# Statistical Analysis and Discovery of Heterogeneous Catalysts Based on Machine Learning from Diverse Published Data

Keisuke Suzuki,[a] Takashi Toyao,[b,c] Zen Maeno,[c] Satoru Takakusagi,[c]

Ken-ichi Shimizu,*[b,c] Ichigaku Takigawa*[d, e]

[a] Mr. K. Suzuki
Graduate School of Information Science and Technology, Hokkaido University, N-14, W-9, Sapporo 001-0021 (Japan)

[b] Dr. T. Toyao, Dr. Z. Maeno, Dr. S. Takakusagi, Prof. Dr. K. Shimizu
Institute for Catalysis, Hokkaido University, N-21, W-10, Sapporo 001-0021 (Japan)
E-mail: kshimizu@cat.hokudai.ac.jp

[c] Dr. T. Toyao, Prof. Dr. K. Shimizu
Elements Strategy Initiative for Catalysis and Batteries, Kyoto University, Katsura, Kyoto 615-8520 (Japan)

[d] Dr. I. Takigawa
RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027 (Japan)
E-mail: ichigaku.takigawa@riken.jp

[e] Dr. I. Takigawa
Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, N-21, W-10, Sapporo 001-0021 (Japan)

## Abstract

The literature provides insights for catalyst design and discovery. Effective analysis of reported data using machine learning (ML) methods offers the ability to gain valuable information. However, utilizing the literature in this way has obstacles such as lack of compositional overlaps, bias from prior published data, and low sample counts for many elements. The present study describes an ML approach that considers elemental features as input representations instead of inputting catalyst compositions directly. This ML method has the potential for catalyst discovery, including catalytic reactions with limited catalyst composition overlap in the available data. Oxidative coupling of methane (OCM), water gas shift (WGS), and CO oxidation reactions were chosen to confirm the effectiveness of the proposed method by analysis using several state-of-the-art ML methods. Among the ML methods tested, gradient boosting regression with XGBoost (XGB) provided the best results, and prediction accuracy was improved by the proposed approach for all three reaction types. In addition, a quantitative value of “feature importance score” was calculated to evaluate the most influential input variables on catalyst performance. Finally, catalyst optimization was explored using ML as a “surrogate” model, and the top 20 promising candidate catalysts were identified for the OCM reaction based on the optimization. The advantages of ML in catalysis analysis as well as the difficulties and limitations originating from the complexity of heterogeneous catalysis were explored.

## Introduction

Catalysis is a highly complex phenomenon.[1] In particular, heterogeneous catalysis remains an empirical science due to the complexity of the surface reactions involved.[2,3] Recent experimental and theoretical studies have produced insights on the atomic level.[4–6] However, discovering new and interesting catalysts remains a formidable task. High-throughput or combinatorial techniques, which have been applied successfully to relevant fields such as materials synthesis and homogeneous catalysis, should serve as powerful tools for the discovery of novel heterogeneous catalysts and catalytic processes.[7–11] Data- and text-mining approaches also show promise for accelerating research.[12,13] For these approaches, refined and automatic search strategies to minimize the number of experiments during exploration are required and need to be a key aspect of the approach. Evolutionary computation (such as generic algorithms) often are used for these statistical approaches.[14–20] However, they normally require a large amount of data and computational time to explore. In this sense, machine learning (ML) techniques, which require a relatively small amount of data and are less computationally expensive, should be effective for searches, especially in fields having little information.

ML methods have gained attention in molecular and materials science fields to predict various physical and chemical properties.[21–23] These methods can serve as fast and high-precision alternatives to first-principles modelling. Several successful examples have already been reported for organic chemistry reactions, including those that involve homogeneous catalysts.[24–29] However, the applicability of ML predictions for heterogeneous catalysis have been limited mainly to computationally determined values such as band gaps,[30–32] d-band centers,[33,34] and adsorption energies.[35–40] For the practical use of ML for discovering new solid catalytic materials, not only first-principles calculated values but also experimental values for specific catalytic reactions are needed, especially in heterogeneous catalysis because an adequate theoretical model for heterogeneous catalysis is not available. Thus, the ML predicted values could not directly lead to novel catalyst designs. Although reports based on ML predictions of experimental results of heterogeneous catalysis are limited in number, some examples are available.[41–48] For example, catalysts for oxidative coupling of methane (OCM) to $C_2$ products such as $C_2H_4$ and $C_2H_6$ have been predicted.[47,48] Yildirim *et al*. analyzed results of various catalytic reactions obtained from the literature, including CO oxidation,[49] water gas shift (WGS),[50,51] and transesterification for biodiesel production.[52] It is worthy of note that a recent report by Schmack and coworkers described a meta-analysis of experimental results of the OCM reaction.[53] The proposed method incorporates general textbook knowledge about fundamental material properties and the experienced intuition of a chemist or material scientist about possible property–performance correlations in addition to the experimental data reported in literature.

Catalyst-performance data are a rich resource that can provide insightful information for catalyst discovery and design. The data usually involve a wide variety of multicomponent catalysts that have different compositions of diverse elements, which make meaningful statistical comparisons difficult and can prevent full use of these catalyst data with less compositional overlaps. For example, the original OCM dataset used for a previous study contained 1868 catalyst-performance data composed of 68 different

elements: 61 cations and 7 anions (Cl, F, Br, B, S, C, and P) excluding oxygen,[47] but each catalyst contained only 2.30 elements on average. In addition, catalyst research has relied heavily on prior published data, and the compositional variations of high-performance catalysts are limited. Out of 1868 catalysts for the OCM reaction, 317 catalysts performed well with $C_2$ yields ≥ 15% and $C_2$ selectivity ≥ 50%, but the frequency of the occurrence of a few specific elements, such as La, Ba, Sr, Cl, Mn, and F, was very high. This implies that, even when the data collections are large, their information coverage could be limited and biased. Elements such as Li, Mg, Na, Ca, and La also frequently appear in the dataset, which suggests that a large dataset prepared from already examined catalyst data from the literature is likely to contain a few widely used elements and many other elements with low sample counts; this situation makes statistically meaningful use of the data challenging. To efficiently utilize the experimental data for development of novel catalysts, establishing a new ML protocol that includes a variety of elements is necessary to find effective elements that have not been thoroughly explored experimentally.

The present study describes the development of an effective ML predictive protocol for experimental catalytic results including OCM, WGS, and CO oxidation reactions. The proposed ML approach considers elemental features as representative input instead of the catalyst compositions themselves, as shown in Figure. 1. Therefore, this method could allow catalyst discovery, even for catalytic reactions with little information and limited catalyst composition overlaps in the literature. In addition, the reactions were analyzed quantitatively to determine correlations between the input variables, such as catalyst compositions and experimental conditions, and reactivity properties. The resulting quantitative "feature importance score" is particularly important for the design of efficient catalytic systems for which various parameters can affect results. Promising candidate catalysts for the OCM reaction also are proposed for future study. This study presents not only the advantages of ML but also the limitations and difficulties of ML for heterogeneous catalysis. The schemes proposed and results obtained provide a valuable contribution to establishing "catalysis Informatics."[54–57]

