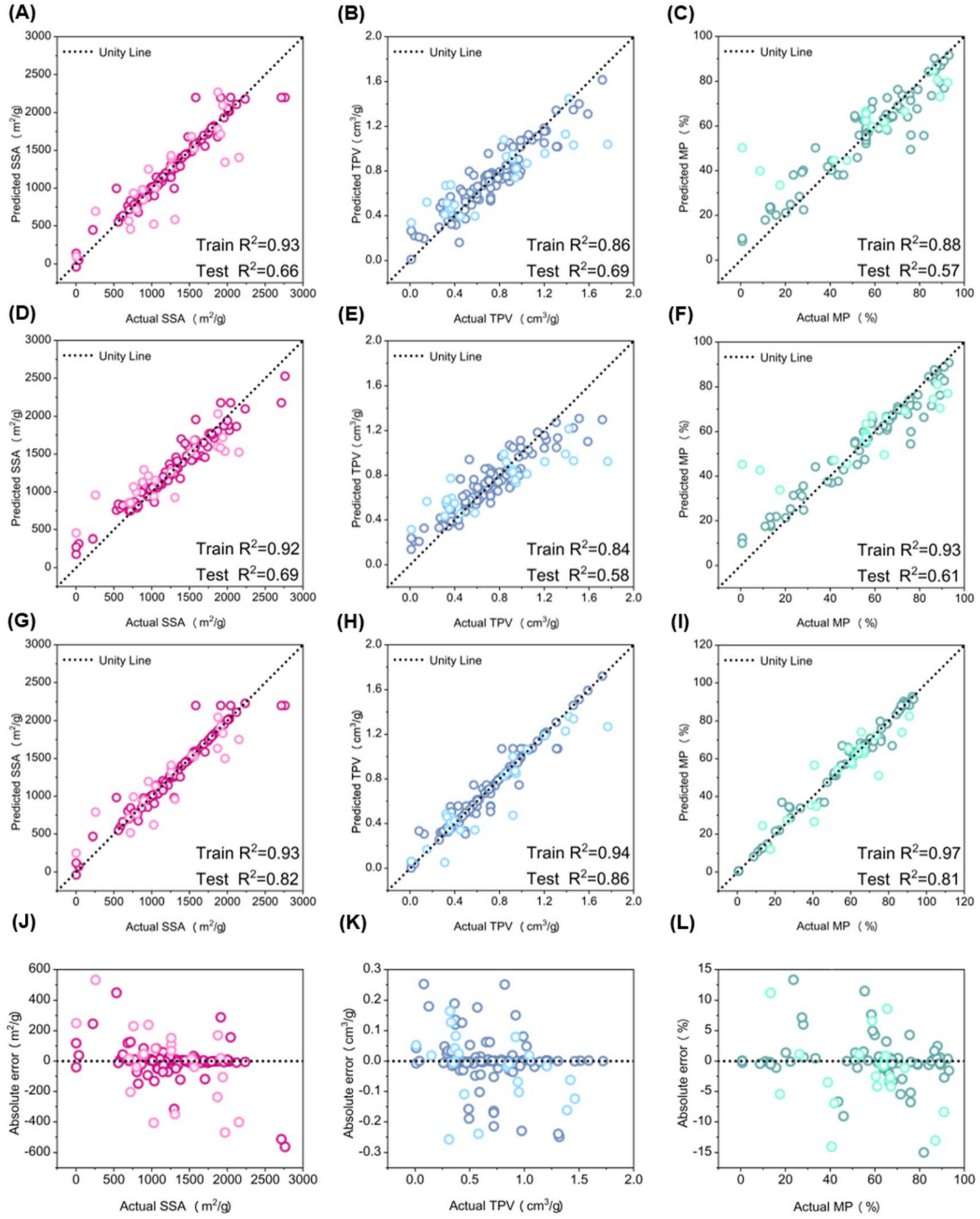

**Fig. 3** Comparison of predicted and actual values for SSA, TPV, and MP using (A–C) RF model, (D–F) GBR model, and (G–I) RF-based GBR model. Absolute error of all samples using RF-based GBR model for (J–L) SSA, TPV, and MP.

