ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Ensemble Model Approach for Predicting the Yield of Dehydrogenation Products during the Oxidative Dehydrogenation of *n*-Butane

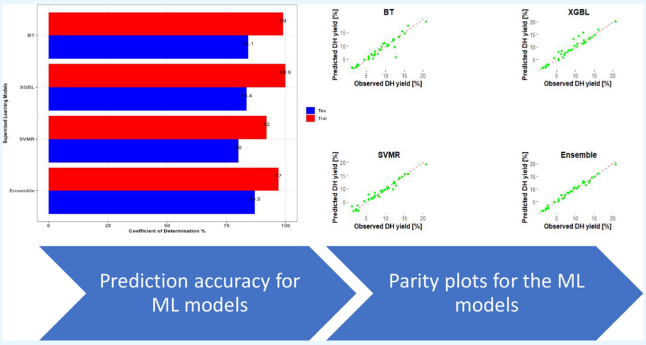
Gazali Tanimu,* Nurudeen A. Adegoke, Jimoh Olawale Ajadi, Yussif Yahaya, and Hassan Alasiri

Cite This: *ACS Omega* 2025, 10, 1198–1206

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Efficient and selective oxidative dehydrogenation (ODH) catalysts are crucial to advance the production of valuable petrochemicals. In this study, we leverage the power of machine learning to predict dehydrogenation (DH) product yield and unravel the factors influencing the product distribution. A comprehensive data set obtained from experiments conducted in a fixed-bed reactor under varying temperatures, feed ratios (O_2/*n*-butane), and metal oxide loadings (Ni, Fe, Co, Bi, Mo, W, Zn, and Mn) on an aluminum oxide support served as the basis for model development. Three supervised machine learning models, Boosted Tree (BT), Extreme Gradient Boosting Linear (XGBL), and Support Vector Machine Radial (SVMR), were evaluated. The ensemble technique of the three models showed remarkable accuracy, with an RMSE of 1.65 and MAE of 1.14 on the test data set, and it demonstrated robust generalization capabilities by capturing 87% of the variation in DH yield. In the feature importance analysis of the selected models, Mo, Co, Ni, and W emerged as critical factors influencing the DH yield. The practical significance of these findings lies in their potential to revolutionize catalysis research and industrial applications. The ability of the ensemble model to predict DH yields opens new avenues for optimizing DH products and designing more advanced catalysts. By providing essential insights into the influential variables governing the ODH reactions, researchers can make informed decisions to achieve higher yields and efficiencies.



INTRODUCTION

The demand for butenes and 1,3-butadiene, which are essential building blocks of the petrochemical industry, continues to grow. The conventional production route via direct dehydrogenation (DDH;eqs 1–5) of *n*-butane using platinum (Pt)-and chromium (CrO_x)-based catalysts[1−3] faces two major challenges: catalyst deactivation due to coke deposition and the need for high reaction temperatures to achieve reasonable conversions. To address these limitations and seek more efficient and sustainable alternatives, researchers have explored oxidative dehydrogenation (ODH) of *n*-butane as a promising alternative.

$$n - C_4H_{10} \rightarrow i - C_4H_8 + H_2 \quad (1)$$

$$n - C_4H_{10} \rightarrow cis - 2 - C_4H_8 + H_2 \quad (2)$$

$$n - C_4H_{10} \rightarrow trans - 2 - C_4H_8 + H_2 \quad (3)$$

$$n - C_4H_{10} \rightarrow 1 - C_4H_8 + H_2 \quad (4)$$

$$n - C_4H_{10} \rightarrow 1, 3 - C_4H_6 + 2H_2 \quad (5)$$

ODH represents a compelling approach, as it involves cofeeding *n*-butane with oxidants, such as O_2, CO_2, or N_2O, which has been shown to minimize catalyst deactivation and allow for operation at lower temperatures, thus enhancing the energy efficiency and prolonging the catalyst lifetime.[4−9] The typical reaction equations for *n*-butane ODH to 1,3-butadiene using cofed oxygen are presented in eqs 6–10. Despite the potential advantages of ODH, the commercialization of this process has been hindered primarily by the challenge of achieving the desired selectivity for the dehydrogenation (DH) products (butenes and 1,3-butadiene). The overoxidation of butane and intermediate butenes to stable oxidation products (CO_x) reduces the selectivity for the desired products.

$$n - C_4H_{10} + 1/2O_2 \rightarrow i - C_4H_8 + H_2O \quad (6)$$

Received: September 24, 2024
Revised: November 30, 2024
Accepted: December 12, 2024
Published: December 26, 2024

https://doi.org/10.1021/acsomega.4c08763

$$n - C_4H_{10} + 1/2O_2 \rightarrow \textit{cis} - 2 - C_4H_8 + H_2O \tag{7}$$

$$n - C_4H_{10} + 1/2O_2 \rightarrow \textit{trans} - 2 - C_4H_8 + H_2O \tag{8}$$

$$n - C_4H_{10} + 1/2O_2 \rightarrow 1 - C_4H_8 + H_2O \tag{9}$$

$$1 - C_4H_8 + 1/2O_2 \rightarrow 1, 3 - C_4H_6 + H_2O \tag{10}$$

Researchers have focused on the emerging field of catalyst informatics to overcome this critical challenge and accelerate the development of high-performance ODH catalysts. Catalyst informatics leverages data science techniques to design optimal catalysts for specific reactions and mechanisms.[10,11] It encompasses diverse aspects such as data set generation, data preprocessing, data visualization, and machine learning (ML). ML plays a pivotal role in identifying meaningful relationships between catalyst descriptors and performance metrics, thereby expediting the exploration of catalytic materials and their properties.[10,12] The use of ML in the field of heterogeneous catalysis dates back to the mid-1990s when it was first used to enhance the discovery of novel catalysts. Over the years, ML techniques have been increasingly applied to predict catalyst performance,[13] optimize reaction conditions through data-driven approaches,[14] and discover previously unknown catalysts,[15] particularly for reactions such as the oxidative coupling of methane (OCM). Recently, a comprehensive review on the predicting ability of machine learning models including predicting product yield, product compositions and product properties, with emphasis on the thermochemical treatment of biomass, was reported by Leng and his co-workers.[16] The review also highlighted various machine learning schemes and provided ways of enhancing the predictive performance, generalizability and application of ML.

