Machine learning analysis of literature dataset on the CO_2 reduction toward prediction of high activity multi-elemental catalysts

Umi Shin*[1] Shinya Mine*[1], Toshiaki Taira[1]
*Equal contribution [1][Research Institute for Chemical Process Technology, National Institute of Advanced Industrial Science and Technology (AIST), 4-2-1, Nigatake, Miyagino-ku, Sendai, 983-8551, Japan]

Correspondence to: [Toshiaki Taira] t-taira@aist.go.jp

Introduction

Carbon dioxide (CO_2) is one of the major contributors for global warming. Catalysts play a crucial role in enabling the efficient conversion of inert CO_2 molecule into value-added chemicals, thereby offering a promising strategy for CO_2 utilization. Catalytic performance strongly depends on both catalyst materials and reaction conditions. Supported metal-oxide catalysts, as major catalysts for CO_2 conversion, enable high-efficiency conversion through the rational design of metal-oxide and support components.

The development of supported metal-oxide catalysts using machine learning (ML) has attracted increasing attention as a promising approach to accelerate catalyst discovery [1,2]. However, the iterative approach of ML prediction and experimentation still faces significant challenges in terms of the time and cost required for data acquisition. Thus, it is essential to establish a machine-learnable data infrastructure from existing literature and to develop a system capable of proposing promising catalyst candidates.

In this study, we develop machine learning models to predict novel catalyst candidates for CO_2 reduction, focusing on the reverse water–gas shift (RWGS) reaction. The models are trained using a dataset compiled from 61 peer-reviewed studies on RWGS published up to 2024, with data selected to match experimentally used catalyst compositions and reaction conditions.

Data and ML Method

The literature dataset was constructed from 61 peer-reviewed papers published up to 2024, resulting in 607 data entries. For each study, catalyst composition, preparation conditions, and reaction conditions were systematically recorded, under the following selection criteria: (i) supported metal-oxide catalysts synthesized by the impregnation method, and (ii) reaction temperatures between 300 and 400 °C.

Catalyst compositions were used as input variables, and CO yield was used as the target variable. Since the catalytic activity of CO_2 reduction strongly depends on metals, support materials, and their interactions, the catalyst compositions need to be appropriately optimized to achieve high performance. To enable systematic optimization of catalyst compositions and improve prediction accuracy, each elemental component was represented using physicochemical descriptors (e.g., melting point, electronegativity, and atomic radius), which were weighted by their elemental fractions in the catalyst. This descriptor-based representation allows the ML model to account for chemical similarities between different elements, rather than treating them as mere atomic symbols. Based on this representation, model performance was evaluated by 10-fold cross-validation using the root mean square error (RMSE) and coefficient of determination (R^2) between predicted and experimental values.

Results

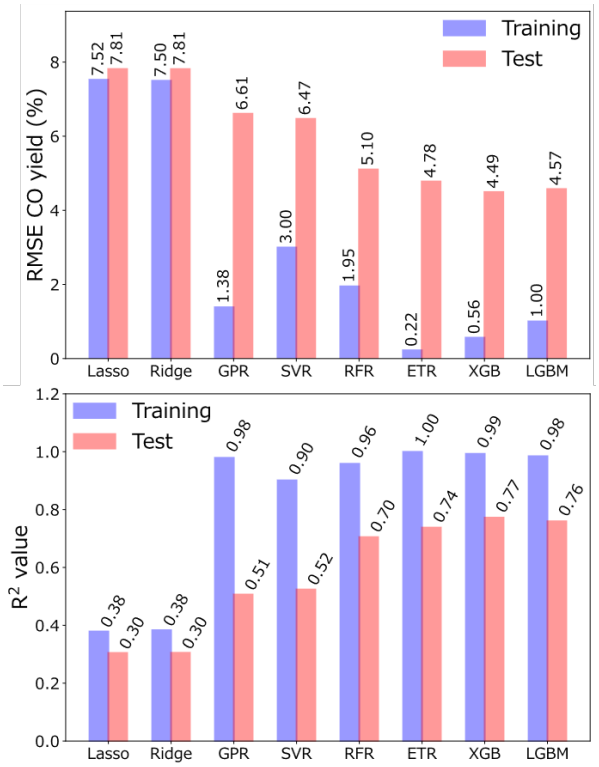

Fig. 1: RMSE and R2 comparisons of different ML models for CO yield in the RWGS reaction

The prediction performance of the eight regression models: Lasso, Ridge, Gaussian Process Regression (GPR), Support Vector Regression (SVR), Random Forest Regression (RFR), Extra Trees Regression (ETR), and two gradient boosting models (XGBoost (XGB) and LightGBM (LGBM)) were evaluated using the averaged RMSE and R^2. Fig. 1 summarizes the RMSE and R^2 values for CO yield prediction. Tree- ensemble methods (RFR, ETR, XGB, and LGBM) generally outperformed others, with XGB showing the best performance (RMSE = 4.49, R^2 = 0.77). This accuracy is comparable to that of a previous study achieving R^2 = 0.82 using an ETR model trained on 300 data points obtained through 44 experimental–ML iterations [1].

