Analysis of Updated Literature Data up to 2019 on the Oxidative Coupling of Methane Using an Extrapolative Machine-Learning Method to Identify Novel Catalysts

Shinya Mine,[a] Motoshi Takao,[a] Taichi Yamaguchi,[a] Takashi Toyao,[a] Zen Maeno,[a] S. M. A. Hakim Siddiki,[a] Satoru Takakusagi,[a] Ken-ichi Shimizu,[a, b] and Ichigaku Takigawa[c, d]

We have constructed and analyzed an updated dataset consisting of 4759 experimental datapoints for the oxidative coupling of methane (OCM) reaction based on literature data reported before 2020 (~2019) using machine learning (ML) methods. Several ML methods, including random forest regression (RFR), extra trees regression (ETR), and gradient boosting regression with XGBoost (XGB), were used in conjunction with our proposed approach, in which elemental features are used as input representations rather than inputting the catalyst compositions directly. A recent research trend, namely, the extensive exploration of Mn/Na-containing catalyst systems in recent years due to their high activity and durability, was clearly reflected in the dataset analysis. An ML model for the prediction of the reaction outcome (C₂ yield) was successfully developed, and feature importance scores and SHapley Additive exPlanations (SHAP) values were calculated based on ETR and XGB, respectively, to identify the input variables with the greatest influence on the catalyst performance and observe how these important variables affect the C₂ yield in the OCM. The discovery and optimization of catalytic processes using ML as a “surrogate” model were explored, and promising catalytic system candidates for the OCM reaction were identified. Notably, the developed ML model predicted catalysts containing elements that do not appear in the OCM dataset. This clearly demonstrates desirably high potential of our ML model to enable extrapolative predictions for ML-aided future catalysis research.

Introduction

Molecular/materials informatics has been gaining popularity in materials and molecular science.[1–2] In the data-intensive materials/molecules design approach (also known as the inverse design approach), the desired functionalities for novel materials are first chosen, and promising candidates are subsequently extracted from an experimental or computational database using data science techniques such as machine learning (ML) methods.[3–5] This new methodology, which is an improved version of “quantitative structure-property relationship (QSPR) modeling”,[6] has become a powerful tool to predict the functionalities of substances, including those of previously unreported substances.[7–10] The recent successes of this approach are exemplified by the ML-based prediction of the physical properties of materials, such as their formation energies,[11–13] band gaps,[14–16] and melting temperatures.[17] In cases where a small number of physical properties dominate the performance of a material, materials informatics could represent an ideal tool to identify novel functional compounds. The effective use of ML not only facilitates the discovery of materials, but also helps to establish a deeper understanding of the relationships between the properties of materials and their functionalities and could thus help to establish principles for the design of new materials.

Although “catalysis informatics” is highly related to molecular/materials informatics (or chemoinformatics), catalysis is a multidimensional, multiscale, and dynamic event controlled by
the structure and chemical nature of the catalytic reaction environment.\cite{29–31} Because of this complexity, theoretical models for catalysis, especially heterogeneous catalysis, with reasonable computational costs are not currently available.\cite{32–34} Therefore, ML-assisted prediction of the rates and selectivities of catalytic reactions remains a formidable task, and catalysis informatics is still in its infancy.\cite{35,36–38} The development of methods that bridge the gap between simplified descriptor-based screening and the real-world complexity of catalytic reactions would be highly beneficial to this field. Namely, ML approaches using “real-world” experimental data, rather than models based on computationally derived (i.e., well-behaved) datasets, would be much more useful for heterogeneous catalysis.