Different groups explored the predicting ability of ML models for different catalytic and noncatalytic applications. For example, Kumar et al.[17] reported the use of different ML algorithms for predicting the binding energies of oxygen and carbon atoms on single atom alloys of Cu, Ag and Au, with gradient boosting regression (GBR) algorithm having the best prediction with an error of ~0.2 eV. Interestingly, combining the ML approach with ab initio microkinetic model (MKM) resulted in a higher turnover frequency for ethanol conversion during the nonoxidative ethanol dehydrogenation. Similarly, Yilmaz et al.[18] utilized random forest (RF) algorithm for predicting CO_2 conversion during methanation reaction, using 23 descriptors that consist of catalyst properties, reaction conditions and preparation methods. Excellent predictions were recorded with an RMSE of 12.7 and R^2 of 85% for the test data sets.

Madaan et al. pioneered the use of descriptor models based on radial distribution functions (RDF) to predict the performance of bimetallic mixed-oxide-supported catalysts for the *n*-butane ODH. Their approach demonstrated impressive accuracy, achieving a prediction accuracy of over 90% for a new set of bimetallic oxides.[19] In a recent study, our group extended this methodology to predict both the conversion of *n*-butane and selectivity for 1,3-butadiene during ODH by employing different ML algorithms, including linear and nonlinear methods. Among these algorithms, the support vector machine with radial basis function (SVMR) model has emerged as the top-performing predictor, yielding high coefficients of determination (R^2), low mean absolute errors (MAE), and root-mean-square errors (RMSE).[20]

Most recently, Liu et al.[21] utilized four different ML algorithms including artificial neural network (ANN), k-nearest neighbor (KNN), support vector regression (SVR) and random forest regression (RFR) to predict propylene space-time yield based on literature reported data for CO_2 assisted oxidative dehydrogenation of propane. RFR model outperformed the other models having the least RMSE value of 0.027 and highest R^2 of 81.8%. Notably, WHSV and temperature were reported as the most important features for the prediction based on SHAP analysis. Similarly, Roh et al.[22] developed eight ML algorithms including CatBoost regressor, Decision tree regressor, DNN, and RF regressor to predict the performance of 5655 data obtained from the literature for the DRM catalysts. CatBoost regressor model surpassed all the models with the highest prediction accuracy of 96% R^2 and 5.2663 RMSE. Likewise, Chen et al.[23] validated four ML models for predicting guaiacol conversion during the catalytic hydrodeoxygenation reaction. Data set were generated based on literature consisting varying reaction conditions and catalyst characteristics. Gradient Boosting Regression demonstrated superior performance with R^2 = 73−95%. Temperature and catalyst surface area were found to exert the most significant influence in the prediction based on permutation and SHAP feature importance analysis. All these studies have reported effective prediction using ML models with large data sets. However, one major drawback is the utilization of literature reported results for the model's development. Because most literature seldom report negative results, hence, the data set will contain majorly positive results thereby increasing bias coupled with ineffective understanding and generalization of the models.

In this study, we further advanced the application of ML to the ODH of *n*-butane by developing ML models to predict the yield of DH products. The novelty of this study lies in the utilization of the ensemble technique for the three developed models to further enhance their prediction ability, based on experimentally generated data set. Typically, the ODH reaction involves three competitive pathways: dehydrogenation (DH), cracking, and partial/complete oxidation. The motivation behind this study is 2-fold: first, to address the challenge of selectivity control and increase the understanding of catalyst design principles for desirable product formation (mainly the DH pathway), and second, to advance the field of catalyst informatics as a valuable and efficient tool for discovering high-performance ODH catalysts. Through a comprehensive analysis of ML model predictions, we aimed to gain insights into the factors influencing product yield, paving the way for future catalyst design and process optimization.

Methodology. Herein, we present a comprehensive data set derived from a series of experiments for the development of efficient and selective catalysts for the ODH of *n*-butane. The data set was collected from a fixed-bed flow reactor using catalysts prepared via coimpregnation and activated via a two-step calcination process. The calcination process involved two heating steps: the first at 350 °C for 1 h and the second at 590 °C for 2 h, with ramping rates of 10 and 15 °C/min, respectively. These experiments involved various combinations of Ni, Fe, Co, Mo, Zn, and W oxides supported on γ-Al_2O_3, with loadings ranging from 0 to 30 wt %. The reaction was tested using 300 mg of each catalyst, and temperatures ranging from 400 to 500 °C were explored, along with O_2/C_4 feed ratios of 1, 2, and 4 mol/mol. Details of the choice of metal oxides and supports, catalyst synthesis, characterization, and performance evaluation can be found in our previous studies.[7−9]

Each data set entry represents a specific combination of metal oxides on the Al_2O_3 support tested under particular temperature and O_2/C_4 feed ratio conditions. One noteworthy aspect of our data set was the inclusion of both positive and negative data. Positive data correspond to successful and selective catalysts, whereas negative data refer to results for nonselective catalysts. Including such diverse data types, which are often underrepresented in the literature, enhances the richness and complexity of the data set. This allows ML algorithms to better understand the underlying relationships and provides a basis for accurate learning and modeling. The primary focus of our study was to predict the DH yield as the outcome of interest. These models exploit the wealth of information in diverse data sets to identify the key features influencing yield and ultimately accelerate the discovery of high-performance catalysts for the ODH of *n*-butane.

Measurement and Calculation of Outcome Variables in ODH of *n*-Butane. This study determined the primary variable of interest, particularly DH yield, using an online gas chromatograph connected to a fixed-bed reactor. We utilized flame ionization and thermal conductivity detectors with various columns to detect the reaction products. The fundamental outcome metric, DH yield, was calculated from the Butane Conversion and DH selectivity based on carbon balance principles (eqs 1−3). Butane conversion (%) is the percentage of butane converted during the ODH reaction.

$$\begin{aligned}&\text{Butane Butane }(\%)\\&\quad=(\text{moles of butane fed}\\&\qquad-\text{ moles of butane in the product})\\&\qquad/(\text{moles of butane fed})\times 100\end{aligned}\tag{11}$$

The DH Selectivity (%) represents the percentage of the DH product formed relative to the extent of butane conversion.

$$\begin{aligned}&\text{DH selectivity }(\%)\\&\quad=(\text{moles of DH product})/(\text{Butane conversion})\times 100\end{aligned}\tag{12}$$

The DH yield indicates the overall efficiency of the ODH process in producing the desired DH product.

$$\begin{aligned}\text{DH yield }(\%)=&\ (\text{Butane conversion}\times\text{DH selectivity})\\&/100\end{aligned}\tag{13}$$

This outcome variable, namely the DH yield, is an essential indicator of the performance of the ODH reaction and plays a crucial role in understanding the activity and selectivity of the catalysts under investigation. By accurately measuring and calculating this metric, we evaluated the effectiveness of different catalyst formulations and optimized the design of high-performance catalysts for ODH of *n*-butane.