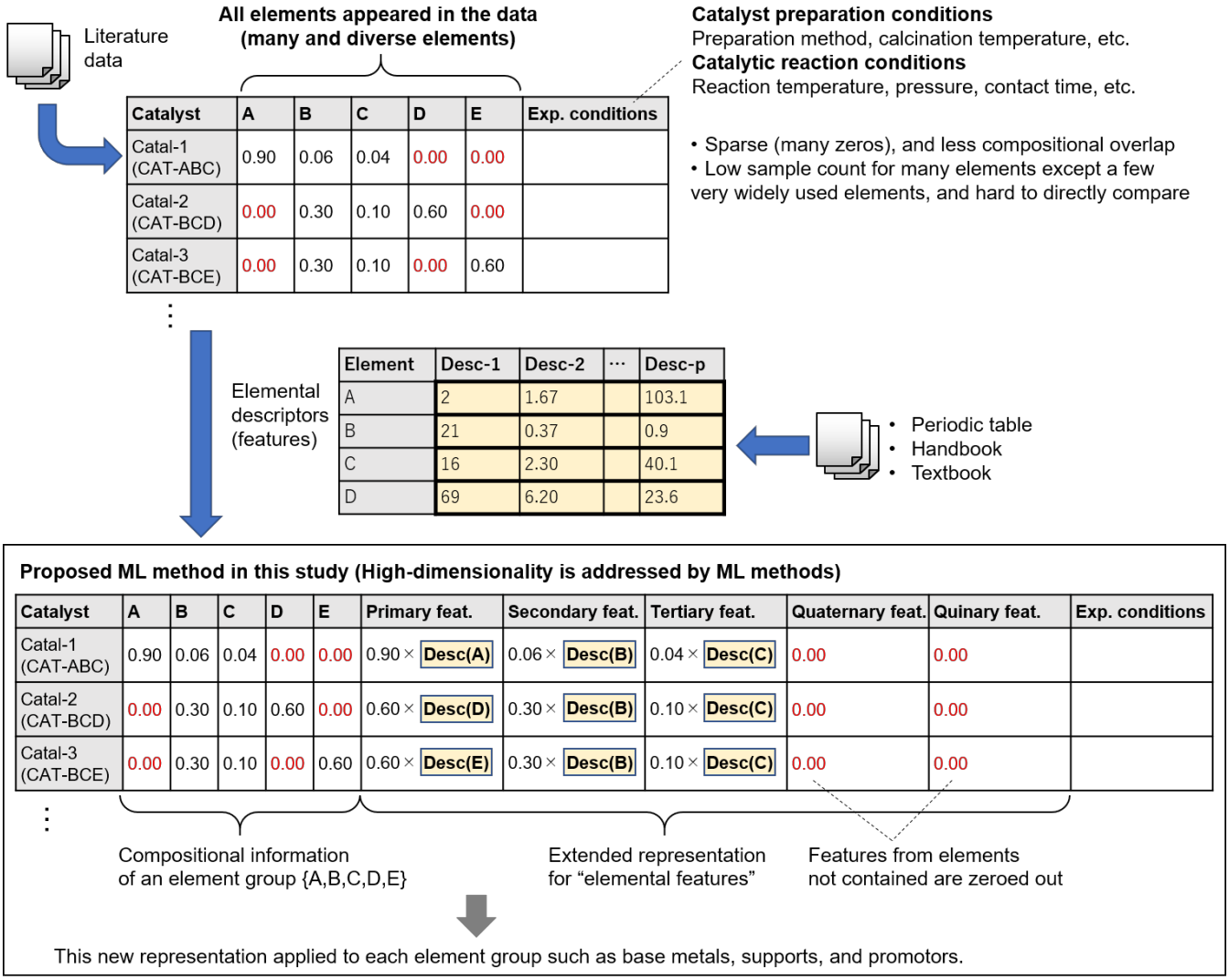



**Figure 1**. The proposed machine learning (ML) approach.

## Data and methods

### Catalyst datasets and elemental descriptors

For quantitative evaluation, three published datasets of heterogeneous catalysts and their performances were analyzed (Table 1) for OCM,[47] WGS,[51] and CO oxidation reactions.[58] The original OCM dataset consisted of 1866 catalysts with information on their compositions, support types, promoter types, experimental conditions (preparation methods, operating temperature, total pressure, and contact time), and reported performance (yield and selectivity of $C_2$ hydrocarbons). To ensure the quality of the statistical data, the following preprocessing was applied. First, catalysts with more than five elements and catalysts including Th (due to its scarcity) were removed. The WGS dataset consisted of 4360 catalysts with their compositions (wt%), support types, promoter types, experimental conditions [*e.g.*, temperature, vol% of $H_2$, $O_2$, and CO, time on stream (TOS), flow per unit weight (F/W)], and performance (% CO conversion). For preprocessing, catalysts containing “YSZ” (yttria-stabilized zirconia) were removed. The CO oxidation dataset consisted of 5610 catalysts with their compositions (wt%), support types, promoter types, experimental conditions (*e.g.*, temperature, vol% of $H_2$, $O_2$, and CO, TOS, and F/W), and performance (% CO conversion). For the three datasets, the indistinguishable catalyst records with the same catalyst compositions and reaction conditions were aggregated into a single record with the best reported performance. After preprocessing, the final numbers of catalyst records for the three datasets resulted in 1833 for OCM, 4185 for WGS, and 5567 for CO oxidation (Table 1).

**Table 1.** The three catalyst datasets investigated in this study

| Dataset | # Catalysts (Original) | # Features | Target |
|---|---|---|---|
| OCM[47] | 1833 (1866) | 105 | $C_2$ yield (%) |
| WGS[51] | 4185 (4360) | 61 | CO conversion (%) |
| CO oxidation[58] | 5567 (5610) | 80 | CO conversion (%) |

OCM: http://www.fhi-berlin.mpg.de/acnew/department/pages/ocmdata.html
WGS: https://www.sciencedirect.com/science/article/pii/S0360319914002407?via%3Dihub
CO oxidation: https://onlinelibrary.wiley.com/doi/abs/10.1002/cctc.201200665

### Machine learning (ML) methods

For performance prediction, seven well-established ML models (Table 2) were used: least absolute shrinkage and selection operator regression (LASSO),[59] ridge regression (Ridge),[60] kernel ridge regression (KRR),[61] support vector regression (SVR),[62] random forest regression (RFR),[63] gradient boosting regression with XGBoost (XGB),[64] and extra trees regression (ETR).[65] Two of these methods involve linear regression methods (Lasso, Ridge), and the other five methods involve nonlinear regression methods [more specifically, two kernel methods (KRR, SVR) and three tree ensemble methods (RFR, GBR, ETR)]. This set of ML models covers a wide spectrum of model types, which can reveal the relevant aspect of diverse data as illustrated in previous research on data-driven predictions for DFT-calculated values such as d-band centers[33,34] and adsorption energies.[40] Widely used implementations of scikit-learn

(version 0.19.1)[66] were employed for all ML models except XGB, and for XGB the original implementation (version 0.81) of XGBoost[64] was used. The key hyperparameters of each model tuned in an exhaustive way (*i.e.*, grid search) within the specified ranges shown in Table 2 (the hyperparameters not explicitly indicated in the table were set to the scikit-learn or XGBoost defaults).

The quantitative evaluations of prediction accuracy were based on the root mean squared errors (RMSEs) calculated by 10-fold cross-validation, the most widely used method for estimating prediction error. Cross-validation uses a portion of the given data to fit the model, and a different portion to test it. In 10-fold cross-validation, the given data are first partitioned into 10 equal-sized parts. For each part, an ML model was fit to the other 9 parts (a training set), and the RMSE of the fitted model calculated when predicting that part (a test set). After repeating this process for each of 10 parts, 10 RMSE values were obtained for the out-of-sample predictions, which were averaged into a single number as the final estimate for the prediction error (test error).

To evaluate the input feature variables that contributed most to the prediction of the target of interest, the "feature importance score," provided by tree ensemble methods (RFR, XGB, ETR), was used. The importance score can be of many types, typically calculated as the weighted mean of improvements in squared errors attributed to the individual feature variables, and represents the relative importance of each feature variable with respect to the predictability of the target variable. The input feature variables were seldom equally relevant, and usually only a few had significant influence on predicting the target variable. The feature importance scores do not require strong assumptions, such as linearity and independence of input variables assumed in conventional linear regression analysis.

**Table 2.** List of ML methods with search ranges for hyperparameter settings (if not specified, default values of scikit-learn were used). 3-fold cross validation with n_estimators = 100 was used for inner evaluations in the grid-search due to time constraints.

| Type | | Method | Hyperparameters [Tested range] |
|---|---|---|---|
| Linear | | Lasso | alpha $\in$ [$10^{-2}$,$10^{-1}$,1.0,$10^{1}$,$10^{2}$] |
| | | Ridge | alpha $\in$ [$10^{-2}$,$10^{-1}$,1.0,$10^{1}$,$10^{2}$] |
| Nonlinear | Kernel methods | KRR | kernel='rbf, alpha$\in$ [1.0,$10^{-1}$,$10^{-2}$,…,$10^{-5}$], gamma $\in$ [1.0,$10^{-1}$,$10^{-2}$,…,$10^{-5}$] |
| | | SVR | kernel='rbf', C $\in$ [1.0,10,$10^{2}$,…, $10^{5}$], gamma $\in$ [1.0,$10^{-1}$,$10^{-2}$,…,$10^{-9}$,$10^{-10}$], epsilon $\in$ [$10^{-2}$,$10^{-1}$,1.0,$10^{1}$,$10^{2}$] |
| | Tree ensemble methods | RFR | n_estimators = 500(OCM) or 1500(WGS, CO oxidation) |
| | | XGB | n_estimators = 500(OCM) or 1500(WGS, CO oxidation), max_depth $\in$ [6, 7, 8], learning_rate $\in$ [0.1, 0.05], subsample$\in$ [0.8, 0.9, 1], colsample_bytree$\in$ [0.8, 0.9, 1], |
| | | ETR | n_estimators = 500(OCM) or 1500(WGS, CO oxidation) |

**Prediction of multicomponent catalyst performance by incorporating descriptors for individual component elements of ML methods**

Trends in literature data, such as diversity in elements, bias from prior published data, few compositional overlaps, and low sample counts for many elements, were observed in the three datasets (Table 1). Figure 2 shows the number of common component elements for all pairwise comparisons of catalyst compositions in OCM, WGS, and CO oxidation. Even though some elements were used frequently as catalyst components, the reported catalysts have a small number of common elements, and the majority have no common elements in their compositions. In addition, the number of catalyst records having each of elements Pt, Tm, N, Ge, Cr, Cd, Re, Be, Ru, Sc, Lu, Pd, Rh, Te, In, Ga, Au, Ar, and I in OCM was less than 5 out of 1868, which was insufficient to affect the statistical analysis. To overcome this difficulty, the descriptors for each element of multicomponent catalyst representations as input variables to ML were used as shown in Figure 1. The elemental compositions were conventionally represented by a vector of compositional ratios for all elements under consideration. For example, for the dataset covering five elements A, B, C, D, and E, a catalyst with 90% A, 6% B, and 4% C would be expressed as Cat-ABC = (0.90, 0.06, 0.04, 0.00, 0.00), a catalyst with 60% D, 30% B, and 10% C was represented as Cat-BCD = (0.00, 0.30, 0.10, 0.60, 0.00), and a catalyst with 60% E, 30% B, and 10% C was represented as Cat-BCE = (0.00, 0.30, 0.10, 0.00, 0.60). However, the Euclidean distance between Cat-ABC and Cat-BCD was equal to that between Cat-ABC and Cat-BCE because only the numbers of elemental composition ratios are visible, not the individual elements. If element A shares some common properties with D but not with E, Cat-BCD and Cat-BCE can be characterized differently. However, meaningful comparisons cannot be made between two catalysts with no common elements (*i.e.*, with no compositional overlap), whereas the majority of arbitrary pairs of catalysts had no common elements (Figure 2).