abilities, with a small absolute error of less than 200 $m^2$ $g^{-1}$ around the mean values. Similarly, Fig. 3H and K show the performance for TPV, where the RF-based GBR model attained an $R^2$ of 0.94 on the training set and 0.86 on the test set, again indicating robust fitting and generalization, with a small absolute error of less than 0.1 $cm^3$ $g^{-1}$. For MP, Fig. 3I and L report an $R^2$ of 0.97 for the training set and 0.81 for the test set, with a small absolute error of less than 5%. Overall, the RF-based GBR model (Fig. 3G–I) exhibited better alignment of predicted and actual values compared to the RF (Fig. 3A–C) and GBR (Fig. 3D–F) models.

## Feature importance of final regressor

Fig. 4 illustrates the contributions of various input features to the output targets using the RF-based GBR model. The input features are categorized into elemental composition, proximate composition, pyrolysis conditions, and chemical activators. The effects of these features on SSA, TPV, and MP are shown in Fig. 4A–C. Chemical activators emerged as the most significant contributors, accounting for 28.79% of the variance in SSA, 21.85% in TPV, and 20.87% in MP. Fig. 4D–F detail the contributions of different chemical activators, with KOH

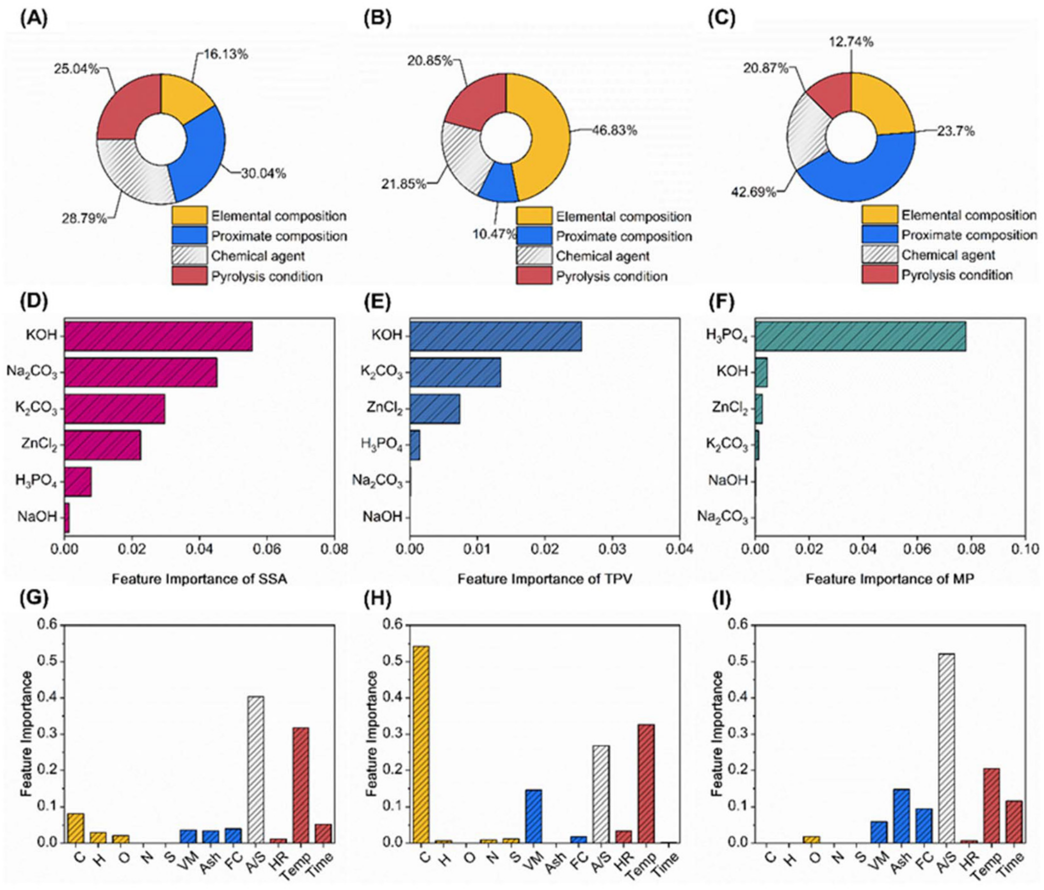

**Fig. 4** Contribution of individual input features to SSA, TPV, and MP using the RF-based GBR model. (A–C): feature contribution of each step; (D–F): feature contribution of different chemical activators; (G–I): feature contribution of other features under the catalytic pyrolysis of KOH or $H_3PO_4$.

showing the highest contributions of 0.06 for SSA and 0.028 for TPV, while $H_3PO_4$ contributed 0.08 for MP. The strong effects of KOH and $H_3PO_4$ on SSA, TPV, and MP are linked to their activation mechanisms. KOH promotes large pore volume and high SSA by reacting with carbon to form pore structures.[32,33] $H_3PO_4$, by catalyzing macromolecular chain fracture and dehydration, mainly contributes to narrow pore formation, which is reflected in the model's results on MP.[11,34]