Descriptor Variables Influencing the Performance of ODH Catalysts. We explored a carefully selected set of descriptors that significantly influenced the performance of the ODH catalysts. These descriptors encompass temperature (ranging from 400 to 500 °C), the feed ratio of O_2 to *n*-butane (ranging from 1 to 4 mol/mol), and varying loadings (0−30 wt %) of several metal oxides, including nickel (Ni; 0−20 wt %), iron (Fe; 0−20 wt %), cobalt (Co; 0−20 wt %), bismuth (Bi; 0−30 wt %), molybdenum (Mo; 0−15 wt %), tungsten (W; 0−30 wt %), zinc (Zn; 0−20 wt %), and manganese (Mn; 0−20 wt %). They play a pivotal role in shaping the outcomes of catalytic processes. The temperature and feed ratio directly affected the degree of feed conversion and selectivity toward the desired and undesired products. Higher temperatures and feed ratios tended to favor cracking and deep oxidation products. In contrast, an optimal temperature and moderate feed ratio facilitated the DH pathway, leading to the desired product formation.[19]

Furthermore, the metal oxide composition and loading in the catalyst exert a profound influence on its properties, such as acidity, reducibility, nature of the oxygen species, and the interaction between the metal(s) and support.[24] These factors, in turn decisively affect the selectivity toward the desired product, ultimately determining the effectiveness of the catalyst in the ODH reaction.[25,26] By exploring and understanding the role of these descriptor variables, we aimed to unlock insights into the intricate mechanisms governing the ODH reactions. This knowledge is instrumental in guiding the rational design of high-performance catalysts for this industrially significant process, paving the way for more efficient and sustainable petrochemical production in the future.

■ STATISTICAL ANALYSIS

Data Exploration. The data set used in this study was derived from experimental testing conducted in a fixed-bed reactor. It comprises 11 variables, with one being the response variable (DH yield) and the remaining ten being predictor variables. The predictor variables included temperature (measured in degrees Celsius), feed ratio (O_2/*n*-butane), and various loadings (ranging from 0 to 30 wt %) of metal oxides, specifically Ni, Fe, Co, Bi, Mo, W, Zn, and Mn. Throughout the experiments, 185 observations were collected for each variable to provide a comprehensive data set for the analysis. This extensive data collection enabled a thorough exploration of the behavior of the fixed-bed reactor and its performance under different conditions.

Our initial analysis started with data exploration to comprehensively understand the data set and identify patterns and relationships among the variables. Two primary methods, namely, the correlation plot and pair plot, were employed for this purpose. The correlation plot provides significant insights into the relationships between the predictor variables. This allowed us to visualize both positive and negative associations among the variables. By examining the correlation coefficients, we identified strongly correlated variables that showed weak or negligible correlations. This information is crucial for understanding the interdependencies among predictor variables and for assessing potential multicollinearity issues.

In addition, a pair plot provided a detailed overview of the distributions and relationships between the selected variables (Figure S1). Each subplot on the diagonal presents a histogram of a single variable, providing a visual representation of its distribution. Additionally, the off-diagonal plots consisted of scatter plots between pairs of variables, revealing the potential correlations between them. Through this visualization, we gained insight into the nature of the relationships and the presence of linear or nonlinear associations. By conducting thorough data exploration, we were able to identify significant trends and patterns within the data set. These insights were instrumental in guiding subsequent analyses and model

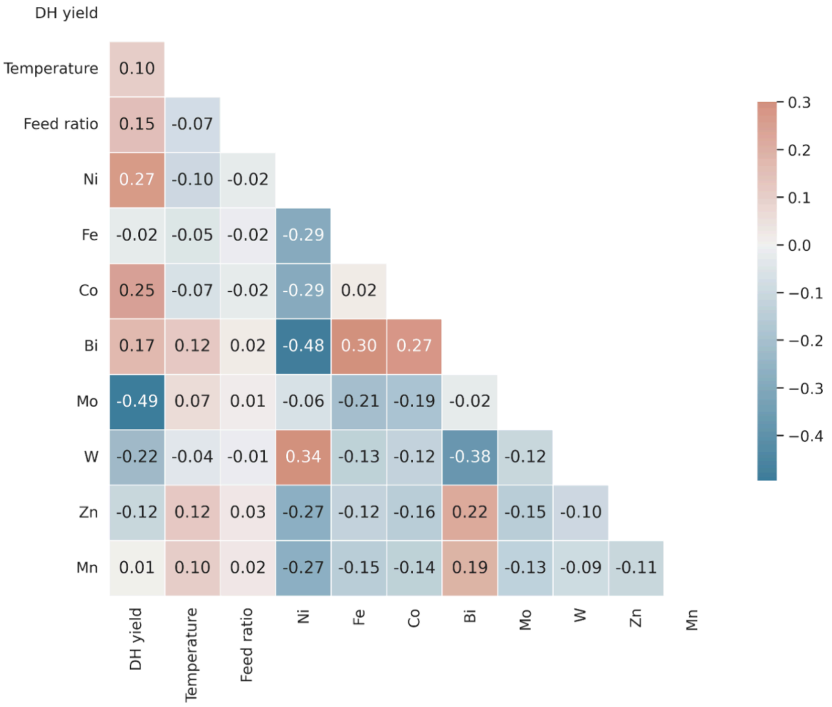


Figure 1. Heatmap of the lower triangular correlation matrix for predictor and outcome variables. Nickel (Ni; 0−20 wt %), iron (Fe; 0−20 wt %), cobalt (Co; 0−20 wt %), bismuth (Bi; 0−30 wt %), molybdenum (Mo; 0−15 wt %), tungsten (W; 0−30 wt %), zinc (Zn; 0−20 wt %), and manganese (Mn; 0−20 wt %).

development, ensuring that DH could be accurately predicted based on predictor variables. The data exploration process laid the foundation for a deeper understanding of the factors that influence the DH yield.