Oxidative coupling of methane (OCM) reactions, in which methane (CH\textsubscript{4}) is partially oxidized to C\textsubscript{2} hydrocarbons such as ethane (C\textsubscript{2}H\textsubscript{6}) and ethylene (C\textsubscript{2}H\textsubscript{4}) in a single step, represent a long-standing challenge in the field of catalysis.\cite{39–41} Successful OCM is difficult because the C\textsubscript{2} hydrocarbon products are more reactive than CH\textsubscript{4} and readily undergo nonselective oxidation to CO\textsubscript{2}.\cite{42–44} Thus, despite the significant efforts of both the industrial and academic communities, none of the OCM catalysts developed so far have met the industrial requirements for selectivity and conversion, i.e., ~80\% ethylene product selectivity at >20\% CH\textsubscript{4} conversion using a non-diluted reaction feed.\cite{45}

The OCM reaction is one of the most studied heterogeneous catalytic reactions using ML and other relevant statistical analysis techniques with datasets obtained from high-throughput experiments, published data, and computational studies.\cite{46–49} Many of these studies have utilized the database of Baerns and coworkers, which consists of 1868 OCM reaction datapoints, includes the catalyst composition, experimental conditions, and the catalytic performance of the reactions, and was compiled from a wide range of data published before 2010.\cite{50} Although this database contains a relatively large amount of experimental data, especially given the typical small datasets for heterogeneous catalytic systems, the number of datapoints is still insufficient for effective statistical analysis. Additionally, this literature-derived dataset suffers from bias towards previously reported approaches, a lack of number of datapoints for many elements, and a lack of compositional overlap. The data are also very noisy and inconsistent due to the use of different instruments, procedures, and platforms. These problems are highly related to the main problem, e.g., the insufficient number of datapoints. In addition, the database consists of experimental data reported before 2010 (~2009) and is therefore outdated. In the decade since the data were collected, many additional reports of the OCM reaction have been published due to the recent focus on the abundance of CH\textsubscript{4} in the form of natural gas (shale gas).

In this study, we provide an updated dataset consisting of 4759 experimental OCM reaction datapoints from reports published before 2020 (~2019) and analyze it using ML approaches. Analysis of the updated dataset illustrates the current-state-of-the-art catalytic systems and provides insight into which chemical spaces have already been explored well, which have not, and which spaces merit further investigation. In addition, our ML approach, in which elemental features are used as the input representations rather than directly inputting the catalyst compositions, is used to predict new catalytic OCM reaction systems.

### Data and Methods

#### Catalytic OCM reaction dataset

In our previous study,\cite{51–53} we employed the catalytic OCM reaction data originally reported by Baerns and coworkers\cite{54} for ML analysis. A large number of new OCM reaction data have been reported since this database was published. To discover more effective catalysts using ML, a greater number of datapoints and the most recent data are necessary. In this study, we collected OCM reaction data reported mainly during the past decade, i.e., 2010–2019, using essentially the same format and rules as the original data reported by Baerns and coworkers in MS Excel format (see the Supporting Material for details).\cite{55,56} We included only standard thermal catalysis processes in the dataset. That is, we excluded reaction systems employing oxidants other than O\textsubscript{2} (e.g., N\textsubscript{2}O),\cite{57} transient reactions including chemical looping approaches,\cite{58} membrane reactors,\cite{59,60} dual reactors,\cite{61–63} electrochemical reactors,\cite{64–66} electric fields,\cite{67,68} or catalysts that had undergone special modifications or treatments or had forms that could not be properly described in the dataset format.\cite{69–72} Reports focusing on reactor or process design or non-steady-state reactions were also excluded. For the contact time, we used a volume-based unit [s]; some values reported with a mass-based unit [s] were converted. The contact time value was determined by assuming that each element of the catalyst was present as its most stable oxide if the density was not specified in the original report. We are aware that catalyst properties such as surface area, shape, and number of defect sites, as well as the reactor specifications, can have a significant impact on the reaction outcomes. However, this information is not always reported, and therefore, it is difficult to represent in the dataset. Reporting this information as fully as possible, as well as the establishment of reaction procedure criteria for future catalysis studies, will be highly important to enable the full exploitation of ML and advance the field of catalysis informatics.