Fig. 4G and H focus on the contributions of other features in the presence of KOH. For SSA, temperature and the activator-to-sample ratio (A/S) were identified as primary influencing factors. Variations in pyrolysis temperature not only affect carbonization levels but also trigger distinct etching reactions between KOH and the carbon framework. Experimental evidence suggests that KOH etches carbon atoms directly to produce hydrogen gas, while the decomposition of $K_2CO_3$ at 700 °C yields $K_2O$, enhancing the etching process.[2]

For TPV, temperature and A/S similarly emerged as key influences. The carbon content significantly affects TPV due to depletion during etching, impacting the integrity of the pore structure. This highlights that the stability of the carbon framework plays a greater role in TPV than the etching effects of the chemical activators, as shown in Fig. 4B.

Fig. 4I shows the contributions of various features to MP in the presence of $H_3PO_4$. The activator-to-sample ratio (A/S) was identified as the primary influencing factor, followed by temperature and carbonization time. The effect of $H_3PO_4$ on MP is primarily linked to the quantity of the chemical activator added, underscoring its critical role in forming microporous structures through reactions with carbon atoms, as illustrated in Fig. 4C.

## Partial dependence of final regressor

Building upon the feature contributions, partial dependence analysis was conducted for the top four features that significantly influence SSA, TPV, and MP of LDPC using different chemical activators. All scatter points are fitted within a cubic polynomial (Table S5†). Fig. 5(A, D, G and J) illustrate the partial dependence plots for the top features affecting SSA under KOH activation. Among these, A/S emerges as the most influential. Initially, SSA increases with increased A/S due to enhanced etching reactions between KOH and carbon at temperatures above 700 °C. However, excessively high A/S can lead to the collapse of mesoporous structures, resulting in decreased SSA. This result highlights the critical role of selecting an optimal doping amount to obtain the maximum SSA.