Model Development. This study aimed to utilize an ensemble model to predict the DH yield by integrating three well-established supervised ML models: Boosted Tree (BT),[27] Extreme Gradient Boosting Linear (XGBL),[28] and Support Vector Machine Radial (SVMR).[29] Brief discussions regarding the models are provided in the Supporting Information. The process involves developing individual models for each algorithm and conducting nested cross-validation for optimal parameter tuning and feature selection. An ensemble model was formed by averaging the predictions of the three individual models. The training data set, comprising 80% of the total data, was employed for model training, whereas the remaining 20% served as a testing data set for performance evaluation. Evaluation metrics, including R^2, RMSE, and MAE, were applied to assess model performance. The models were built using the caret package (version 6.0.94) in R software (version 4.31).[30,31] The mathematical expressions for the metrics R^2, RMSE, and MAE are provided in the Supporting Information.

Data Scaling. Data scaling was performed to ensure that the ML models were not biased toward specific features owing to the magnitude differences. Continuous variables in the training set were transformed using the z-score algorithm, standardizing the data to have a mean of 0 and a standard deviation of 1. This transformation facilitates fair comparison and analysis of the features. The test set variables were scaled similarly to ensure that the models were evaluated on data with consistent scaling, promoting better performance and training stability, particularly for sensitive models, such as support vector machines.

Model Cross-Validation. To assess the generalization performance of the model and address overfitting, a robust cross-validation strategy was employed. For this purpose, a nested cross-validation method was used. The data set was divided into ten equally sized subsets, or folds, for the outer loop. Within each fold, further subsetting occurred for the inner loop of the nested cross-validation process. In the inner loop, the model parameters were tuned to identify the optimal configuration, which was then used in the outer loop for the model development. Throughout each iteration, the models underwent training on specific training subsets and were subsequently evaluated on the validation sets to comprehensively assess their performance. This iterative process aids in obtaining reliable performance estimates by minimizing the impact of random partitioning. The use of ten folds and nested structures contributes to a thorough evaluation, enhancing the robustness of the cross-validation approach.

Optimal Feature Selection and Hyperparameters. Backward elimination (BE) was used to identify the most relevant features. The least significant variable was removed iteratively based on the R^2, leading to a subset of features with the greatest impact on DH yield, enhancing model simplicity and interpretability. Grids of the hyperparameters were generated for each algorithm, and different configurations were explored. The wrapper selection identified the optimal model based on the R^2, and the selected features with the best-performing model were recorded. The hyperparameter results are shown in Table S1.

■ RESULTS AND DISCUSSION

Descriptive Statistics. This section details the descriptive statistical analysis of the data sets. Figure 1 presents a heatmap of the lower triangular correlation matrix for both predictor and outcome variables. This visualization offers a comparative analysis of the linear relationships among the descriptor variables as well as between the descriptor variables and the response variable (DH yield). Each cell in the heatmap represents the correlation coefficient between two variables, ranging from −1 (perfect negative correlation, shown in blue) to +1 (perfect positive correlation, shown in red). Values close to 0 (white color) indicate little to no linear relationship. Ideally, we aim for the DH yield to have a stronger correlation with the descriptor variables while seeking less correlation among the descriptor variables.

As shown in Figure 1, moderate positive correlations were observed between DH yield and the descriptors Ni (0.27), Co (0.25), and Bi (0.17). These findings suggest that augmentation of these variables could potentially increase DH yield. A similar yet weaker positive correlation was found with the feed ratio and temperature variables, suggesting that although these may affect DH yield, they are not the predominant influencers. In contrast, the Mn and Fe parameters exhibited almost negligible correlations with the DH yield variable, implying an insignificant linear association. However, the Zn and W parameters exhibited weak negative correlations with DH yield, suggesting a potential decrease in DH yield with an increase in these variables. Notably, Mo was found to have a strong negative correlation with DH yield, indicating a significant inverse relationship.

Furthermore, we observed positive and negative associations with varying correlation coefficients, indicating distinct dynamics among the predictor variables. Notably, the feed ratio and W correlation were minimal (coefficient: −0.01), suggesting a negligible negative relationship. In contrast, Ni and Bi displayed a relatively strong negative correlation (coefficient: −0.5), indicating an inverse relationship between these variables. In general, the data set used did not show any significant multicollinearity among the predictor variables. We determined this through the correlation matrix analysis, which indicated low correlation coefficients between the predictors. Multicollinearity, which involves strong linear relationships between predictors, can affect the interpretability and stability of regression models. It is worth noting that we did not encounter any issues with multicollinearity, and the regression models we used (BT, XGBL, and SVMR) were not significantly affected by it. The absence of multicollinearity issues and the ability of the models to withstand them ensure the reliability of the predictor variables in our regression models for making predictions. The Pairwise relationships and distributions of the selected variables are shown in Figure S1 of the Supporting Information.

Model Performance. Table 1 and Figure 2 present a comprehensive overview of the performance of the individual models and the ensemble method on both the training and test data sets. For the training data, the BT model achieved an RMSE of 0.39 and MAE of 0.25 for the training data, along with an RMSE of 1.78 and MAE of 1.27 for the test data. Notably, R^2 values of 99.4 and 85% were obtained for the training and test data sets, respectively. For the XGBL model, the training data showed an RMSE of 0.05 and MAE of 0.04, whereas the test data exhibited an RMSE of 1.87 and MAE of 1.30. The model achieved R^2 values of 99.9% for the training data set and 83.4% for the test data set. In the case of SVMR, the training data yielded an RMSE of 1.45 and MAE of 0.95, whereas the test data reported an RMSE of 2.05 and MAE of 1.38. Interestingly, the ensemble of the three models demonstrated an RMSE of 0.56 and MAE of 0.37 for the training data, and an RMSE of 1.65 and MAE of 1.14 for the test data. Remarkably, the R^2 values of 97 and 86.9% for the training and test data sets, respectively. An RMSE of 0.56, on average, suggests that the ensemble method predictions deviate from the actual dehydrogenation yields by approximately 0.56 percentage, whereas an R^2 of 97% suggests that the features employed in the study within the ensemble method's prediction of dehydrogenation yield can explain 97% of the variability observed in the actual dehydrogenation yields.