The number of OCM publications meeting the aforementioned criteria for each publication year is shown in Figure S1 in the Supporting Information. The OCM was researched extensively in the 1990s. After a period of stagnation in the 2000s, it has become popular again, probably due to the recent rise of shale gas. The periodic tables in Figure S2a, b show the frequency with which each element appears in the pre-2010 and full OCM dataset, respectively. Figure 1 shows histograms of the component elements and important experimental descriptors. The entire OCM dataset is now composed of 4759 experimental data, including 70 elements and 1986 unique catalyst compositions. In the original pre-2010 dataset, the elements Li, La, and Mg appeared most frequently, as catalysts...
such as Li/MgO and La$_2$O$_3$ were known to be highly active for the OCM reaction. In contrast, during the last decade (2010–2019), Na, Mn, W, and Si were most frequently used due to the great recent interest in Mn/Na$_2$WO$_4$/SiO$_2$ catalyst systems. Mn/Na$_2$WO$_4$/SiO$_2$ is currently regarded as the only catalyst system that demonstrates promising activity and stability over long periods of operation; Li/MgO catalysts suffer from intrinsic instability due to excessive loss of Li during the OCM reaction at high temperatures, and La-oxide-based catalysts suffer from selectivity issues. These trends in catalyst composition depending on the research period and previous findings are clearly reflected in the dataset. Variation in the catalyst composition is also clearly limited, as catalyst research relies heavily on previously published data. This implies that, although the addition of recently published data increased the size of the dataset, its information coverage is still limited and biased. That is, it contains a limited number of frequently used elements and many other elements with a small number of datapoints.
Figure 2 illustrates the relationship between the CH₄ conversion and the C₂ selectivity for the entire OCM dataset. As shown in Figure 2a, the C₂ yield is clearly limited by the tradeoff between CH₄ conversion and C₂ selectivity, as the OCM reaction is a selective oxidation reaction. With increasing CH₄ conversion, the C₂ selectivity decreases. In Figure 2b–d, the data have been classified according to relevant criteria such as the reaction temperature, $P_{\text{CH}_4}/P_{\text{O}_2}$, and catalyst type (i.e., Mn and W, Li and Mg, La, and other catalysts). Catalysts containing Mn and W are mainly Mn/Na₂WO₄/SiO₂-type catalysts, while those containing Li and Mg are mainly Li/MgO-type catalysts. The La category consists mainly of La₂O₃-type catalysts. Several trends can be observed in the scatter plots and histograms. For instance, the conversion tends to be higher for reactions performed at high temperature and low $P_{\text{CH}_4}/P_{\text{O}_2}$.

Before evaluating the capability of the various ML methods to predict the catalysis performance, we removed outliers that in which the catalyst contained thorium (Th) or had a contact time [s] ≥ 10 or $P_{\text{CH}_4}/P_{\text{O}_2}$ ≥ 10 because of the scarcity of these data. Datapoints lacking reaction temperature, $P_{\text{CH}_4}/P_{\text{O}_2}$, $P_{\text{total}}$, or contact time values were also removed. In addition, datapoints with a C₂ yield of greater than 35% were removed due to the scarcity of such data and the potential difficulty in reproducing these data, especially under industrially meaningful conditions. As a result, 4387 OCM datapoints were used for the subsequent analyses.

Figure 2. Visualization of entire OCM dataset as scatter plots. (a) CH₄ conversion vs C₂ selectivity; the color of the points corresponds to the C₂ yield. Dashed lines indicate C₂ yield cutoffs. CH₄ conversion vs C₂ selectivity plots in which the points are colored according to (b) reaction temperature, (c) $P_{\text{CH}_4}/P_{\text{O}_2}$, and (d) catalyst type (Mn and W, Li and Mg, La, or other). Histograms of the datapoints are shown at the upper and right sides.
Exploitative and explorative ML models using SWED representations