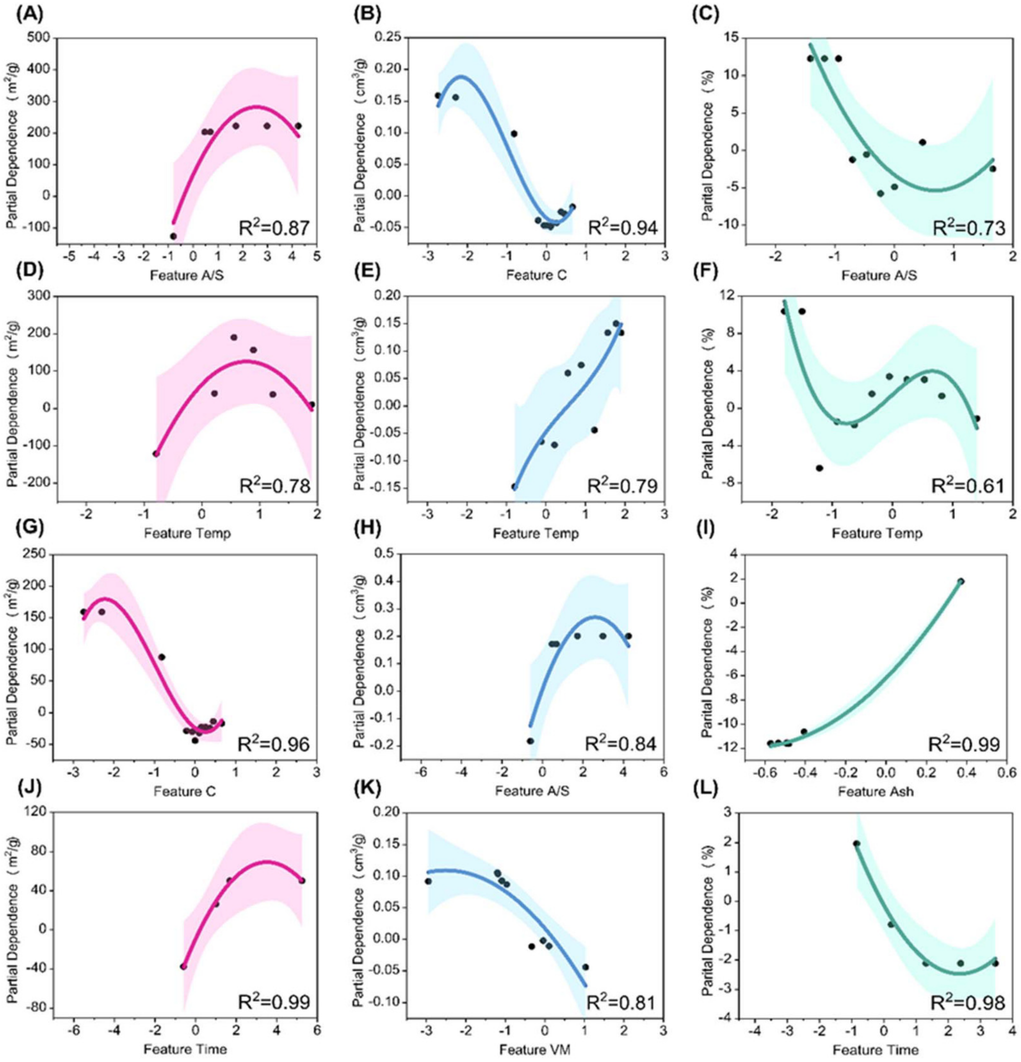

**Fig. 5** Partial dependence analysis of top four input features on each target: SSA of LDPC with KOH (A, D, G and J); TPV of LDPC with KOH (B, E, H and K); and MP of LDPC with $H_3PO_4$ (C, F, I and L).

Similarly, A/S significantly influences TPV under KOH (Fig. 5H) and MP (Fig. 5C) under $H_3PO_4$. With an increase in A/S, TPV initially increases and then decreases, while an opposite trend was observed on MP. This indicates a transition from micropore to mesopore structures with increased A/S. However, excessive etching at higher A/S levels can deplete carbon atoms with the collapse of the carbon framework. The trend of A/S affecting SSA and TPV under KOH is similar to findings by Dai *et al.*, where increasing A/S from 1 to 3 resulted in higher SSA and TPV, before decreasing at A/S of 4.[35] While The trends of A/S, Temp, and Time affecting MP under $H_3PO_4$ are supported by Liao *et al.*'s study, where MP decreased as A/S increased from 20% to 60%, and varied with temperature and time.[36]

