Study on the electrocatalytic CO_2 reduction performance of covalent organic framework materials based on machine learning

Xiangying Zhang, Sitan Wang*

College of Environmental Science and Engineering, Liaoning Technical University, Fuxin 123000, PR China

Abstract: In order to accurately predict the catalytic performance of covalent organic framework materials (COFs) for electrocatalytic carbon dioxide and analyze the influencing factors affecting the catalytic effect, this study collected COFs structure data and experimental data from 44 literatures, and used machine learning methods. Six regression models were trained and evaluated with COFs structure data and experimental data as features and Faraday efficiency as output. By evaluating the fitting coefficient, mean absolute error and the fitting effect of the test set, the extreme gradient boosting (XGB) model has the best performance. Through the visual analysis of the partial dependence diagram and the individual expectation condition diagram of the XGB model, the coordination metal is Ni, the coordination metal content is greater than 10 %, the pore limit diameter is in the range of 2.5nm-12.5nm. The COFs with tetragonal crystal system have high Faraday efficiency. This research method can not only accurately predict the catalytic performance of covalent organic framework materials (COFs) for electrocatalytic carbon dioxide, but also provide a reference for the screening of catalysts according to the structural characteristics.

1 Introduction

In recent years, due to the rapid development of industry, CO_2 emissions have increased significantly [1]. The greenhouse effect caused by it has caused a series of serious climate changes [2], such as the gradual rise of sea level, frequent landings of typhoons and hurricanes, and global warming, which pose a serious threat to human survival. Electrocatalytic CO_2 reduction can be converted into various valuable products [3], such as carbon monoxide (CO), formate (HCOO-), methane (CH_4) and so on.

Covalent organic frameworks (COFs) have high specific surface area, adjustable structure, high crystallinity and customizable functionalization [4]. However, due to the complexity of the factors affecting the catalytic effect of COFs materials, the long cycle and high cost of synthesizing COFs materials by artificial experimental methods [5].

Machine learning (ML) is the basis for driving computer intelligence [6]. It can enable machines to have automatic learning capabilities and realize non-display programming. However, the field of electrocatalysis is still in its infancy [7], especially the prediction of the catalytic performance of COFs for electrocatalytic carbon dioxide is rarely reported.

In this study, machine learning methods were used to predict the catalytic performance of COFs for electrocatalytic carbon dioxide. The structural characteristics of COFs and the experimental parameters of electrocatalytic CO_2 in 44 literatures were collected and collated as data sources. Six regression models are selected. According to the evaluation results, the best model was selected to analyze the influencing factors of COFs electrocatalytic CO_2, and to predict the electrocatalytic CO_2 catalytic performance of COFs, so as to provide a reference for efficient synthesis of COFs.

2 Data and Methods

2.1 Data Collection

In this study, 44 articles on COFs electrocatalytic CO_2 reduction were collected from Web of Science, Science Direct, ACS publications and other databases by manual screening, and 1057 data points were extracted. The database uses Faraday efficiency (FE) as the output value. It involves 26 characteristics: pore volume (Vt), pore size(PS),modification method (MT), specific surface area (BET), temperature (TEM), CO_2 adsorption capacity (UT), electric double layer capacitance (Cdl), Tafel slope (TF), pH value (pH), type of electrolytic cell (TC), crystal type (CS), electrolyte concentration (NM), electrolyte (EL), potential (PT), current density (J), coordination number (CN), space configuration (SC), functional group (FG), coordination metal (CM), coordination metal content (CT), coordination bond (CB), construction block 1 (BB1), construction block 2 (BB2), compound (COM), pore limit diameter (PLD), product (PD).

*wangsitan@lntu.edu.cn

2.2 Data pre-processing

In this study, the data set was divided into training set and test set. 92.6 % of the data was selected for model training and parameter adjustment, and the remaining 7.4 % of the data was used for model evaluation. Because the linear algorithm is simple, it is only suitable for dealing with problems with linear relationship, while the nonlinear algorithm is suitable for dealing with problems with nonlinear relationship and more complex problems. Therefore, AdaBoost, RF, SVR, KNN, GBDT and XGB nonlinear algorithms are used for modeling in this study. The 10-fold cross-validation and grid search methods were used to adjust the parameters to ensure that the optimal parameters were obtained. The fitting coefficient (R^2) and the mean absolute error (MAE) were selected as the evaluation indicators. By comparing the R^2 and MAE of different models, the model suitable for this data set was selected for subsequent analysis.