As mentioned above, the OCM dataset contains a small number of frequently used elements such as Na, Mg, Si, and Mn, along with many other infrequent elements. For example, the elements Tl and I appear only one each in the dataset (Figure S3). This is not surprising, as experimental approaches tend to be biased toward previously reported successful examples. This situation is very common in catalysis research; we observed similar or even more severe elemental overlap in the oxidation of CO by O$_2$ and water gas shift (WGS) reactions.\(^{[49]}\) However, this situation significantly complicates statistically meaningful analysis. Directly applying ML techniques would result in the discovery of a narrow range of small variations on previously studied catalysts, as the predictive models built by ML are representative of the training data. To make full use of the experimental data for the design of new catalysts and identify useful elements that have not been thoroughly explored experimentally, the establishment of a new ML protocol that includes a variety of elements is necessary.

In our previous study,\(^{[49]}\) we proposed a new ML approach, which is denoted as “Proposed method (exploitative)” in Figure 3. This model uses elemental features as the input representations rather than inputting the catalyst compositions directly; in contrast, in the “conventional method”, the elemental compositions of the catalysts were represented by a vector of compositional ratios for all elements under consideration. Namely, in the proposed method (exploitative), the elemental composition ratios are multiplied by elemental descriptors, which are unique for each element. We call this approach the Sorted Weighted Elemental Descriptor (SWED) representation. This SWED representation was applied to base metals and supports in the OCM dataset, and provided higher prediction accuracy than conventional methods using the catalyst compositions as inputs.\(^{[49]}\) Importantly, the newly developed method has the potential to guide catalyst design and discovery and promote the creation of new catalysts in areas where limited catalyst composition overlap exists in the given data.

In this study, we present a ‘proposed method (explorative)’ as depicted in Figure 3, which represents catalysts only with respect to their physico-chemical properties via certain descriptors without directly specifying the individual contributions of distinct elements. As each dimension of the compositional-ratio vector represents the contribution of a distinct element, methods utilizing this vector are incapable of handling elements that are absent or statistically infrequent in the training data. However, the compositional-ratio vector was included in the ‘proposed method (exploitative)’ because the composition information is indispensable to synthesize a catalyst. To overcome these conflicting requirements, the SWED representation was used together with a new procedure to recover the catalyst composition from the SWED representation. The difference between the ‘proposed method (exploitative)’ and ‘proposed method (exploitative)’ might seem subtle but dropping the compositional-ratio input is only made possible by the use of this newly developed recovery procedure. More importantly, focusing only on the physico-chemical properties instead of the types of elements enables a more extrapolative and ambitious exploration beyond the training data. It should be also emphasized that in practice, the descriptors used to generate the SWED representation should be carefully selected and controlled, given that some descriptors, such as atomic number, are too specific and result in the same effect as simply specifying each individual element. It should also be noted here that although relevant “feature-engineering” approaches using

---

**Figure 3. Outline of the three ML methods. Note that the numbers and types of descriptors can be tuned to achieve particular objectives and to control the ‘ambitiveness’ of the predictions.**

---

K is the maximum number of elements in a catalyst. K = 5 in this example. K = 8 used in this paper.
structural and electronic properties have been used to improve the prediction accuracy in the field of materials science,\textsuperscript{[16-25]} they are not aimed at performing predictions that extrapolate across the periodic table.

Methods, descriptors, and parameter settings

The following physical properties, which are readily available from the periodic table and chemical handbooks, were used as the elemental descriptors for the proposed ML models: atomic radius in Å, atomic weight (AW) in g mol$^{-1}$, electronegativity, melting point (m.p.) in K, boiling point (b.p.) in K, density at 25°C in g cm$^{-3}$, enthalpy of fusion ($\Delta H_{\text{fus}}$) in J g$^{-1}$, and ionization energy in eV. The interdependencies of the descriptors were confirmed by constructing the correlation matrix shown in Figure S4.