Figure 1 illustrates the scheme of the proposed approach. As shown at the bottom of Figure 1, the proposed input representation contains additional variables for considering the “elemental features” of component elements. Due to this, Cat-BCD can be distinguished from Cat-BCE. For Cat-BCD that contains 60% D, 30% B, and 10% C, the primary, secondary, and tertiary components are respectively D, B, and C. Thus, the elemental descriptors for D, B, and C were set in this order, each of which was multiplied by ratio 0.60, 0.30, and 0.10, respectively. In this case, the resultant vector is represented by Cat-BCD=(0.00, 0.30, 0.10, 0.60, 0.00, $0.60\times\text{desc-}D_1$, $0.60\times\text{desc-}D_2$,…, $0.60\times\text{desc-}D_p$, $0.30\times\text{desc-}B_1$, $0.30\times\text{desc-}B_2$,…, $0.30\times\text{desc-}B_p$, $0.10\times\text{desc-}C_1$, $0.10\times\text{desc-}C_2$,…, $0.10\times\text{desc-}C_p$), where $\text{desc-}X_i$ indicates the i-th descriptor value for element X (p=11 in this paper as described later). This extended representation was applied to each group of elements, such as base metals, supports, and promoters. The multiplication and reordering are needed because of the following: (1) no information should be included from elements not contained in the catalyst; and (2) the elemental descriptor values are “constant” for each element, and, hence, just replacing elemental ratios by elemental descriptors could not add any more information without catalyst-specific manipulations. Note that soft weighting using numbers between 0 and 1 with a sum of 1 is a common technique to represent statistical interactions such as those between compositional ratios and descriptors of the component element in this study.

As elemental descriptors for the proposed ML model, the following 11 physical properties (readily available from the periodic table and chemical handbook) were used for each element:[67] atomic number (AN), atomic weight (AW) in g $mol^{-1}$, group, period, atomic radius in Å, electronegativity, melting point (m.p.) in K, boiling point (b.p.) in K, enthalpy of fusion in J $g^{-1}$, density ($\rho$) at 25 °C in g $cm^{-3}$, and ionization energy in eV.

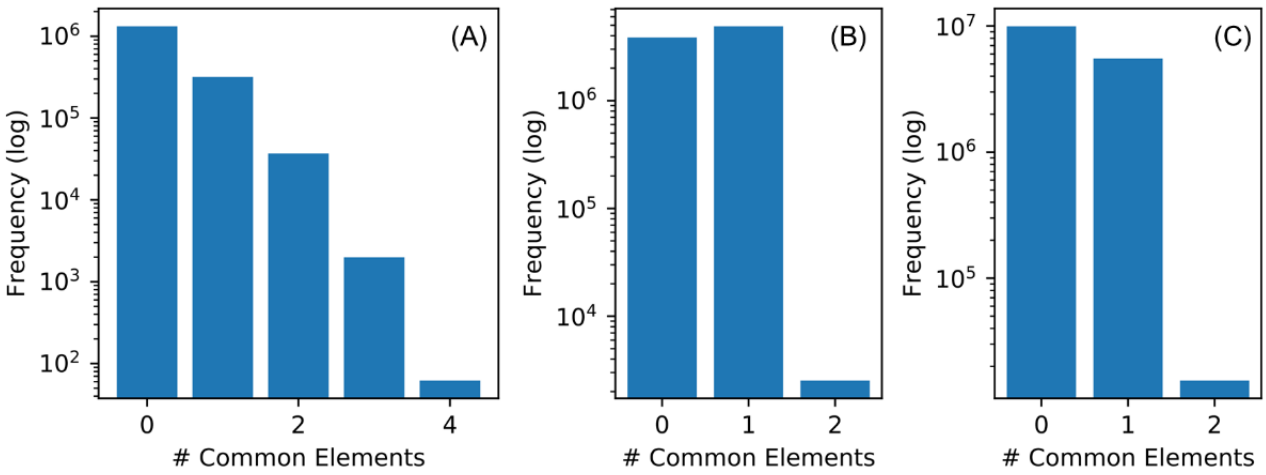



**Figure 2**. Number of common elements for all pairwise comparisons of catalyst compositions for (A) OCM, (B) WGS, and (C) CO oxidation. Frequencies (y axis) are on a log scale.

## Results and Discussion

### Quantitative evaluations for performance prediction of OCM, WGS, and CO oxidation catalysts

To evaluate performance predictions using the proposed approach, a quantitative investigation was done based on the OCM dataset,[47] which had previously published analysis results.[48,68] In a recent study by Takahashi and coworkers,[68] the catalysts were classified into four groups with $C_2$ yields of 0-10%, 10-20%, 20-30%, and greater than 30%, and the performance by classification version of RFR was evaluated. This was done because the OCM dataset from the literature is noisy and inconsistent due to the variety of data sources from different instruments, procedures, platforms, and researchers. In contrast, prediction of $C_2$ yield performance directly as a regression problem rather than a classification problem was pursued in the present study, allowing for quantitative analysis of the noise variance and predictability across multiple state-of-the-art ML methods.

First, the prediction performance of the proposed approach was evaluated using seven different ML methods (Table 2). All evaluations were conducted using 10-fold cross-validation and the root mean square error (RMSE) of the difference between the predicted value and ground truth; standard deviations also were calculated. Figure 3 shows the prediction errors (RMSEs) for the predictions of $C_2$ yield from OCM catalysts. Non-linear methods performed better than linear methods, and, in particular, tree ensemble methods (RFR, XGB, ETR) resulted in small training and test errors. The best prediction performance by XGB was RMSE of 0.52 (training error) and 4.15 (test error) for predicting $C_2$ yield (%). A previous study by Takahashi *et al*. regarded catalysts as being at roughly the same level if the performance was within the same 10% intervals;[68] therefore, this 4.15% test error in RMSE was informative. In addition, Figure 4 provides a visual representation of the details of the ML predictions by plotting actual $C_2$ yield (x axis) against predicted yield by ML (y axis). Linear methods revealed underfitting, and continuous nonlinear methods, such as KRR and SVR, were difficult to fit to the proposed representations that were discontinuous. In contrast, tree ensemble methods, such as RFR, XGB, and ETR, worked fairly well, and therefore a combination of these methods was used with the proposed representations. The data were distributed mainly in low-performance intervals, which cannot simply be identified from quantitative evaluations such as classification performances and RMSEs.

Second, the three representation patterns for input to ML were evaluated using the XGB model, which showed the best performance, combined with different input representations: (1) conventional ML method using only catalyst compositions, (2) conventional ML method using both catalyst compositions + experimental conditions, and (3) proposed ML method using both catalyst compositions + experimental conditions. The proposed method used the elemental features shown in Figure 1. Figure 5 shows the results for predictions of 10% catalysts (in red) from the remaining 90% catalysts (in blue), which strongly suggests that the performance prediction for OCM relied heavily on experimental conditions; using only catalyst compositional information resulted in poor prediction results. The visual assessment also indicated that the improvement from (2) to (3) was marginal compared to that from (1) to (2), but the data shown in Tables 3 and 4 confirm quantitatively that the proposed representation improves the prediction

performance for all 3 cases of OCM, WGS, and CO oxidation, even though the datasets contained unavoidable intrinsic noise originating from the variety and uncertainty of multiple data sources, which limited predictability.