Temp exhibits a similar influence on SSA under KOH activation and MP under $H_3PO_4$ (Fig. 5D and F), with both initially increasing and then decreasing with increasing carbonization temperatures. This trend is also experimentally verified. Liao *et al.* investigated the influence of carbonization temperature between 450–750 °C on the SSA of porous carbons using $H_3PO_4$ as the activation agent.[36] The results indicated that with an increase in carbonization temperature, more mesoporous structures appear in the prepared porous carbons, leading to a decrease in SSA from the highest of 1215.82 $m^2\ g^{-1}$ at 550 °C to 980.88 $m^2\ g^{-1}$ at 750 °C. They attributed this to the aggregation and destruction of pore structures at high temperatures. At elevated temperatures, a large number of carbon atoms gradually enter an activated state, resulting in the gradual erosion of the carbon framework and collapse of the internal structure of porous carbons. The transition from micropores to mesopores, transitional consumption of carbon atoms, and collapse of the carbon framework collectively contribute to the decrease in SSA. This effect is also evident in the influence of carbonization temperature on TPV of porous carbons synthesized under KOH activation (Fig. 5E), with TPV tending to increase as the temperature rises, indicating a transition from micropores to mesopores with higher temperatures. The trend of Temp and Time affecting SSA and TPV under KOH is similar to findings by Li *et al.* found that higher pyrolysis temperatures increase both SSA and TPV, and increasing pyrolysis time from 1 h to 2 h led to higher SSA, consistent with our machine learning model.[37]

Regarding the influence of differences in elemental and proximate composition on SSA, TPV, and MP, Fig. 5G and K highlights VM and C as the most significant features. VM is decomposed to small gas molecules to physically create pores. This physical stripping is often uncontrollable, which has a lower pore formation efficiency compared to that of catalytic pyrolysis. Therefore, with an increase in VM, TPV tends to decrease (Fig. 5K). The negative correlation between VM and TPV is also supported by previous studies.[37,38] For the element C, as the activation agent mainly etches carbon atoms to form pore structures, fewer carbon atoms are more easily etched into pores. Conversely, a higher number of carbon atoms implies higher stability of pore structures, which was less unaffected from excessive etching (Fig. 5G and B). The negative correlations between C content and SSA/TPV from our model are consistent with experimental results from Zhang *et al.*, Xi *et al.*, and Li *et al.*[37,39,40]

## Conclusions

In this study, we developed a hybrid machine learning framework incorporating a pre-trained interpolation model and a final regressor using elemental composition, proximate composition, chemical activation, and pyrolysis conditions as inputs to predict SSA, TPV, and MP of LDPC synthesized *via* catalytic pyrolysis. The pre-trained interpolation model effectively handled missing data and optimized model hyperparameters. Despite the modest dataset size of 112 samples, our final regressor demonstrated robust predictive accuracy across six different chemical activators: KOH, $ZnCl_2$, $H_3PO_4$, $K_2CO_3$, NaOH, and $Na_2CO_3$, yielding $R^2$ values of 0.82 for SSA, 0.86 for TPV, and 0.81 for MP. Interpretability analysis highlighted the significant influence of KOH on SSA and TPV, while $H_3PO_4$ notably affected MP. Additionally, Temp and A/S emerged as critical factors influencing all three properties. This research presents a practical approach for accurately predicting LDPC properties under catalytic pyrolysis conditions, facilitating advancements in their tailored application across various fields.

## Author contributions

Zihao Xie: writing – original draft, investigation, validation, formal analysis. Yue Cao: conceptualization, methodology. Zhicheng Luo: writing – review & editing, supervision, project administration.