3 Model and algorithm

3.1 Model selection

In this study, the data set was divided into training set and test set. 92.6 % of the data was selected for model training and parameter adjustment, and the remaining 7.4 % of the data was used for model evaluation. Therefore, AdaBoost, RF, SVR, KNN, GBDT and XGB nonlinear algorithms are used for modeling in this study. By comparing the R^2 and MAE of different models, the model suitable for this data set was selected for subsequent analysis.

3.2 Analysis of relationship

Through correlation analysis, it helps to understand whether there is a relationship between variables and the strength of the relationship, and helps to identify redundant features, avoid using them in the modeling process. It shows the Pearson correlation matrix between the variables, in which there is a strong positive correlation between the pore volume and the specific surface area of the catalyst.

3.3 Model evaluation

AdaBoost, RF, SVR, KNN, GBDT and XGB are adjusted to the optimal parameters. After the training is completed, the test set is tested. The fitting coefficient (R^2) and the mean absolute error (MAE) are used as evaluation indicators. The closer the fitting coefficient is to 1, the smaller the mean absolute error is, and the better the model prediction performance is, so as to obtain the optimal model. The fitting effect is shown in Figure 1. Comparing the fitting coefficients of the six models with the mean absolute error and the fitting of the test set, the XGB model has the best prediction performance. The optimal parameters are: n _ estimators = 3000. Learning _ rate = 0.06, max _ depth = 5, min _ child _ weight = 3, min _ split _ loss = 0.3, subsample = 0.7, colsample _ bytree = 0.8.

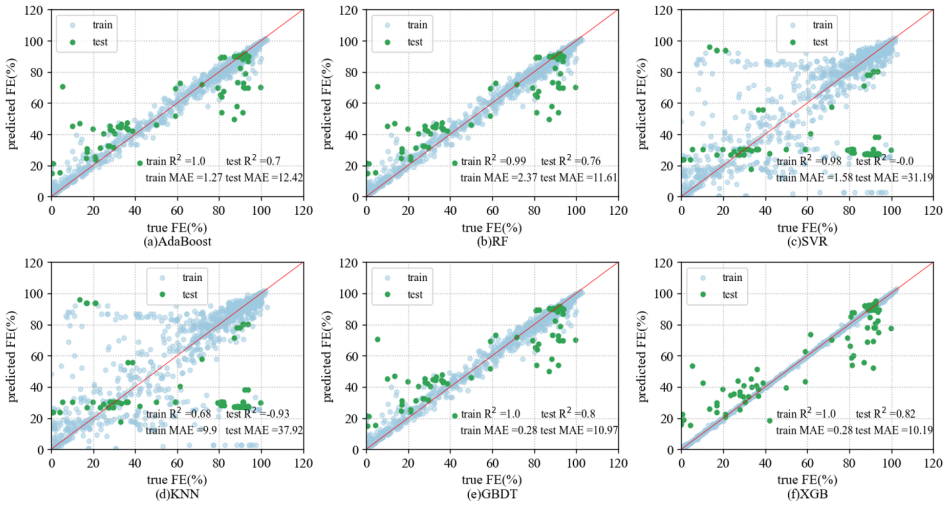

Figure 1. Fitting effect diagram of six machine learning models to test database

4 Analysis

4.1 Analysis of feature importance

Through the trained XGB model, the contribution of each feature to the model prediction results is measured, and the correlation between the feature and the target variable is highlighted. As shown in Figure 2, the top ten importance of the XGB model for data characteristics are : J > PLD > CT > UT > TF > Cdl > pH > PT > BET > PS.

First, the current density (J) contributes the most to the output Faraday efficiency (FE), indicating that the current density generated by the catalyst affects the rate of the electrocatalytic reaction, which in turn affects the Faraday efficiency. Secondly, the contribution of pore limit diameter (PLD) ranks second, and the size of pore limit diameter affects the dispersion and activity of the catalyst, thus affecting the reaction rate. Thirdly, the coordination metal

content (CT), carbon dioxide adsorption capacity (UT), Tafel slope (TF), capacitance value (Cdl), pH value (pH) and potential (PT) are all experimental parameters, and due attention should be paid to the adjustment of parameters. In addition, the specific surface area (BET) and pore size (PS) are the structural characteristics of the catalyst, and the specific surface area and pore size should be considered.