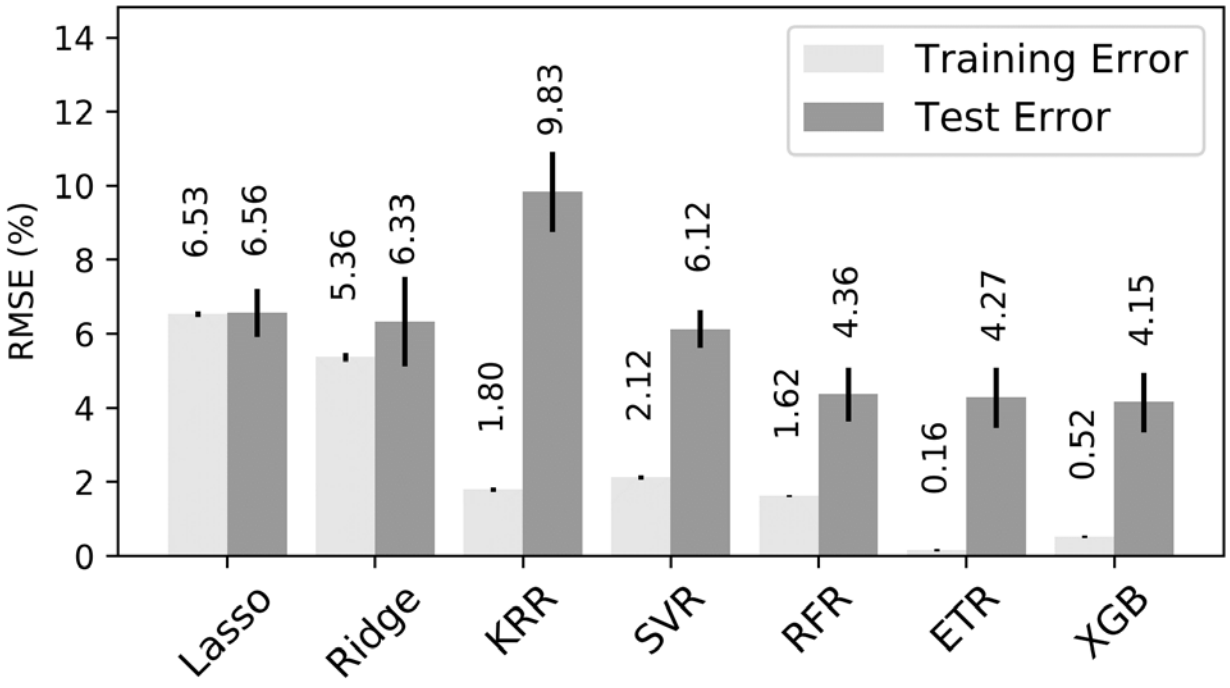



**Figure 3**. Prediction-error comparisons for catalyst performance ($C_2$ yield) for OCM. The proposed approach was tested with seven state-of-the-art machine learning methods. Evaluations were done with respect to RMSEs estimated by 10-fold cross-validation. The error bar indicates the 95% CI (±2σ).

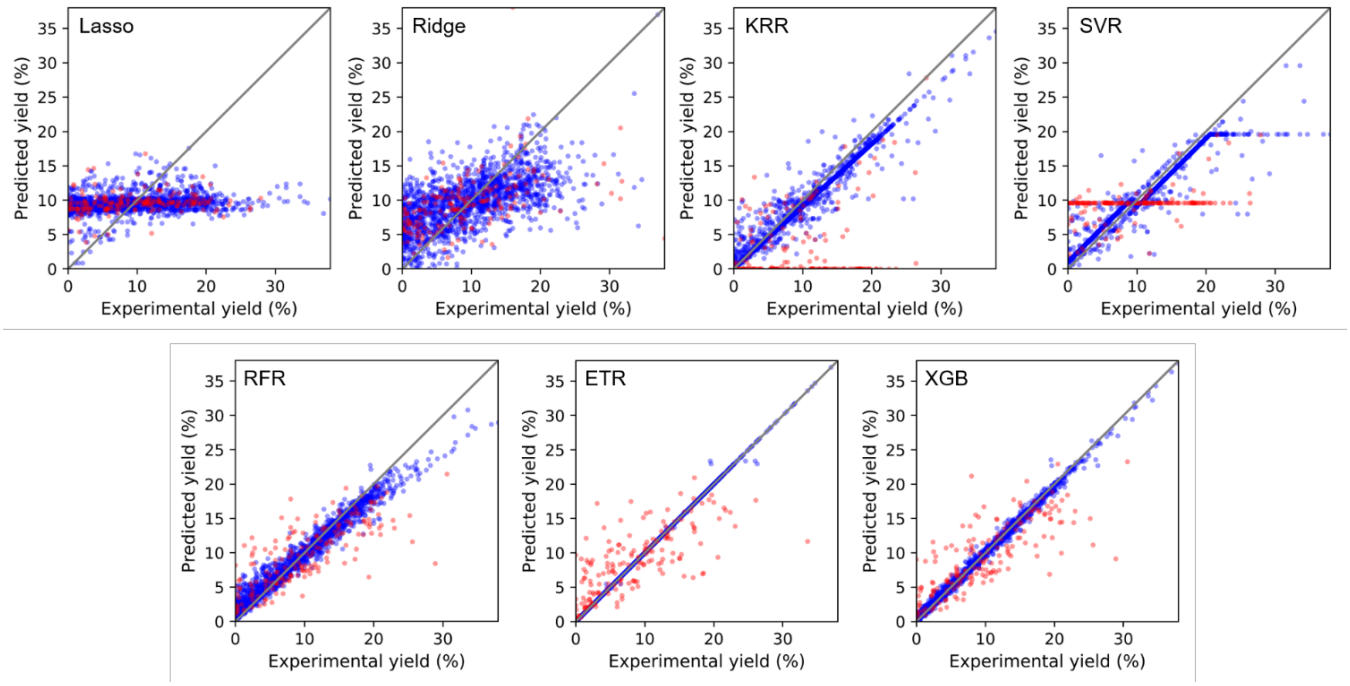



**Figure 4**. The 90%/10% training-test error plots for catalyst performance ($C_2$ yield) for OCM. The proposed approach was tested using seven ML methods. Training data (blue), test data (red).

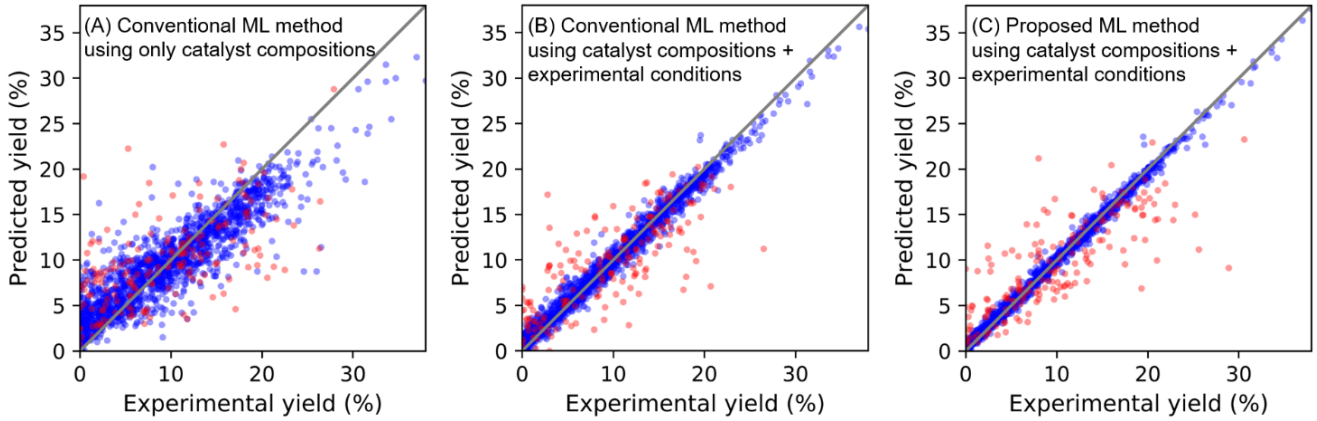



**Figure 5**. Comparison of 90%/10% training-test error plots for three representation patterns for catalyst performance ($C_2$ yield) for OCM. (A) Conventional ML method using (A) only catalyst compositions, (B) catalyst compositions + experimental conditions, and (C) proposed ML method using catalyst compositions + experimental conditions. XGB was used for the ML method. Training data (blue), test data (red).

**Table 3.** Comparison of prediction accuracy (RMSE, 10-fold cross-validation) for $C_2$ yield (%) of OCM. Three representation patterns (composition only, composition + condition, and the proposed) were tested with seven ML methods. The numbers shown in parentheses are the corresponding σs.

| Method | Lasso | Ridge | KRR | SVR | RFR | ETR | XGB |
|---|---|---|---|---|---|---|---|
| Conventional ML method | | | | | | | |
| Only catalyst composition | | | | | | | |
| Training Error | 6.23 (0.05) | 6.09 (0.05) | 6.21 (0.05) | 2.65 (0.04) | 2.54 (0.02) | 1.85 (0.03) | 3.01 (0.03) |
| Test Error | 6.37 (0.41) | 6.36 (0.43) | 6.33 (0.41) | 5.98 (0.30) | 5.33 (0.37) | 5.48 (0.38) | **5.23** (0.35) |
| Both catalyst composition & experimental conditions | | | | | | | |
| Training Error | 6.55 (0.03) | 5.56 (0.05) | 5.95 (0.04) | 5.55 (0.04) | 1.65 (0.02) | 0.16 (0.02) | 0.97 (0.02) |
| Test Error | 6.57 (0.33) | 6.64 (0.82) | 6.11 (0.38) | 6.06 (0.41) | 4.41 (0.34) | 4.32 (0.27) | **4.23** (0.24) |
| Proposed ML method | | | | | | | |
| Both catalyst composition & experimental conditions | | | | | | | |
| Training Error | 6.53 (0.04) | 5.36 (0.06) | 1.80 (0.03) | 2.21 (0.03) | 1.62 (0.02) | 0.16 (0.02) | 0.52 (0.02) |
| Test Error | 6.56 (0.33) | 6.33 (0.60) | 9.83 (0.54) | 6.12 (0.26) | 4.36 (0.37) | 4.27 (0.41) | **4.15** (0.40) |