## Data availability

The data supporting this article have been included as part of the ESI.†

## Conflicts of interest

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

## Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2024YFB4206400), the National Natural Science Foundation of China (grant number 52206236), the Joint Funds of the National Natural Science Foundation of China (U23A2090), the Natural Science Foundation of Jiangsu Province (grant number BK20220837), the Fundamental Research Funds for the Central Universities (3203002211A1, 2242024k30025).

## References

1 W.-J. Liu, H. Jiang and H.-Q. Yu, *Chem. Rev.*, 2015, **115**, 12251–12285.

2 B. Zhang, Y. Jiang and R. Balasubramanian, *J. Mater. Chem. A*, 2021, **9**, 24759–24802.

3 Z. Luo, C. Liu, A. Radu, D. F. de Waard, Y. Wang, J. T. Behaghel de Bueren, P. D. Kouris, M. D. Boot, J. Xiao, H. Zhang, R. Xiao, J. S. Luterbacher and E. J. M. Hensen, *Nat. Chem. Eng.*, 2024, **1**, 61–72.

4 S. Yu, L. Bie and Z. Luo, *Chem. Eng. J.*, 2024, **494**, 153030.

5 S. Chu, A. V. Subrahmanyam and G. W. Huber, *Green Chem.*, 2013, **15**, 125–136.

6 W.-J. Chen, C.-X. Zhao, B.-Q. Li, T.-Q. Yuan and Q. Zhang, *Green Chem.*, 2022, **24**, 565–584.

7 H. Y. Yang, Z. J. Han, S. F. Yu, K. L. Pey, K. Ostrikov and R. Karnik, *Nat. Commun.*, 2013, **4**, 2220.

8 L. Sun, Y. Gong, D. Li and C. Pan, *Green Chem.*, 2022, **24**, 3864–3894.

9 H. Shao, Y.-C. Wu, Z. Lin, P.-L. Taberna and P. Simon, *Chem. Soc. Rev.*, 2020, **49**, 3005–3039.

10 G. Zhang, X. Liu, L. Wang and H. Fu, *J. Mater. Chem. A*, 2022, **10**, 9277–9307.

11 W. Zhang, J. Yin, C. Wang, L. Zhao, W. Jian, K. Lu, H. Lin, X. Qiu and H. N. Alshareef, *Small Methods*, 2021, **5**, 2100896.

12 N. Díez, M. Sevilla and A. B. Fuertes, *Carbon*, 2021, **178**, 451–476.

13 C. Wang, D. Yang, S. Huang, Y. Qin, W. Zhang and X. Qiu, *Green Chem.*, 2022, **24**, 5941–5951.

14 Y. Gao, Q. Yue, B. Gao, Y. Sun, W. Wang, Q. Li and Y. Wang, *Chem. Eng. J.*, 2013, **217**, 345–353.

15 Y. Li, R. Gupta and S. You, *Bioresour. Technol.*, 2022, **359**, 127511.

16 X. Yuan, M. Suvarna, S. Low, P. D. Dissanayake, K. B. Lee, J. Li, X. Wang and Y. S. Ok, *Environ. Sci. Technol.*, 2021, **55**, 11925–11936.

17 X. Yang, C. Yuan, S. He, D. Jiang, B. Cao and S. Wang, *Fuel*, 2023, **331**, 125718.

18 W. A. M. Wickramaarachchi, M. Minakshi, X. Gao, R. Dabare and K. W. Wong, *Chem. Eng. J. Adv.*, 2021, **8**, 100158.

19 K. M. Jablonka, D. Ongari, S. M. Moosavi and B. Smit, *Chem. Rev.*, 2020, **120**, 8066–8129.

20 T. Wang, R. Pan, M. L. Martins, J. Cui, Z. Huang, B. P. Thapaliya, C.-L. Do-Thanh, M. Zhou, J. Fan, Z. Yang, M. Chi, T. Kobayashi, J. Wu, E. Mamontov and S. Dai, *Nat. Commun.*, 2023, **14**, 4607.