Tree ensemble methods provided high prediction accuracy in our previous study\textsuperscript{[69]} and have been widely used for understanding data\textsuperscript{[66-68]} in the field of materials science, structural and electronic properties have been used to improve the prediction accuracy.\textsuperscript{[16-25]} Cross-validation is an exhaustively method for finding the optimal hyperparameters of a ML model. A model is built for every combination of the hyperparameters, which are prepared in advance, and each model is evaluated using 5-fold cross-validation. In this study, we applied grid search to the training data in each of the 10-fold cross-validation trials and used the best grid search model to estimate its prediction error. The key hyperparameters tested in the grid searches and their ranges are listed in Table 1. The default values of scikit-learn and XGBoost were used for tuning all other hyperparameters.

### Results and Discussion

#### Quantitative evaluation of the ML predictions for the OCM reactions

The prediction performance of the three ML methods, RFR, ETR, and XGB, in conjunction with the conventional method, proposed method (explorative), and proposed method (exploitative) models described above, was evaluated using the averaged RMSE and the R$^2$ for the training and test parts in 10-fold cross validation. This validation was performed for both the entire OCM dataset and the pre-2010 dataset to evaluate the improvement in the prediction accuracy. The obtained results are shown in Table 2 and Figure 4. The prediction accuracy using the entire OCM dataset was better than that obtained using a limited sample. This evaluation procedure was as follows: The dataset was shuffled randomly and split into 10 equal-sized subsets. One subset was used as test data, and the rest were used as training data. The predicted values for the test data were calculated using the ML models fitted to the training data. The root mean square error (RMSE) and coefficient of determination ($R^2$) between the true values and the predicted ones were computed. This process was repeated ten times, with each of ten subsets being used once as the test data. The results were averaged, and their standard deviations were calculated. Grid search is an exhaustively method for finding the optimal hyperparameters of a ML model. A model is built for every combination of the hyperparameters, which are prepared in advance, and each model is evaluated using 5-fold cross-validation. In this study, we applied grid search to the training data in each of the 10-fold cross-validation trials and used the best grid search model to estimate its prediction error. The key hyperparameters tested in the grid searches and their ranges are listed in Table 1. The default values of scikit-learn and XGBoost were used for tuning all other hyperparameters.

#### Table 1. Ranges used in the hyperparameter searches for the ML methods (if not specified, the default values of scikit-learn were used). 5-Fold cross-validation was used for the inner evaluations in the grid-search.

| Method          | Hyperparameter (tested range)                  |
|-----------------|------------------------------------------------|
| RFR             | n_estimators ∈ [500, 1000]                     |
| XGB             | n_estimators ∈ [500, 1000] max_depth ∈ [6, 7, 8], learning_rate ∈ [0.1, 0.05], subsample ∈ [0.8, 0.9, 1], colsample_bytree ∈ [0.8, 0.9, 1] |
| ETR             | n_estimators ∈ [500, 1000]                     |

#### Table 2. Comparison of prediction accuracy (RMSE) and coefficient of determination ($R^2$) for the C$_2$ yields (%) of the OCM reaction. Two datasets and three ML methods were tested using 10-fold cross-validation. The numbers shown in parentheses are the corresponding $R^2$.

| ML model          | Pre-2010 dataset | Entire OCM dataset |
|-------------------|------------------|--------------------|
|                   | RFR              | ETR                | XGB                |
| Training Error [%] | 1.66 (0.02)      | 0.17 (0.03)        | 1.07 (0.37)        |
| Test Error [%]    | 4.50 (0.38)      | 4.65 (0.50)        | 4.34 (0.34)        |
| Test $R^2$        | 0.536            | 0.504              | 0.567              |
| Proposed Method   |                  |                    |                    |
| Training Error [%] | 1.63 (0.02)      | 0.17 (0.02)        | 0.55 (0.31)        |
| Test Error [%]    | 4.39 (0.43)      | 4.30 (0.52)        | 4.25 (0.41)        |
| Test $R^2$        | 0.557            | 0.575              | 0.583              |
| Proposed Method   |                  |                    |                    |
| Training Error [%] | 1.68 (0.02)      | 0.17 (0.02)        | 0.29 (0.10)        |
| Test Error [%]    | 4.50 (0.48)      | 4.44 (0.50)        | 4.43 (0.52)        |
| Test $R^2$        | 0.536            | 0.547              | 0.547              |