**Table 4.** Comparison of prediction accuracy (RMSE, 10-fold cross-validation) for CO conversion (%) of WGS and CO oxidation. Three representation patterns (composition only, composition + condition, and the proposed) were tested with three tree ensemble methods.

| Method | WGS | | | CO oxidation | | |
|---|---|---|---|---|---|---|
| | RFR | ETR | XGB | RFR | ETR | XGB |
| Conventional ML method | | | | | | |
| Only catalyst composition | | | | | | |
| Training Error | 29.62 (0.07) | 29.61 (0.07) | 29.61 (0.07) | 29.33 (0.09) | 29.32 (0.09) | 29.36 (0.09) |
| Test Error | 31.15 (0.65) | 31.17 (0.64) | 31.17 (0.65) | 30.33 (0.86) | 30.36 (0.87) | 30.37 (0.85) |
| Both catalyst composition & experimental conditions | | | | | | |
| Training Error | 4.24 (0.04) | 0.01 (0.00) | 2.95 (0.05) | 5.23 (0.04) | 0.02 (0.01) | 2.65 (0.09) |
| Test Error | 11.48 (0.91) | 10.45 (0.79) | 8.33 (0.60) | 14.30 (0.84) | 12.58 (0.69) | 12.12 (0.67) |
| Proposed ML method | | | | | | |
| Both catalyst composition & experimental conditions | | | | | | |
| Training Error | 4.11 (0.06) | 0.01 (0.00) | 2.41 (0.03) | 5.02 (0.04) | 0.02 (0.01) | 2.66 (0.06) |
| Test Error | 11.39 (0.77) | 10.12 (0.98) | **8.28** (0.81) | 13.71 (0.59) | 12.13 (0.69) | **12.05** (0.55) |

Finally, the input variables that provided the most contribution to prediction of the catalyst performance were evaluated. For this purpose, feature importance scores were calculated from the optimized XGB models from the proposed ML method. This score was used to assess the relative importance of a descriptor with respect to predictability of the catalytic performances, and therefore, it is important for the systematic design of efficient catalytic processes. The default type of XGBoost ('weight') was used among options in this analysis. Figure 6 shows the top 20 contributing variables for predicting the performance for OCM, WGS, and CO oxidation reactions. The variables for elemental features were Metal_i or Promotor_i, which represent the type of active metal or promoter, respectively, in the catalyst. The i indicated the order of the abundance of the element in the catalyst. The top descriptor was reaction temperature for all three reactions. Reaction temperature for WGS and CO oxidation, especially, had relatively strong contributions compared to reaction temperature of the OCM reaction. The OCM reaction can be classified as a "selective" catalytic reaction, whereas the other two reactions could not. The $C_2$ yield for the OCM reaction decreased and undesirable coke or $CO_x$ formation occurred when reaction temperature was too high. Therefore, the feature importance score suggests that the catalytic performance is not dependent only on reaction temperature, especially for selective catalytic reactions such as the OCM. Experimental conditions other than catalyst composition also can contribute to reaction outcomes. Contact time, $P_{CH4}$, and $P_{CH4}/P_{O2}$ were the next top three contributors to OCM reaction results. This is probably because that the OCM reaction mechanism involves a reactive radical species.[47,69] The initial step of the OCM reaction is formation of $CH_3$ radical species. After formation of the $CH_3$ species on a catalytic surface, gas-phase reactions proceed, and $CH_3$ radical species are expected to combine to form the $C_2$ products. For catalyst compositions, alkali metals and alkaline earth metals, such as Li, Ca, Na, Ba, and Mg, influenced $C_2$ yield. In addition, La was a top-12 contributing descriptor. These observations agree with past reports on multicomponent catalysts containing a host basic metal oxide (MgO or $La_2O_3$) promoted with metal oxide dopants that positively influenced $C_2$ selectivity (Li, Na, Cs, Sr, Ba) .[47,48] The present approach permits further quantitative analysis that includes catalyst compositions and experimental conditions and element features, and thus, provides greater insights into catalytic reactions.

For the drawbacks and limitations of feature importance analysis, it should be noted that the results could depend on the choice of importance criteria, the choice of tree learning algorithms, as well as the sample variability, and thus careful evaluations would be required for further confident interpretation. For example, permutation test[63,70] as well as Y-randomization and pseudo-descriptor analysis[71] would be possible options. However, it should also be kept in mind that forcibly projecting *multivariate* trends into independent *univariate* contributions of individual features always entails information loss due to the cross correlation of features. These 'interpretability or explainability' aspects for blackbox ML algorithms have also been a recent hot topic in the ML community.

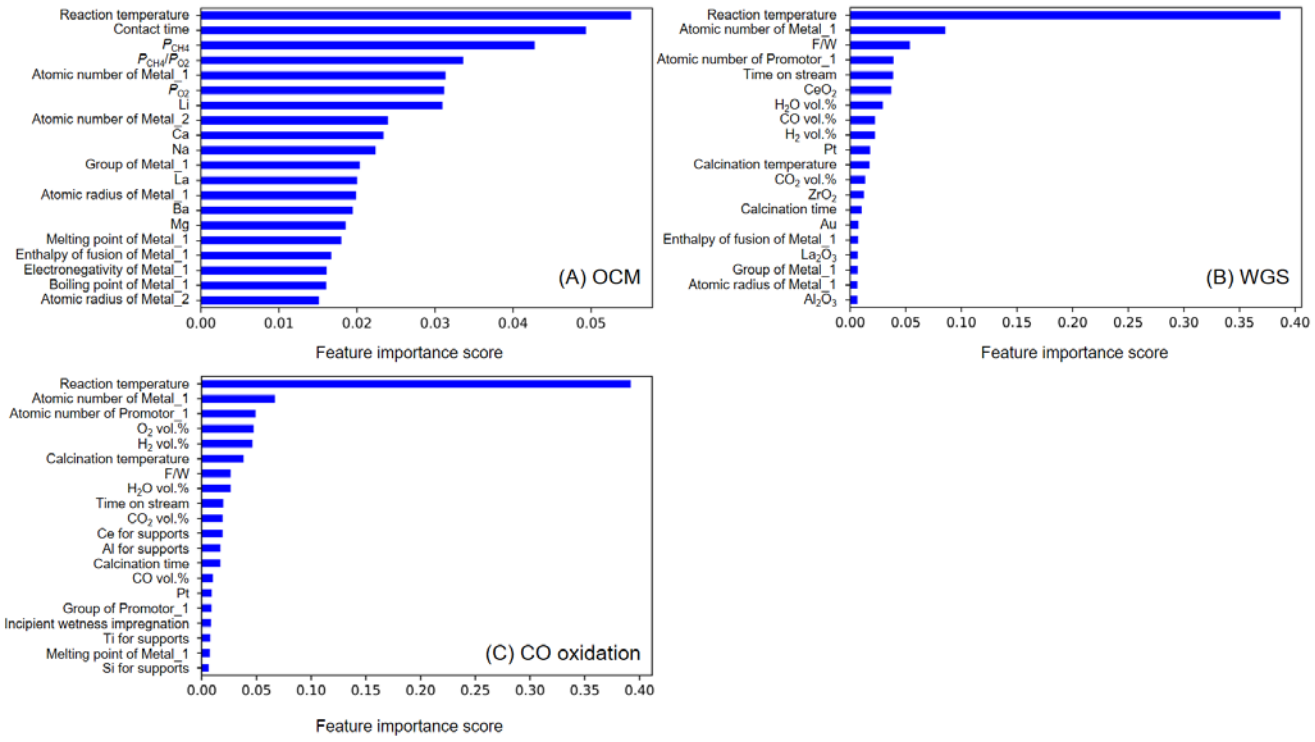



**Figure 6**. Top 20 contributing descriptors for predicting (A) $C_2$ yield from OCM, (B) CO conversion from WGS, and (C) CO conversion from CO oxidation, based on the average feature importance from the best XGB models in 10-fold cross-validation.

**Simulated evaluation for catalyst optimization based on ML surrogates**

Optimization of catalysts using ML predictions also was investigated. Since catalyst research has relied heavily on prior published data, the research tends to be biased toward variations in catalyst composition that were successful. Since ML builds predictive models that are representative of the training data, using ML predictions directly would result only in finely tuned variations of previously investigated catalysts and would narrow the scope of catalysts.