21 W. Zhang, Q. Chen, J. Chen, D. Xu, H. Zhan, H. Peng, J. Pan, M. Vlaskin, L. Leng and H. Li, *Bioresour. Technol.*, 2023, **370**, 128547.

22 X. Zhu, Z. Xu, S. You, M. Komárek, D. S. Alessi, X. Yuan, K. N. Palansooriya, Y. S. Ok and D. C. W. Tsang, *Chem. Eng. J.*, 2022, **428**, 131967.

23 H. Li, Z. Ai, L. Yang, W. Zhang, Z. Yang, H. Peng and L. Leng, *Bioresour. Technol.*, 2023, **369**, 128417.

24 L. Leng, L. Yang, X. Lei, W. Zhang, Z. Ai, Z. Yang, H. Zhan, J. Yang, X. Yuan, H. Peng and H. Li, *BioChar*, 2022, **4**, 63.

25 X. Zhu, Y. Li and X. Wang, *Bioresour. Technol.*, 2019, **288**, 121527.

26 R. Zou, Z. Yang, J. Zhang, R. Lei, W. Zhang, F. Fnu, D. C. W. Tsang, J. Heyne, X. Zhang, R. Ruan and H. Lei, *Bioresour. Technol.*, 2024, **399**, 130624.

27 C. Wang, W. Jiang, G. Jiang, T. Zhang, K. He, L. Mu, J. Zhu, D. Huang, H. Qian and X. Lu, *Ind. Eng. Chem. Res.*, 2023, **62**, 11016–11031.

28 J. Lee, S. Hong, H. Cho, B. Lyu, M. Kim, J. Kim and I. Moon, *Energy Convers. Manage.*, 2021, **244**, 114438.

29 C. Ribeiro and A. A. Freitas, *Artif. Intell. Rev.*, 2021, **54**, 6277–6307.

30 L. Leng, W. Zhang, T. Liu, H. Zhan, J. Li, L. Yang, J. Li, H. Peng and H. Li, *Bioresour. Technol.*, 2022, **358**, 127348.

31 J. Li, L. Pan, M. Suvarna and X. Wang, *Chem. Eng. J.*, 2021, **426**, 131285.

32 W. Zhang, X. Qiu, C. Wang, L. Zhong, F. Fu, J. Zhu, Z. Zhang, Y. Qin, D. Yang and C. C. Xu, *Carbon Res.*, 2022, **1**, 14.

33 N. Guo, M. Li, X. Sun, F. Wang and R. Yang, *Green Chem.*, 2017, **19**, 2595–2602.

34 M. A. Yahya, Z. Al-Qodah and C. W. Z. Ngah, *Renewable Sustainable Energy Rev.*, 2015, **46**, 218–235.

35 J. Dai, A. Xie, R. Zhang, W. Ge, Z. Chang, S. Tian, C. Li and Y. Yan, *J. Mol. Liq.*, 2018, **256**, 203–212.

36 Z. Liao, Y.-H. Zhu, G.-T. Sun, L. Qiu and M.-Q. Zhu, *Ind. Crops Prod.*, 2022, **175**, 114266.

37 M. Li, X. Liu, C. Sun, L. Stevens and H. Liu, *J. Environ. Chem. Eng.*, 2022, **10**, 107471.

38 D.-W. Lee, M.-H. Jin, J.-H. Park, Y.-J. Lee and Y.-C. Choi, *ACS Sustainable Chem. Eng.*, 2018, **6**, 10454–10462.

39 B. Zhang, D. Yang, X. Qiu, Y. Qian, M. Yan and Q. Li, *J. Ind. Eng. Chem.*, 2020, **82**, 220–227.

40 Y. Xi, D. Yang, X. Qiu, H. Wang, J. Huang and Q. Li, *Ind. Crops Prod.*, 2018, **124**, 747–754.