[a] The electronegativity, density, and $\Delta H_{\text{fus}}$ were used as descriptors.
using the pre-2010 dataset, regardless of the ML method or model used. This indicates that the prediction accuracy was improved by the additional data. Among the three ML methods tested in this study, ETR coupled with the proposed method (exploitative) provided the best RMSE values of 3.52% (test error) and $R^2$ value of 0.736 for predicting the C$_2$ yields. More importantly, the proposed method (explorative) successfully gave the comparable predictive accuracies of the test RMSEs, even when using only three descriptors, i.e., electronegativity, density, and $\Delta H_{\text{fus}}$. This implies that the catalyst performance can be predicted even when the catalysts are represented by only these characteristic features without directly specifying individual contributions of distinct elements. Therefore, we can conclude that the proposed method (explorative) is a good surrogate to search for promising prospective novel catalysts. Furthermore, we can expect more ambitious and extrapolative exploration for the proposed method (explorative) with merely three descriptors (electronegativity, density, and $\Delta H_{\text{fus}}$), which is much less than needed for previous results.\[43]

Feature importance scores represent the relative impact of an input value on the output value, and thus provide useful insights into the underlying chemistry of a system. The feature

![Figure 4](image-url). Comparison of the 90%/10% training-test error plots for the three ML methods in the prediction of the OCM catalyst performance (C$_2$ yields) based on ETR. (a) Conventional ML method; (b) proposed ML method (exploitative); (c) proposed ML method (explorative) with all the eight descriptors; (d) proposed ML method (explorative) with three descriptors (electronegativity, density, and $\Delta H_{\text{fus}}$). Blue points represent training data, while red points represent test data.
importance scores were computed from the optimized ETR models using the pre-2010 and the entire OCM dataset. The scores obtained using the conventional method are shown in Figure 5, as this method directly considers the elemental composition and is thus easier to understand the contributing elements in the given data at a glance. The scores based on the proposed methods are provided in Figure S5 and S6, respectively, and indicate which elemental features are important as input representations. Experimental conditions, such as reaction temperature, $P_{\text{CH}_4}/P_{\text{O}_2}$, and contact time, were identified as the most influential descriptors for both datasets. This result clearly shows that OCM reactions are strongly controlled by the experimental conditions, probably due to the involvement of a reactive radical species$^{[42,70]}$. In the initial step of OCM, a CH$_3$ radical species is formed on a catalytic surface; subsequently, gas-phase reactions take place, and the CH$_3$ radical species are believed to combine to form the C$_2$ products. In terms of catalyst composition, alkali metals and alkali earth metals, such as Li, Na, Mg, Ca, Sr, and Ba, were found to be important in both datasets. The importance of elements associated with Mn/Na$_2$WO$_4$/SiO$_2$-type catalysts was greater in the entire dataset, probably due to the great recent attention this catalyst has attracted. Feature importance scores were also calculated for temperature ranges of $T \leq 700$ °C, $700$ °C $< T \leq 800$ °C, $800$ °C $< T \leq 900$ °C, and $900$ °C $< T$ to decrease the contribution of the reaction temperature and are shown in Figure S7. The inclusion of these scores provides additional insights into the most effective catalytic elements in each temperature range, as the optimal temperature differs depending on the catalyst type. It should be noted that high importance scores are typically obtained for features that appear frequently in the dataset. This is especially true for elements, as each catalytic system contains only a few elements.