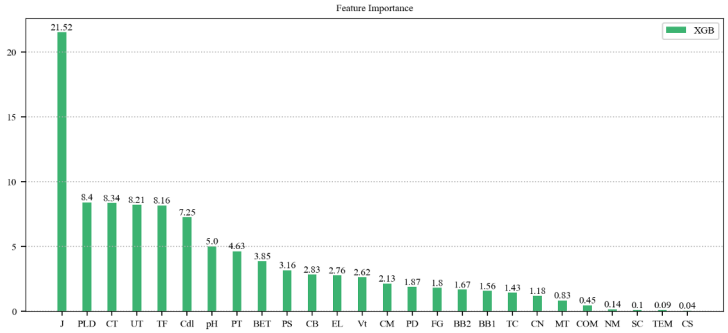

Figure 2. Feature importance

4.2 Model explanation

Individual Conditional Expectation Plot (ICE) is a visual tool that describes the relationship between the predicted value of each individual and a single variable. It shows a curve for each instance, showing how the prediction of the instance changes when the feature changes, so as to visually predict the relationship between the features of each instance.

Figure 3 is the individual conditional expectation diagram of quantitative characteristics. It can be concluded from the figure that the structural conditions of COFs with high Faraday efficiency are as follows : specific surface area (BET) greater than 1500 m^2 g^{-1}, pore volume (Vt) greater than 1.0 cm^3 g^{-1}, pore size (PS) greater than 2.0 nm, coordination number (CN) greater than 4, coordination metal content (CT) greater than 10 %, pore limit diameter (PLD) greater than 2.5 less than 12.5 nm. With the increase of temperature (TEM), carbon dioxide adsorption capacity (UT), electric double layer capacitance (Cdl), electrolyte solution concentration (NM), potential (PT), current density (J) and other experimental conditions, the Faraday efficiency keeps a high trend. When the Tafel slope is less than 600 and the pH value is in the range of 0 ~ 7.5, the Faraday efficiency is higher.

Figure 4 shows the individual expectation condition diagram of qualitative characteristics. It can be concluded from the figure that the structural conditions of COFs with high Faraday efficiency are as follows : crystal form (CS) is tetragonal system, group (FG) is ' amino ' and ' aminophenyl ', coordination metal (CM) is Ni, coordination bond (CB) is ' Ar-N ', ' C = N ', ' Co-N ', ' Co-N, Ni-N ', structure (BB) is ' DAPP ' and ' MTAPP ' ; the experimental conditions are as follows : using carbonization and composite modification method (MT), using h-type electrolytic cell (TC), electrolyte (EL) is K_2SO_4, $KHCO_3$, the product is carbon monoxide.