Table 1. Performance Evaluation of the Individual Models and the Ensemble for Predicting DH Yield

	RMSE		MAE	
	train	test	train	test
BT	0.39	1.78	0.25	1.27
XGBL	0.05	1.87	0.04	1.30
SVMR	1.45	2.05	0.95	1.38
ensemble	0.56	1.65	0.37	1.14

In summary, the results indicate exceptional performance from each individual model, with the ensemble of the three models showing superior outcomes for the test data, achieving the lowest test RMSE and MAE, coupled with the highest R^2.

Figure 3 provides a visual representation of the feature importance in predicting the DH yield across the various models. Feature importance analysis was used to identify the critical variables for DH yield prediction within each model. In the BT model (Figure 3a), Mo loading emerged as the most crucial factor, and Co loading also showed a significant influence. Zn and Fe loadings had relatively less impact on the DH yield predictions. In the XGBL model (Figure 3b), similarly, Mo loading was the most important variable, followed by Ni and W loadings. Mn and Fe loadings were the least significant predictors of DH yield. Similarly, in the SVMR model (Figure 3c), Mo loading was highly significant in predicting the DH yield. Conversely, Zn and Fe loadings had comparatively less of an impact. Mo loading consistently appeared as the most significant feature in all models. This feature importance analysis offers valuable insights into the key factors that influence the DH yield. The selected features and those that were not selected are listed in Supplementary Table S2.

Figure 4 illustrates parity plots, visually comparing the predicted DH yield values with actual observed values for individual models (BT, XGBL, and SVMR) and the ensemble method. Parity plots offer a holistic assessment of the predictive accuracy, with each data point representing a predicted DH yield value plotted against its corresponding actual value. In the case of the ensemble method, a notable concentration of data points was observed around the equality line, indicating strong agreement between the ensemble method predictions and actual values. This clustering underscores the capability of the ensemble method to capture the underlying patterns and trends in the DH yield data set accurately. Although some deviations from the equality line exist, the overall proximity suggests the ensemble method's reliable and consistent predictive performance.

The parity plot also revealed that individual models demonstrated close alignment between the predicted and observed DH yield values. Despite occasional deviations, the overall pattern along the equality line signifies reasonable predictive capabilities based on specified predictor variables.

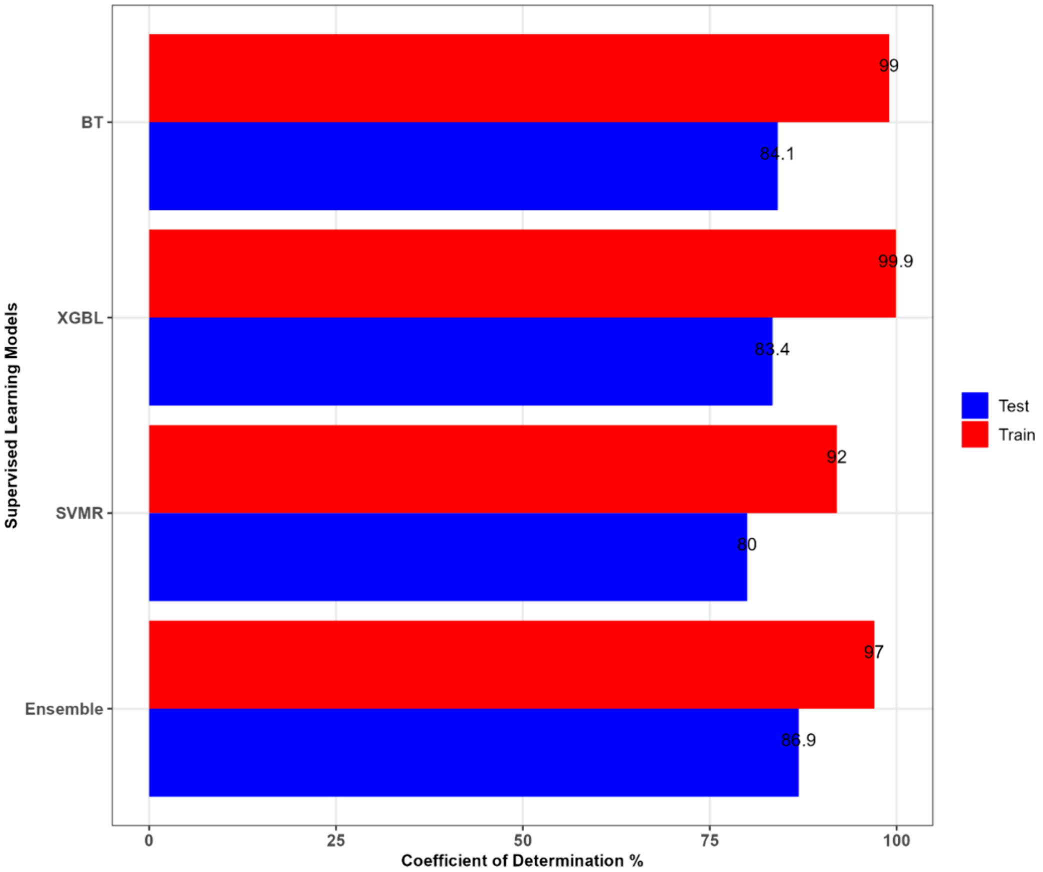


Figure 2. Coefficient of determination (R^2) for supervised ML models.

Collectively, parity plots visually confirmed the predictive performance of BT, XGBL, SVMR models, and their ensemble method for DH yield. These visualizations validate the ability of the models to approximate true values and offer insights into their accuracy and reliability in real-world scenarios. In particular, the ensemble method is the most effective choice for predicting the DH yield, evident in its consistent alignment with the equality line and minimal prediction errors. The practical use of the ensemble method lies in its ability to predict the DH yield accurately, as demonstrated by the close alignment between the predicted and observed values in the parity plots. This robust predictive capability translates into real-world benefits, where process optimization can be achieved more confidently. Accurate predictions can guide researchers and practitioners in setting optimal process conditions and designing tailored catalyst compositions. Consequently, integrating ML models into catalysis research offers a pathway to accelerate the discovery and optimization of high-performance catalysts. The implications of this study extend beyond the laboratory setting to industrial applications. Companies in the chemical and petrochemical sectors can enhance their catalytic processes by deploying ensemble methods or similar ML approaches. An improved DH yield not only leads to higher product yields but also minimizes waste and reduces the environmental impact of these processes. Ultimately, this study will contribute to the development of sustainable and eco-friendly industrial practices.