To inspect the ML predictions in greater detail, we also utilized SHapley Additive exPlanations (SHAP) analysis$^{[71–74]}$. As illustrated in Figure 6, SHAP is a method in which the predicted value is decomposed into the additive sum of contributions from individual feature values. Rather than considering the importance of a feature to the entire model (‘global’ explanations), the SHAP values for the features are obtained for individual predictions (‘local’ explanations). Thus, SHAP can be thought of as a tool to quantitatively analyze the origin of the predicted value for a specific case in terms of individual feature-value contributions. Technically speaking, SHAP is grounded in optimal credit allocation in cooperative game theory and is computed as the difference between the predicted values with and without the feature value averaged over all possible combinations for other feature values. In general, this quantity is computationally intractable (NP-hard), but for tree models such as RFR, ETR, and XGB, efficient (polynomial-time) exact algorithms are available$^{[71–74]}$.

The SHAP values were calculated using the entire OCM dataset based on the conventional ML method and were compared with the feature importance scores. The 20 descriptors with the greatest contributions to predicting the C$_2$ yields are shown in Figure 7. The most important descriptor was the reaction temperature, followed by $P_{\text{CH}_4}/P_{\text{O}_2}$ and La. The results were similar to those obtained using the feature importance score analysis, confirming the high reliability of the method. As described above, SHAP analysis can also be used to visualize the dependence of the predicted C$_2$ yield on the value of each descriptor$^{[71–74]}$. For example, high values (red color in Figure 7) for the feature “reaction temperature” are correlated to high C$_2$ yield (SHAP value). To better understand dependence of the predicted C$_2$ yield on important descriptors, the effects of the four most influential descriptor values are further highlighted in Figure 8. The most important descriptor, reaction temperature, positively influences the C$_2$ yield at high temperatures ($T > 1000$ K), and the C$_2$ yield increases with increasing reaction temperature. In contrast, the C$_2$ yield decreases with increasing $P_{\text{CH}_4}/P_{\text{O}_2}$; specifically, the effect of $P_{\text{CH}_4}/P_{\text{O}_2}$ on the C$_2$ yield is only positive at $P_{\text{CH}_4}/P_{\text{O}_2}$ values below ca. 3.5. In terms of catalyst

![Figure 5](image-url)

**Figure 5.** 20 descriptors with the greatest contributions in predicting the C$_2$ yields calculated using ETR based on the “conventional method” and 10-fold cross-validation from (a) the pre-2010 dataset and (b) the entire OCM dataset.
composition, higher C₂ yield is obtained when a small but non-zero amount of Na is used, while La has a positive effect over almost the entire composition ratio range, with high La content being especially effective for the OCM reaction. This is probably because Na is mainly used as a supported species in catalysts such as Mn/Na₂WO₄/SiO₂, while La catalysts are used directly or hosted as La₂O₃. To obtain further insight into the effect of the catalyst composition, the temperature dependence of the effect of the SHAP descriptors with the greatest contributions, such as La, Na, and Li, on the predicted C₂ yields are provided in Figure 9. Na and Li were found to have positive effects in the middle of the temperature range, especially at 900–1100 K and 850–1050 K, respectively. On the other hand, La has positive effects at temperatures below 850 K. This result shows that the optimal temperature range is highly dependent on the catalyst type; this information should contribute to the understanding of the underlying catalytic processes.

La, Na, Li, and W were identified as the four element descriptors with the greatest contribution in the OCM reaction. They are associated with popular OCM catalysts such as La₂O₃, Mn/Na₂WO₄/SiO₂, and Li/MgO. Although the optimal reaction conditions can differ for each catalyst type, this information should contribute to the understanding of the underlying catalytic processes.