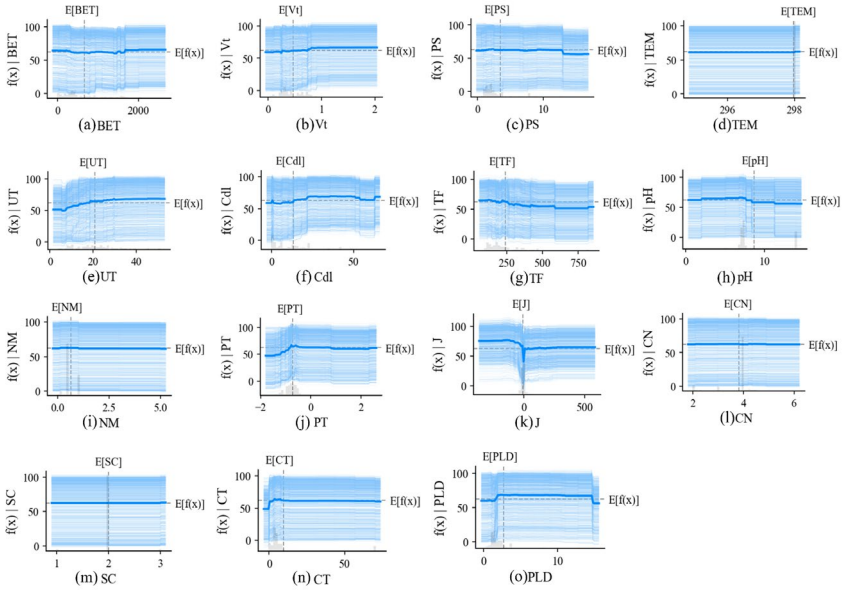

Figure 3. Individual conditional expectation graph of quantitative features

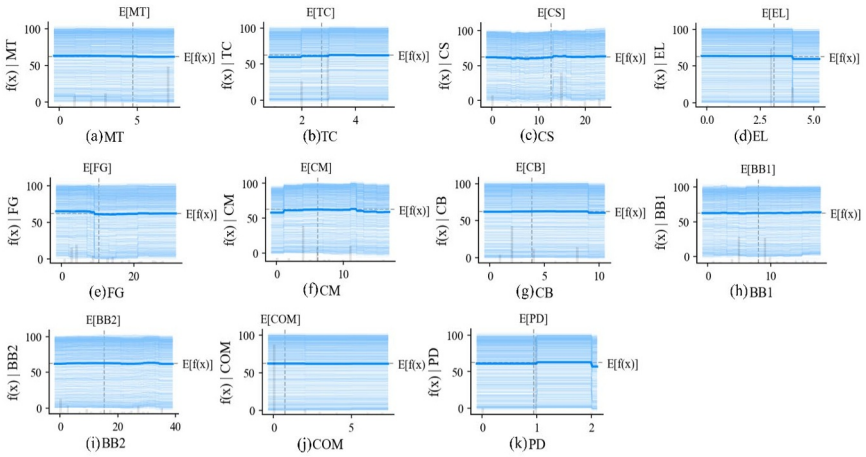

Figure 4. Individual conditional expectation graph of qualitative features

5 Conclusion

In this study, a data set of COFs electrocatalytic carbon dioxide was established, and KNN interpolation method was used to train and evaluate six models : adaptive boosting algorithm (AdaBoost), random forest algorithm (RF), support vector regression (SVR), K-nearest neighbor algorithm (KNN), gradient boosting decision tree (GBDT) and extreme gradient boosting (XGB) six regression models. The results show that the XGB model has the best prediction performance. The fitting coefficient of the training set reaches 1.00 and the average absolute error is 0.28. The fitting coefficient of the test set reaches 0.82 and the average absolute error is 10.19.

The top ten terms of the importance of the XGB model for data characteristics are : current density (J) > pore limit diameter (PLD)>coordination metal content (CT)> carbon dioxide adsorption (UT) > Tafel slope (TF) > electric double layer capacitance (Cdl) > pH (pH) > potential (PT) > specific surface area (BET)>pore size (PS). It can be concluded that the experimental characteristics such as current density, Tafel slope, carbon dioxide adsorption capacity, electric double layer capacitance, potential and pH value have great influence on Faraday efficiency. The pore size, coordination metal content, pore limit diameter, specific surface area and pore volume have great contribution to Faraday efficiency.

Through the visualization analysis of the individual expectation condition map, it is concluded that the conditions for COFs with high Faraday efficiency are as follows : the coordination metal is Ni, the coordination metal content is more than 10 %, the coordination number is more than 4.2, the structure is metal porphyrin (MTAPP), the specific surface area is more than $1500 m^2 g^{-1}$, the pore volume is more than $1.0 cm^3 g^{-1}$, the pore diameter is more than 3.0nm, the pore limit diameter is in the range of 2.5nm ~ 12.5nm, and the crystal form is tetragonal.

In view of the increasingly rich covalent organic framework, it is necessary to continuously enrich the data set and improve the generalization ability of the model to achieve better prediction ability. This research method can not only accurately predict the catalytic performance of covalent organic framework materials (COFs) for electrocatalytic carbon dioxide, but also provide a reference for the screening of catalysts according to the structural characteristics.

References

1. GHADIKOLAEI S S C. An enviroeconomic review of the solar PV cells cooling technology effect on the CO2 emission reduction [J]. Sol Energy, 2021, 216: 468-492.
2. Shah R, Ali S, Raziq F, et al. Exploration of metal organic frameworks and covalent organic frameworks for energy-related applications [J]. Coordination Chemistry Reviews, 2023, 477: 214968.
3. LI Xia, ZHANG Sainan, GAO Jia, et al. Research progress and application of chiral covalent organic framework materials [J]. China Science: Chemistry, 2019,49 (05) : 662-671.
4. HE Xiaoman, GUO Jingyuan, SHEN Dekui, et al. The application progress and prospect of machine learning in anaerobic digestion of organic solid waste [J / OL]. Journal of Southeast University (Natural Science Edition), 1-12 [2025-01-09].
5. YUAN Zefei, ZHANG Zhengjun, JIANG Guolin. Adaptive spectral clustering algorithm based on relative proximity [J / OL]. Computer science, 1-12.
6. DING Peng, WANG Bin, ZHU Sulei. Trajectory user link prediction based on siamese hierarchical attention network model [J]. Computer application and software, 2024,41 (11): 213-219 + 227.
7. LI Zitong, MENG Xiaofeng, WANG Leixia, et al. Review of Machine Forgetting [J / OL]. Journal of Software, 1-28.