To take into account this data bias problem, a catalysis optimization procedure using ML as "surrogate" models was developed and evaluated. In many practical optimization cases in engineering and scientific applications, obtaining accurate target values through laborious experiments or high-fidelity simulations is extremely time-consuming and costly. Therefore, using them directly for comprehensive exploration of high-performance catalysts from an infinite number of theoretically possible catalysts is difficult. Accurate and rapid ML predictions can be used as surrogate models for these costly experiments and time-consuming simulations. In the simplest scenario, called the "exploitation only" strategy, ML models were fit to the training data (literature data) and fed many possible inputs to search for new catalyst candidates. The input with the best predicted catalyst performance can then be tested. But this strategy is often unsuccessful because training data usually are too limited and potentially biased and thus are insufficient to cover all potential catalyst variations. In many practical cases, "exploration" steps also are needed to gain new information by testing new catalysts not close to any of previously tested catalysts. The 'exploitation' makes the best prediction given current information (data), and the 'exploration' gather more information (data). This fundamental tradeoff between exploitation and exploration is the key to identifying promising catalyst candidates to be tested next. To balance between exploiting known data and exploring for new data, global optimization procedures have been investigated in many forms such as surrogate-based optimization (SBO),[72,73] Bayesian optimization (BO),[74–76] response surface methods,[77,78] sequential design of experiments and multi-armed bandits,[79–81] kriging,[82] and derivative-free and blackbox optimization.[83]

In the present study, an SBO strategy based on the *sequential model-based algorithm configuration* (SMAC) procedure[84] was developed and evaluated with *expected improvement* (EI)[85] as the infill criterion (acquisition function). It iterated on the following general steps. First, ML models equipped with predictive variances were fit to the given samples, EIs of candidate samples were calculated based on the fitted ML model, and samples with the greatest EI value as well as some random samples were selected for further testing. For quantitative comparisons, a "random" strategy was used by evaluating additional samples randomly, and we evaluated "exploitation only" strategies by fitting ML models and selecting samples with the best predicted performance (without the EI criterion and random sample inclusion), as well as "exploitation + exploration" strategies that maximize EI by SMAC with RFR or ETR and by BO with Gaussian process regression (GPR).[86] For RFR, ETR, and XGB, RFR and ETR were used with n_estimators=300 with other hyperparameters defaults, and XGB was used with n_estimators=300, max_depth=8, subsample=0.8, colsample_bytree=0.9, and learning_rate=0.05. For

GPR, the scikit-learn implementation was used with the Matern 5/2 kernel with alpha=0.01, n_restarts_optimizer=10, and normalize_y=True. Note that the original SMAC algorithm is based on RFR, but the present study also used ETR for estimating the predictive variance using the formula in 4.3.2 from a report by Hutter and coworkers.[87]

To evaluate the developed catalyst optimization strategies, a simulated situation was set up based on the OCM dataset. In this random simulation, 10 catalysts are provided initially; then each optimization strategy sequentially selects catalyst candidates to test (other than the initial 10 catalysts). More specifically, the following random simulation is repeated 10 times for individual cases to determine how early each strategy can find high-performance catalysts: 1) 10 catalysts are selected randomly from the dataset as initial samples; 2) an ML model is fit to the samples and the criterion value (*i.e.*, EI or predicted value) calculated for catalysts in the dataset that have not yet been selected; 3) the catalyst with the best criterion values is added to the initial samples; and then step 2 is repeated. Note that SMAC also requires addition of newly generated random samples at step 3 for exploration purposes. Figure 7 shows the averaged curve plots of the highest values found for the first 400 selections from the “exploitation only” and “exploitation + exploration” strategies. Each simulation was run 10 times and the average values were used. Both strategies could identify high-performance catalysts much earlier than the random strategy, which demonstrates that ML surrogate-based optimization is a promising approach for a sequential design of experiments. The “exploitation only” strategy worked well in this simulation, although no ML models can perform well in principle when extrapolating catalysts dissimilar to any set of given samples, which implies that the high-performance catalysts in the current OCM dataset have limited variations and a strong bias. Since this simulation is based on the currently available dataset for OCM, the training and test datasets have some similarity. In this situation, wider “exploration” is not needed, and efforts can focus on fine-tuning the currently obtained catalysts. However, the biased datasets for simulating catalysts optimization is a limitation, and potential high-performance catalysts are not necessarily similar to our limited training data. Ideally, for real-world situations, the datasets should be unrestricted, and the exploitation-exploration tradeoff [Figure 7(B)] would play an important role in practical situations.

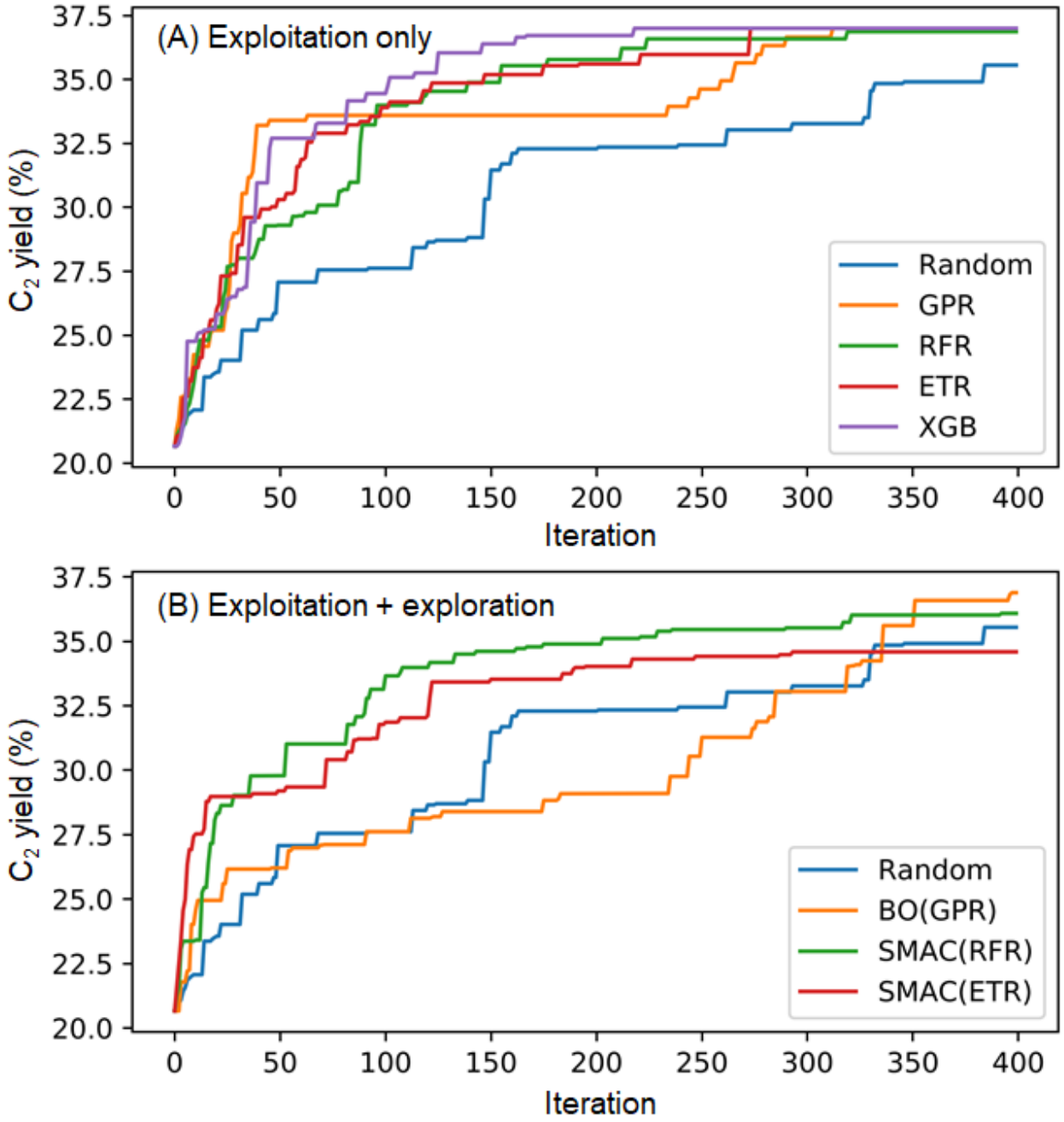



**Figure 7.** Performance comparisons of the first 400 catalysts identified using the simulated catalyst optimization process by an (A) “exploitation only” strategy with 4 ML methods and a random strategy as a baseline and an (B) “exploitation + exploration” strategy with the proposed methods based on SMAC with RFR or ETR, a BO strategy with GPR, or a random strategy as a baseline. Each curve is the average of curves over 10 simulations.

Finally, applying the “exploitation + exploration” strategy using a modified SMAC procedure with ETR to the entire OCM dataset suggested several potential catalyst candidates with high EI values to be tested (Figure 8). In previous situations (such as that shown in Figure 7), only the EIs for those finite candidates needed to be calculated to select the next samples to evaluate. However, in general situations, an infinite number of candidates are available to test next. Therefore, points with the greatest EIs need to

be identified by maximizing the EI function. However, due to the discontinuity and multimodality of the EI function, obtaining the local maximizers is very difficult. The SMAC procedure performs both local searches started from current samples with high EI values and random searches from the entire feature space. To search for OCM catalysts, input variables were grouped into “catalyst composition,” “support,” “promotor,” “preparation,” and “experimental condition.” For each group, a local search by slightly changing the variables in the group was performed. For “catalyst composition,” after slightly changing the variables, the four largest values were selected (search was done for up to four components as base metals), all other variables were set to zero, and the variable set was normalized so that the total sum of elements equaled 1. For “support,” a search was done for up to two components as support metals because no data having more than 3 supports were available in the OCM datasets. For “promoter,” the variables were set randomly to have a single 1 and all others 0s, or all 0s. For “preparation,” one of six options was selected randomly by setting the variables to have a single 1 and all others 0s. For “experimental condition,” a random sampling was used as the original SMAC procedure but with renormalization. Note that the standard “one-hot encoding” was used to represent non-numeric variables such as “promoter” and “preparation.” For example, a variable taking one of ‘a’,’b’, and ‘c’ is represented as a 3-dimensional binary vector taking (1, 0, 0) for ‘a’, (0, 1, 0) for ‘b’, and (0, 0, 1) for ‘c’.