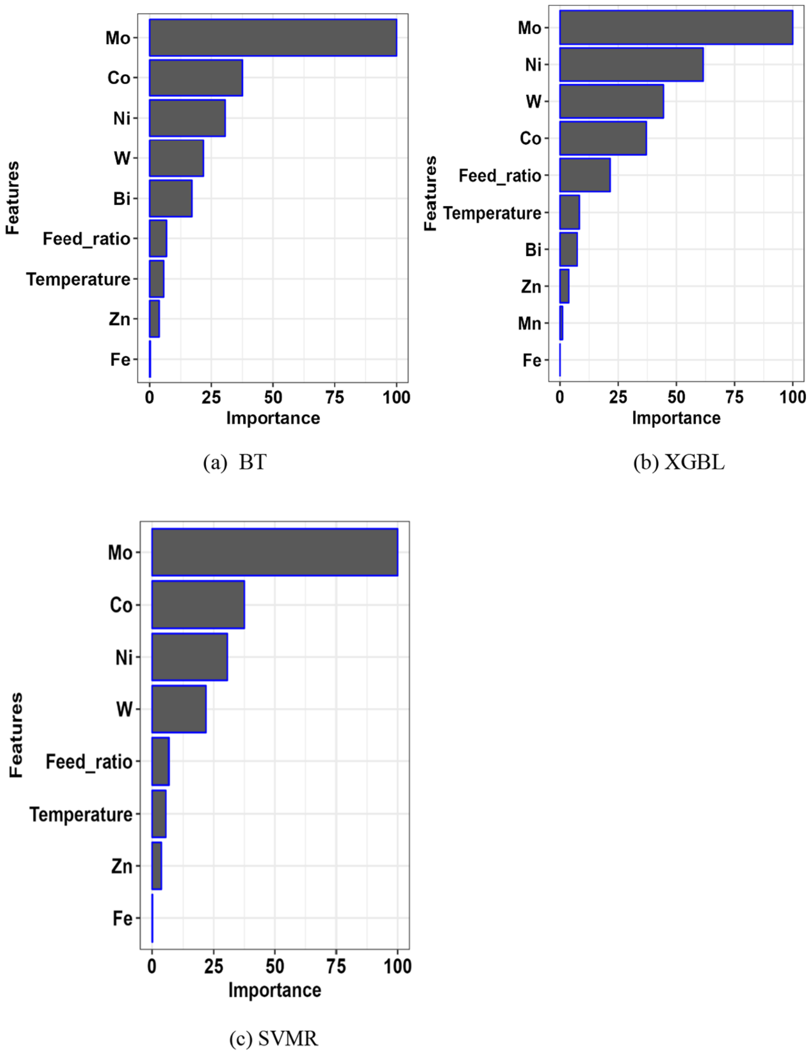
It is important to acknowledge the limitations of this study. Although the Ensemble method demonstrated remarkable predictive performance, the choice of other ML models and tuning approaches may yield different results. Careful consideration of the specific applications and data set characteristics is essential to ensure that the most suitable model is selected. In addition, the experimental data used in this study were obtained from a controlled fixed-bed reactor. Real-world industrial processes may involve more complex and dynamic conditions that could influence the predictive accuracy of the models. Further validation and refinement of the models using data from pilot- or industrial-scale reactors would enhance their applicability and reliability in practical scenarios.

Model Comparison with Literature. This section compares the model performance results obtained with the different models and ensemble method together with other ML models reported in the literature for various catalytic applications. Table 2 presents the results in terms of models employed, catalytic application and performance metrics.

It is clear from Table 2 that the ML models perform reasonably well in predicting the various output features. Similarly, the models developed in this work fall within the limits of the high R^2 and low RMSE values. It is important to note that, the data source (experimental or literature), nature of the data set (input and output features), data preprocessing and hyperparameter tuning of the models, all contribute in the overall model performance. Hence, there is no best model that is suitable for all catalytic application.

■ CONCLUSIONS AND PERSPECTIVES

The findings of this study underscore the practical significance of employing ML models, particularly their ensemble



(a) BT

(b) XGBL

(c) SVMR

Figure 3. Feature importance analysis of different models. [Nickel (Ni; 0−20 wt %), iron (Fe; 0−20 wt %), cobalt (Co; 0−20 wt %), bismuth (Bi; 0−30 wt %), molybdenum (Mo; 0−15 wt %), tungsten (W; 0−30 wt %), zinc (Zn; 0−20 wt %), and manganese (Mn; 0−20 wt %)].

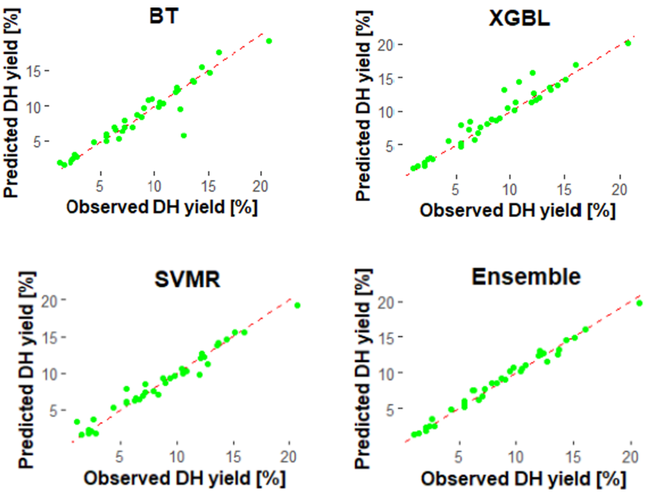


Figure 4. Parity plots for BT, XGBL, SVMR, and ensemble models.

technique, in oxidative dehydrogenation (ODH) reactions. The successful application of the ensemble method to predict the DH yield holds immense promise for catalysis research and industrial applications. By accurately predicting the DH yield, this method offers a valuable tool for optimizing the ODH process and catalyst design, with real-world implications for enhancing the process efficiency and sustainability. One of the key insights from feature importance analysis is the identification of critical predictor variables that influence DH yield. Notably, Mo, Co, and Ni were identified as significant factors influencing the DH yield in the BT model. In the XGBL model, Mo, Ni, and W are the most significant factors. Similarly, for the SVMR model, Mo, Co, and Ni were recognized as the most important factors in predicting DH yield. Understanding the impact of these variables provides essential contextual meaning as it sheds light on the underlying mechanisms governing the ODH reaction. Such insights will enable researchers and process engineers to make informed decisions regarding the design and operation of catalytic systems for maximum DH yield. Moreover, this knowledge contributes to the exploration of novel catalyst materials and development of more efficient and sustainable DH processes.