To inspect the ML predictions in greater detail, we applied SHapley Additive exPlanations (SHAP) analysis to the best-performing XGB model, as shown in Fig. 2.

Among the experimental conditions, reaction temperature was identified as the most important factor: high values of "reaction temperature" (red colour in Fig. 2) are associated with high CO yield (positive SHAP values), whereas low values (blue colour) contribute more strongly to high CH_4 yield. Since RWGS reaction is an endothermic reaction[4], higher temperatures thermodynamically favour the formation of CO over CH_4, in contrast to the exothermic methanation reaction.

In addition to reaction temperature, pretreatment temperature was identified as the second most influential factor, where lower pretreatment temperatures are associated with higher CO yields, showing an opposite trend to that observed for reaction temperature. This behaviour suggests that high temperature promotes catalyst sintering, which reduces the number of active sites available for the RWGS reaction.

These results indicate that both the pretreatment and reaction temperature have critical role on CO yields, so they must be carefully optimized to achieve high RWGS performance.

As for catalyst composition, Pt and Cu were identified as important descriptors for CO yield, whereas Ni was found to promote CH_4 formation. Cu is a key component of in industrial RWGS catalysts (Cu–Zn/Al_2O_3), and is well known to favour CO production from CO_2. Pt is also known to be a highly active metal for the RWGS reaction[5]. In contrast, Ni is a highly active metal for the Sabatier reaction, which converts CO_2 into CH_4. Therefore, Ni-based catalysts often exhibit high CO_2 conversion, while their CO yield is typically low. These results indicate that appropriate combinations of metal elements are required to achieve high RWGS selectivity.

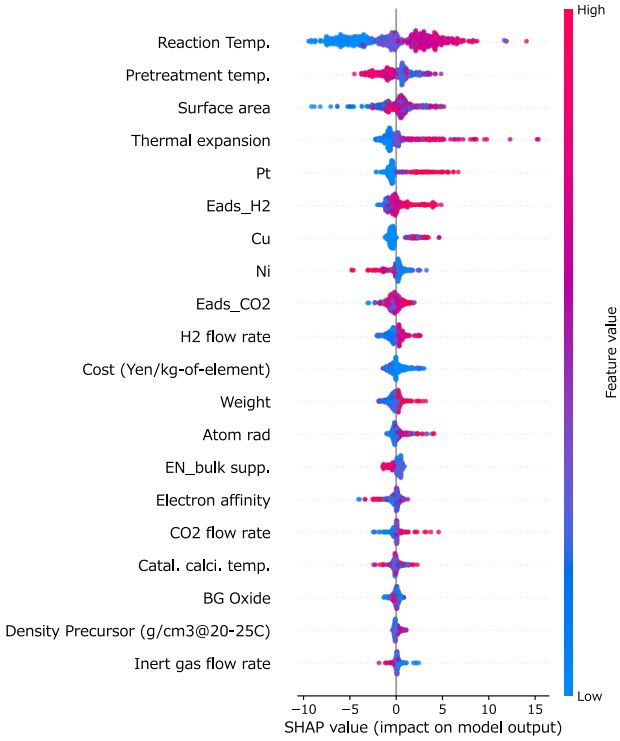

Fig. 2: The 20 descriptors with the greatest contributions in predicting the relative differences between CO and CH_4 production were calculated using XGB, based on the dataset. Here, the feature values are sorted in descending order according to the sum of the absolute SHAP values for global explanation.

In summary, we constructed a ML model based on a literature dataset for the RWGS reaction and achieved high predictive accuracy. Further analysis enabled us to identify and discuss key experimental parameters that strongly influence CO yield. These results provide useful guidelines for the rational design of high-performance RWGS catalysts.

In this presentation, we also report catalyst candidates predicted by the developed models and validation of their performance.

References

[1] Bahmanpour, A., et al., Applied Catalysis B: Environmental, 2021, 295(15), 120319.
[2] Wang, G., et al., Nature Communications, 2023. 14(1): p. 5861.
[3] Zhao, S., et al., Journal of the American Chemical Society, 2025. 147(34): p. 31121-31135.
[4] Kaiser, P., et al., Chemie Ingenieur Technik, 2013. 85(4): p.489-499.
[5] Kim, S.S., Applied Catalysis B: Environmental, 2012 , 119, p. 100-108.