Figure 8 shows a list of the top 20 promising candidate catalysts worth testing, suggested from the entire OCM dataset using the present method. In standard SBO situations, only one point with the highest EI value is needed, but for now, multiple candidates with high EI values were obtained by k-means clustering (k = 421, # of the OCM literature) to aggregate the searched results from multiple starts, and the highest points in each cluster were selected as representatives. A catalyst composed of 68.2 mol% Al, 14.6 mol% Mg, 9.7 mol% Li, 4.9 mol% Mo, and 2.6 mol% Na had the greatest EI values. The second best performing catalyst consisted of 60.7 mol% Ti, 30.3 mol% Ba, 6.0 mol% Cl, and 3.0 mol% Sn. Note that the labels of each catalyst contain only the compositions of active components, and O was not included in this expression. In addition to the group 1 and 2 elements, which are expected to appear for high-performing catalysts, these investigations imply that other elements, which do not appear frequently in the literature, also can be effective in the OCM process. The ultimate goal of this effort was to utilize the present approach to obtain fundamental knowledge about the factors determining catalytic performance to design ideal catalysts on the atomic level, even for catalysts with less explored or unexplored compositions. Since the catalytic properties of materials should be determined by their elemental features, such as physical properties and electronic structures,[88–90] the strategy used here was to find novel catalysts having ideal elemental features by changing the composition of the materials.

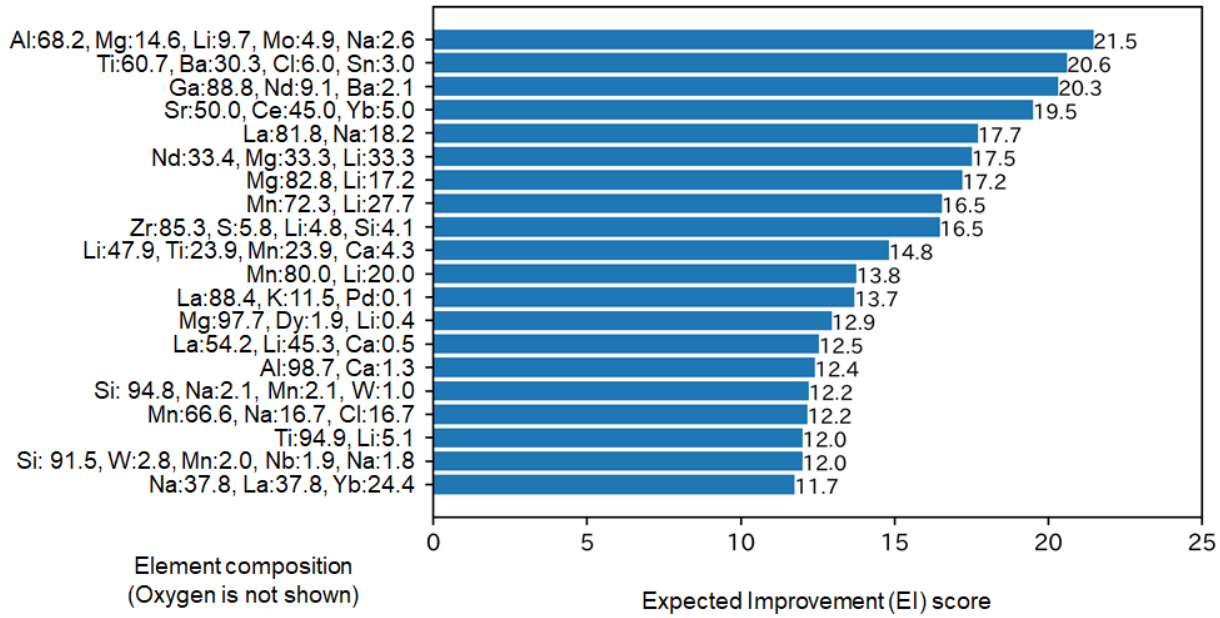



**Figure 8.** Top 20 promising candidate catalysts for OCM, worth testing next, as suggested by the entire OCM dataset using the proposed method. Similar catalysts are clustered, and representative catalysts with the greatest EI in each group are shown.

## Conclusions

The value of a machine learning (ML) approach that involves elemental features as representational input instead of catalyst compositions was demonstrated using previous experimental catalytic data on oxidative coupling of methane (OCM), water gas shift (WGS), and CO oxidation reactions. Prediction accuracy was improved using this approach when compared to conventional methods utilizing catalyst compositions as input. Among several the-state-of-the-art ML methods tested in this study, gradient boosting regression with XGBoost (XGB) performed the best for predicting catalytic performance. A feature importance score also was calculated to determine quantitatively the most influential input feature variables and predict catalyst performance. In addition, a catalyst optimization procedure was explored by using ML as “surrogate” models. Based on this optimization, the top 20 promising catalyst candidates for the OCM reaction were identified for future study. This study provides fundamental knowledge about heterogeneous catalytic processes and is expected to identify truly novel catalysts, even in the field of heterogeneous catalysis where the literature data are very noisy and inconsistent due to the variety of data arising from different instruments, procedures, platforms, and researchers, which can be limiting and biased.

## Acknowledgements

This work was supported by the KAKENHI grants JP17H01341, JP17H01783, and JP17K19953, JP18K14051, JP18K14057 and a Grant-in-Aid for Scientific Research on Innovative Areas “Nano Informatics” (No. 25106010) from the Japan Society for the Promotion of Science (JSPS), by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) within the projects "Integrated Research Consortium on Chemical Sciences (IRCCS)" and "Elements Strategy Initiative to Form Core Research Center" as well as by the JST-CREST projects JPMJCR15P4 and JPMJCR17J3, and JST-PRESTO project JPMJPR15N9. The authors thank Prof. Kondratenko for providing catalytic performance data on the OCM reaction. The authors also thank Prof. Yildirim and Prof. Günay for providing catalytic performance data on the WGS and CO oxidation reactions.

## Supporting Information

All our scripts and data are available on github (https://github.com/itakigawa/ml-catalysis).

**Keywords**: Machine learning • Catalysis informatics • Methane oxidative coupling • Water gas shift • CO oxidation

## References


[1] N. Mizuno, M. Misono, *Chem. Rev.* **1998**, *98*, 199–218.

[2] J. K. Nørskov, T. Bligaard, J. Rossmeisl, C. H. Christensen, *Nat. Chem.* **2009**, *1*, 37–46.

[3] R. Schlögl, *Angew. Chemie - Int. Ed.* **2015**, *54*, 3465–3520.

[4] E. J. Ras, G. Rothenberg, *RSC Adv.* **2014**, *4*, 5963–5974.

[5] R. A. Van Santen, I. Tranca, E. J. M. Hensen, *Catal. Today* **2015**, *244*, 63–84.

[6] L. Liu, A. Corma, *Chem. Rev.* **2018**, *118*, 4981−5079.

[7] J. M. Caruthers, J. A. Lauterbach, K. T. Thomson, V. Venkatasubramanian, C. M. Snively, A. Bhan, S. Katare, G. Oskarsdottir, *J. Catal.* **2003**, *216*, 98–109.

[8] S. Katare, J. M. Caruthers, W. N. Delgass, V. Venkatasubramanian, *Ind. Eng. Chem. Res.* **2004**, *43*, 3484–3512.

[9] G. M. Diamond, K. A. Hall, A. M. Lapointe, M. K. Leclerc, J. Longmire, J. A. W. Shoemaker, P. Sun, *ACS Catal.* **2011**, *1*, 887–900.

[10] P. Kondratyuk, G. Gumuslu, S. Shukla, J. B. Miller, B. D. Morreale, A. J. Gellman, *J. Catal.* **2013**, *300*, 55–62.

[11] J. R. Kitchin, A. J. Gellman, *AIChE J.* **2016**, *62*, 3826–3835.

[12] G. Rothenberg, *Catal. Today* **2008**, *137*, 2–10.

[13] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, *Chem. Mater.* **2017**, *29*, 9436−9444.

[14] D. Wolf, O. V. Buyevskaya, M. Baerns, *Appl. Catal. A Gen.* **2000**, *200*, 63–77.

[15] Y. Yamada, A. Ueda, K. Nakagawa, T. Kobayashi, *Res. Chem. Intermed.* **2002**, *28*, 397–407.

[16] Y. Watanabe, T. Umegaki, M. Hashimoto, K. Omata, M. Yamada, *Catal. Today* **2004**, *89*, 455–464.

[17] S. Moehmel, N. Steinfeldt, S. Engelschalt, M. Holena, S. Kolf, M. Baerns, U. Dingerdissen, D. Wolf, R. Weber, M. Bewersdorf, *Appl. Catal. A Gen.* **2008**, *334*, 73–83.

[18] J. Llamas-Galilea, O. C. Gobin, F. Schüth, *J. Comb. Chem.* **2009**, *11*, 907–913.

[19] J. E. Kreutz, A. Shukhaev, W. Du, S. Druskin, O. Daugulis, R. F. Ismagilov, *J. Am. Chem. Soc.* **2010**, *132*, 3128–3132.