In summary, this study demonstrated the practical potential of ML models, particularly the ensemble method, for catalysis research and industrial applications. The ability to accurately predict DH yield and identify critical predictor variables provides a powerful tool for enhancing process efficiency and

Table 2. Performance Comparison of Various ML Models in Catalyst Prediction

s/no	catalytic application	ML models	performance metrics R^2 (%)	RMSE	ref
1	CO_2 methanation (predicting CO_2 conversion)	random forest	85	12.7	Yilmaz et al.[16]
2	CO_2 oxidative dehydrogenation of propane (propylene yield prediction)	artificial neural network	6.2	0.14	Liu et al.[21]
		support vector regression	39.4	0.09	
		K-nearest neighbor	76.5	0.03	
		random forest	81.8	0.03	
3	dry reforming of methane (CH_4 conversion prediction)	CatBoost regressor	96	5.27	Roh et al.[22]
		XGBoost regressor	95	6.10	
		random forest	90	8.45	
		decision tree	71	14.19	
		deep neural network	54	17.92	
		Gaussian process regressor	34	21.25	
		support vector machine	15	28.14	
4	CO_2 hydrogenation to methanol (methanol yield prediction)	XGBoost	88	0.09	Suvarna et al.[32]
		random forest	84	0.08	
		Gradient Boosting Decision tree	82	0.09	
5	oxidative dehydrogenation of *n*-butane (predicting yield of dehydrogenation products)	Boosted Tree	84.1	1.78	this work
		XGBoost linear	83.4	1.87	
		Support Vector Machine Radial	80	2.05	
		ensemble model	86.9	1.65	

optimizing catalyst design, and we continue to explore the synergies between experimental data and ML techniques. The field of catalysis benefits from accelerated catalyst discovery, improved process efficiency, and a more sustainable approach to chemical production. Integrating ML approaches into catalysis research is not just a technological advancement, but a transformative shift toward a greener and more efficient future for industrial processes.

Future research directions worth investigating include combining catalyst physicochemical properties with catalyst compositions and reaction conditions to capture more information relevant for robust ML prediction and validation. Also, explorative ML using atomic properties generated using DFT calculations with correlations to the catalyst performance is an interesting area that will facilitate the development of new and more promising catalysts. Bayesian optimization also holds promise in facilitating the development of selective catalysts, especially after successful cycles of prediction and experimental validations. Finally, the use of large language models (LLM) and generative AI will speed-up new catalyst development and will improve the chances of commercializing various applications that rely solely on the development of highly active, selective and stable catalysts.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c08763.

Pairwise relationship of features, data distribution of variables, model descriptions, model performance, and Shapley Additive exPlanations (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Gazali Tanimu − *Interdisciplinary Research Center for Refining & Advanced Chemicals and Department of Chemical Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia;* orcid.org/0000-0002-5019-0151; Email: gazali.tanimu@kfupm.edu.sa

Authors

Nurudeen A. Adegoke − *Department of Statistics, The Federal University of Technology Akure, Akure PMB 704 Ondo State, Nigeria*

Jimoh Olawale Ajadi − *Interdisciplinary Research Center for Refining & Advanced Chemicals and Department of Mathematics, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; Department of General Sciences, Deanship of Support Studies, Alasala Colleges, Dammam 32324, Saudi Arabia*

Yussif Yahaya − *Department of Mathematics, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia*

Hassan Alasiri − *Interdisciplinary Research Center for Refining & Advanced Chemicals and Department of Chemical Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia;* orcid.org/0000-0003-4043-5677

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c08763

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the support received from the Interdisciplinary Research Center for Refining and Advanced Chemicals (IR-CRAC) at KFUPM, Saudi Arabia under the funded project INRC 2205.

REFERENCES

(1) Burrington, J. D.; Grasselli, R. K. Aspects of selective oxidation and ammoxidation mechanisms over bismuth molybdate catalysts. *J. Catal.* **1979**, *59*, 79−99.

(2) Bender, M. An Overview of Industrial Processes for the Production of Olefins−C_4 Hydrocarbons. *Chem. Bio. Eng. Rev.* **2014**, *1*, 136−147.

(3) Ajayi, B. P.; Jermy, B. R.; Ogunronbi, K. E.; Abussaud, B. A.; Al-Khattaf, S. n-Butane dehydrogenation over mono and bimetallic MCM-41 catalysts under oxygen free atmosphere. *Catal. Today.* **2013**, *204*, 189−196.

(4) Jung, J. C.; Kim, H.; Kim, Y. S.; Chung, Y.-M.; Kim, T. J.; Lee, S. J.; Oh, S.-H.; Song, I. K. Catalytic performance of bismuth molybdate catalysts in the oxidative dehydrogenation of C_4 raffinate-3 to 1,3-butadiene. *Appl. Catal. A: Gen.* **2007**, *317*, 244−249.

(5) Gottifredi, J. C.; Sham, E. L.; Murgia, V.; Farfa, E. M. Sol−gel synthesis of V_2O_5−SiO_2 catalyst in the oxidative dehydrogenation of *n*-butane. *Appl. Catal., A* **2006**, *312*, 134−143.

(6) Rossetti, I.; Mancini, G. F.; Ghigna, P.; Scavini, M.; Piumetti, M.; Bonell, B.; Cavani, F.; Comite, A. Spectroscopic Enlightening of the Local Structure of VO_X Active Sites in Catalysts for the ODH of Propane. *J. Phys. Chem. C* **2012**, *116* (42), 22386−22398.

(7) Jermy, B. R.; Ajayi, B. P.; Abussaud, B. A.; Asaoka, S.; Al-Khattaf, S. Oxidative dehydrogenation of n-butane to butadiene over Bi-Ni-O/γ-alumina catalyst. *J. Mol. Catal. A Chem.* **2015**, *400*, 121−131.