[20] T. C. Le, D. A. Winkler, *Chem. Rev.* **2016**, *116*, 6107−6132.

[21] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–36.

[22] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547–555.

[23] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, et al., *Nat. Rev. Mater.* **2018**, *3*, 5–20.

[24] E. Burello, D. Farrusseng, G. Rothenberg, *Adv. Synth. Catal.* **2004**, *346*, 1844–1853.

[25] M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res.* **2016**, *49*, 1292−1301.

[26] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2016**, *2*, 725–732.

[27] Z. Zhou, X. Li, R. N. Zare, *ACS Cent. Sci.* **2017**, *3*, 1337−1344.

[28] C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

[29] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610.

[30] J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, *Phys. Rev. B* **2016**, *93*, 115104.

[31] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2016**, *6*, 19375.

[32] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, *J. Phys. Chem. Lett.* **2018**, *9*, 1668–1673.

[33] I. Takigawa, K. Shimizu, K. Tsuda, S. Takakusagi, *RSC Adv.* **2016**, *6*, 52587–52595.

[34] I. Takigawa, K. Shimizu, K. Tsuda, S. Takakusagi, in *Nanoinformatics*, **2018**, pp. 45–64.

[35] X. Ma, Z. Li, L. E. K. Achenie, H. Xin, *J. Phys. Chem. Lett.* **2015**, *6*, 3528−3533.

[36] R. Jinnouchi, R. Asahi, *J. Phys. Chem. Lett.* **2017**, *8*, 4279–4283.

[37] R. Gasper, H. Shi, A. Ramasubramaniam, *J. Phys. Chem. C* **2017**, *121*, 5612–5619.

[38] F. Göltl, P. Müller, P. Uchupalanun, P. Sautet, I. Hermans, *Chem. Mater.* **2017**, *29*, 6434–6444.

[39] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, et al., *ACS Catal.* **2017**, *7*, 6600−6608.

[40] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. Shimizu, I. Takigawa, *J. Phys. Chem. C* **2018**, *122*, 8315–8326.

[41] T. Hattori, S. Kito, *Catal. Today* **1991**, *10*, 213–222.

[42] T. Hattori, S. Kito, *Catal. Today* **1995**, *23*, 347–355.

[43] M. Holeňa, M. Baerns, *Catal. Today* **2003**, *81*, 485–494.

[44] J. M. Serra, A. Corma, A. Chica, E. Argente, V. Botti, *Catal. Today* **2003**, *81*, 393–403.

[45] A. Corma, J. M. Serra, P. Serna, S. Valero, E. Argente, V. Botti, *J. Catal.* **2005**, *229*, 513–524.

[46] K. Omata, *Ind. Eng. Chem. Res.* **2011**, *50*, 10948–10954.

[47] U. Zavyalova, M. Holena, R. Schlögl, M. Baerns, *ChemCatChem* **2011**, *3*, 1935–1947.

[48] E. V. Kondratenko, M. Schlüter, M. Baerns, D. Linke, M. Holena, *Catal. Sci. Technol.* **2015**, *5*, 1668–1677.

[49] M. E. Günay, R. Yildirim, *Appl. Catal. A Gen.* **2013**, *468*, 395–402.

[50] M. E. Günay, F. Akpinar, Z. I. Onsan, R. Yildirim, *Int. J. Hydrogen Energy* **2012**, *37*, 2094–2102.

[51] Ç. Odabaşi, M. E. Günay, R. Yildirim, *Int. J. Hydrogen Energy* **2014**, *39*, 5733–5746.

[52] N. Alper Tapan, R. Yıldırım, M. E. Günay, *Biofuels, Bioprod. Biorefining* **2016**, *10*, 422–434.

[53] R. Schmack, A. Friedrich, E. V. Kondratenko, J. Polte, A. Werwatz, R. Kraehnert, *Nat. Commun.* **2019**, *10*, 441.

[54] A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders, R. Fushimi, *ACS Catal.* **2018**, *8*, 7403–7429.

[55] B. R. Goldsmith, J. Esterhuizen, C. J. Bartel, C. Sutton, J.-X. Liu, *AIChE J.* **2018**, *64*, 2311–2323.

[56] S. L. Philomena, K. Winther, J. A. G. Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, *ChemCatChem* **2019**, DOI:10.1002/cctc.201900595.

[57] K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, et al., *ChemCatChem* **2019**, *11*, 1146–1152.

[58] M. E. Günay, R. Yildirim, *ChemCatChem* **2013**, *5*, 1395–1406.

[59] R. Tibshirani, *J. R. Stat. Soc. B* **1996**, *58*, 267–288.

[60] A. E. Hoerl, R. W. Kennard, *Technometrics* **1970**, *12*, 55–67.

[61] C. Saunders, A. Gammerman, V. Vovk, in *Proc. 15th Int. Conf. Mach. Learn.*, **1998**, pp. 515–521.

[62] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, V. Vapnik, in *Adv. Neural Inf. Process. Dystems*, **1997**, pp. 155–161.

[63] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.

[64] T. Chen, C. Guestrin, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*, **2016**, pp. 785–794.

[65] P. Geurts, D. Ernst, L. Wehenkel, *Mach. Learn.* **2006**, *63*, 3–42.

[66] F. Pedregosa, G. Varoquaux, *J. Mach. Learn. Res* **2011**, *12*, 2825–2830.

[67] D. R. Lide, *Handb. Chem. Phys.* **2003**, *53*, 2616.

[68] K. Takahashi, I. Miyazato, S. Nishimura, J. Ohyama, *ChemCatChem* **2018**, *10*, 3223–3228.

[69] K. D. Campbell, E. Morales, J. H. Lunsford, *J. Am. Chem. Soc.* **1987**, *109*, 7900–7901.

[70] A. Fisher, C. Rudin, F. Dominici, *arXiv:1801.01489* **2018**.

[71] C. Rücker, G. Rücker, M. Meringer, *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

[72] N. V. Queipo, R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, P. Kevin Tucker, *Prog. Aerosp. Sci.* **2005**, *41*, 1–28.

[73] A. I. J. Forrester, A. J. Keane, *Prog. Aerosp. Sci.* **2009**, *45*, 50–79.

[74] J. Mockus, *The Application of Bayesian Methods for Seeking the Extremum*, **1989**.

[75] J. Snoek, H. Larochelle, R. P. Adams, in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, **2012**, pp. 2951–2959.

[76] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. De Freitas, in *Proc. IEEE*, **2016**, pp. 148–175.

[77] G. E. P. Box, N. R. Draper, *Emperical Model-Building and Response Surfaces*, **1987**.

[78] R. H. Myers, D. C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, **1995**.

[79] H. Robbins, *Bull. Am. Math. Soc.* **1952**, *58*, 527–535.

[80] D. A. Berry, B. Fristedt, *Bandit Problems, Sequential Allocation of Experiments*, **1985**.

[81] N. Srinivas, A. Krause, S. M. Kakade, M. Seeger, in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, **2010**, pp. 1015–1022.

[82] J. P. C. Kleijnen, *Eur. J. Oper. Res.* **2009**, *192*, 707–716.

[83] W. Hare, B. Columbia, *Derivative Free and Black Box Optimization Overview of the Field*, **2017**.

[84] F. Hutter, H. H. Hoos, K. Leyton-Brown, *Sequential Model-Based Optimization for General Algorithm Configuration*, **2010**.

[85] D. R. Jones, M. Schonlau, W. J. Welch, *J. Glob. Optim.* **1998**, *13*, 455–492.

[86] C. E. Rasmussen, C. K. I. Williams, *MIT Press* **2005**, DOI 10.1142/S0129065704001899.

[87] F. Hutter, L. Xu, H. H. Hoos, K. Leyton-Brown, *Artif. Intell.* **2015**, *206*, 79–111.

[88] J. R. Kitchin, J. K. Nørskov, M. A. Barteau, J. G. Chen, *Phys. Rev. Lett.* **2004**, *93*, 156801.

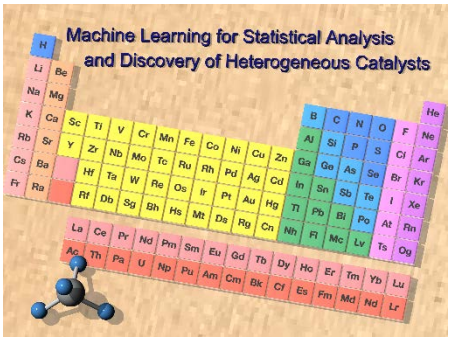
[89] O. Inderwildi, S. Jenkins, *Chem. Soc. Rev.* **2008**, *37*, 2274–2309.

[90] F. Abild-pedersen, *Catal. Today* **2016**, *272*, 6–13.

**Entry for the Table of Contents**

## FULL PAPER

A novel machine learning (ML) approach that uses elemental features as representational inputs instead of catalyst compositions directly was developed using previous experimental catalytic data on oxidative coupling of methane (OCM), water gas shift (WGS), and CO oxidation reactions.




Keisuke Suzuki, Takashi Toyao, Zen Maeno, Satoru Takakusagi, Ken-ichi Shimizu,* Ichigaku Takigawa*

***Page No. – Page No.***

**Statistical Analysis and Discovery of Heterogeneous Catalysts Based on Machine Learning from Diverse Published Data**