(8) Jermy, B. R.; Asaoka, S.; Al-Khattaf, S. Influence of calcination on performance of Bi-Ni-O/gamma-alumina catalyst for n-butane oxidative dehydrogenation to butadiene. *Catal. Sci. Technol.* **2015**, *5* (9), 4622−4635.

(9) Tanimu, G.; Jermy, B. R.; Asaoka, S.; Al-Khattaf, S. Composition effect of metal species in (Ni, Fe, Co)-Bi-O/gamma-Al_2O_3 catalyst on oxidative dehydrogenation of n-butane to butadiene. *J. Ind. Eng. Chem.* **2017**, *45*, 111−120.

(10) McCullough, K.; Williams, T.; Mingle, K.; Jamshidi, P.; Lauterbach, J. High-throughput experimentation meets artificial intelligence: a new pathway to catalyst discovery. *Phys. Chem. Chem. Phys.* **2020**, *22*, 11174−11196.

(11) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403−7429.

(12) Takahashi, K.; Takahashi, L.; Miyazato, I.; Fujima, J.; Tanaka, Y.; Uno, T.; Satoh, H.; Ohno, K.; Nishida, M.; Hirai, K.; Ohyama, J.; Nguyen, T. N.; Nishimura, S.; Taniike, T. The Rise of Catalyst Informatics: Towards Catalyst Genomics. *ChemCatChem.* **2019**, *11*, 1146−1152.

(13) Nishimura, S.; Ohyama, J.; Kinoshita, T.; Le, S. D.; Takahashi, K. Revisiting Machine Learning Predictions for Oxidative Coupling of Methane (OCM) based on Literature Data. *ChemCatChem* **2020**, *12*, 5888−5892.

(14) Ohyama, J.; Nishimura, S.; Takahashi, K. Data Driven Determination of Reaction Conditions in Oxidative Coupling of Methane via Machine Learning. *ChemCatChem.* **2019**, *11*, 4307−4313.

(15) Takahashi, K.; Miyazato, I.; Nishimura, S.; Ohyama, J. Unveiling Hidden Catalysts for the Oxidative Coupling of Methane based on Combining Machine Learning with Literature Data. *ChemCatChem.* **2018**, *10*, 3223−3228.

(16) Li, H.; Chen, J.; Zhang, W.; Zhan, H.; He, C.; Yang, Z.; Peng, H.; Lang, L. Machine Learning Aided Thermochemical Treatment of Biomass: A Review. *Biofuel Res. J.* **2023**, *10*, 1786−1809.

(17) Kumar, A.; Iyer, J.; Jalid, F.; Ramteke, M.; Khan, T. S.; Haider, M. A. Machine Learning Enabled Screening of Single Atom Alloys: Predicting Reactivity Trend for Ethanol Dehydrogenation. *ChemCatChem* **2022**, *14*, No. e202101481.

(18) Yilmaz, B.; Oral, B.; Yildirim, R. Machine Learning Analysis of Catalytic CO_2 Methanation. *Int. J. Hydrogen Energy* **2023**, *48* (64), 24904−24914.

(19) Madaan, N.; Shiju, N. R.; Rothenberg, G. Predicting the performance of oxidation catalysts using descriptor models. *Catal. Sci. Technol.* **2016**, *6*, 125−133.

(20) Tanimu, G.; Ajadi, J. O.; Yahaya, Y.; Alasiri, H.; Adegoke, N. A. Developing Machine Learning Models for Catalysts in Oxidative Dehydrogenation of n-butane. *ChemCatChem* **2023**, *15*, No. e202300598.

(21) Liu, H.; Liu, K.; Zhu, H.; Guo, W.; Li, Y. Explainable Machine-Learning Predictions for Catalysts in CO_2-assisted Propane Oxidative Dehydrogenation. *RSC Adv.* **2024**, *14*, 7276−7282.

(22) Roh, J.; Park, H.; Kwon, H.; Joo, C.; Moon, I.; Cho, H.; Ro, I.; Kim, J. Interpretable Machine Learning Framework for Catalyst Performance Prediction and Validation with Dry Reforming of Methane. *Appl. Catal. B: Environ.* **2024**, *343*, No. 123454.

(23) Chen, X.; Shafizadeh, A.; Shahbeik, H.; Rafiee, S.; Golvirdizadeh, M.; Moradi, A.; Peng, W.; Tabatabaei, M.; Aghbashlo, M. Machine Learning-based Optimization of Catalytic Hydrodeoxygenation of Biomass Pyrolysis Oil. *J. Clean. Prod.* **2024**, *437*, No. 140738.

(24) Madeira, L. M.; Portela, M. F. Catalytic Oxidative Dehydrogenation of n-butane. *Catal. Rev.: Sci. Eng.* **2002**, *44* (2), 247−286.

(25) Gambo, Y.; Adamu, S.; Tanimu, G.; Abdullahi, I. M.; Lucky, R. A.; Ba-Shammakh, M. S.; Hossain, M. M. CO_2-mediated oxidative dehydrogenation of light alkanes to olefins: Advances and perspectives in catalyst design and process improvement. *Appl. Catal. A: Gen.* **2021**, *623*, No. 118273.

(26) Tanimu, G.; Asaoka, S.; Al-Khattaf, S. Effect of support in Ni-Bi-O/support catalyst on oxidative dehydrogenation of *n*-butane to butadiene. *Mol. Catal.* **2017**, *438*, 245−255.

(27) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29* (5), 1189−1232.

(28) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *International Conf. on Knowledge Disco & Data Mining*; ACM: 2016; pp 785−794.

(29) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Computational Learning Theory (COLT'92)*; ACM: 1992; pp 144−152.

(30) Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Soft.* **2008**, *28* (5), 1−26.

(31) *R Core Team _R: A Language and Environment for Statistical Computing_*; R Foundation for Statistical Computing: Vienna, Austria, 2023. https://www.R-project.org/.

(32) Suvarna, M.; Araujo, T. P.; Perez-Ramirez, J. A Generalized Machine Learning Framework to Predict the Space-time Yield of Methanol from Thermocatalytic CO_2 Hydrogenation. *Appl. Catal. B: Environ.* **2022**, *315*, No. 121530.