ML models for exploration of novel catalysts using SWED representations

The discovery and optimization of both catalysts and reaction conditions using ML techniques were then explored. Because catalyst research relies heavily on previously published data, catalyst composition tends to be biased towards previous successful catalysts. Since the predictive models built by ML reflect the training data, direct use of ML predictions will result in only small variations on previously investigated catalysts, narrowing the scope of the predicted catalysts.
In our previous study to alleviate this problem,[49] we introduced the SWED representation and an ML model, i.e., the proposed method (exploitative), as a good “surrogate” model to determine the C\textsubscript{2} yield over wider ranges of catalysts beyond the training data. However, it is widely recognized that the direct use of the C\textsubscript{2} yields predicted by ML is usually inadequate to determine which catalysts are worthy of further investigation. Figure 11 shows a typical situation in which the ML regression models are fitted to the training data. Because the models are fitted to represent the average trends in the dataset by minimizing the mean squared errors, the predicted values (blue curve) usually fall between the maximum and minimum values of the training data, and thus cannot exceed the known maximum values (known best y\textsuperscript{*}) by definition. This clearly conflicts with our intention to find catalysts that can improve upon the known catalysts and are hopefully better than any known catalysts. Hence, we instead use the expected improvement (EI)[75] [Eq. (1)]:

\[
EI(x) = E \{ \max (\mu(x) - y^*, 0) \} = \mu(x) - y^* \cdot \Phi \left( \frac{\mu(x) - y^*}{\sigma(x)} \right) + \sigma(x) \cdot \phi \left( \frac{\mu(x) - y^*}{\sigma(x)} \right)
\]

where \( \mu(x) \) and \( \sigma(x) \) are the predicted value and the standard deviation of an ML surrogate for an input \( x \), while the expectation \( E \) assumes a Gaussian distribution with a PDF of \( \phi \) and CDF of \( \Phi \). As illustrated in Figure 11, EI scores can be intuitively considered as a quantity that indicates how much improvement over the current best \( y^* \) can be expected for an input \( x \). This suggests that we should not only investigate the candidates with high statistical confidence from the currently available data (“exploitation”), but also explore beyond the data (“exploration”) by collecting more data from statistically incon-
Figure 8. Effects of the values of the four descriptors with the greatest contribution to the predicted \( C_2 \) yields: (a) reaction temperature, (b) \( P_{\text{CH}_4}/P_{\text{O}_2} \), (c) Na, and (d) La, calculated based on SHAP using the entire OCM dataset. The horizontal axis (x axis) represents the “Feature value”, and the vertical axis (y axis) represents the “SHAP values” + the base value of 9.482. This base value of 9.482 is the average \( C_2 \) yield value of the dataset. Red represents positive SHAP values, while blue represents negative values.
inclusive cases (i.e., those with conflicting data, scarce data, or even no data).

For balancing this exploitation-exploration tradeoff, long-standing efforts have been devoted to many different but interrelated approaches such as sequential model-based optimization (SMBO)\cite{76–78} Bayesian optimization (BO)\cite{79–81}, kriging\cite{82}, design of experiments and multi-armed bandits\cite{83–85}, response surface methods\cite{86,87} and derivative-free and black-box optimizations\cite{88}. In this study, we have also developed an SMBO strategy based on the sequential model-based algorithm configuration (SMAC)\cite{78} to maximize EI. For an ML surrogate, we used RFR or ETR; it should be noted here that the original SMAC\cite{78} used a prediction \( \mu(x) \) by RFR. Since RFR and ETR are both additive ensembles of \( T \) regression trees for which each \( i \)-th tree naturally has a mean prediction \( \mu_i(x) \) and standard deviation \( \sigma_i(x) \), we can calculate Equations (2) and (3)\cite{89}:

\[
\mu(x) = \frac{1}{T} \cdot \sum_{i=1}^{T} \mu_i(x) \quad (2)
\]

\[
\sigma^2(x) = \frac{1}{T} \cdot \left( \sum_{i=1}^{T} \mu_i^2(x) + \sigma_i^2(x) \right) - \mu^2(x) \quad (3)
\]

Although our SMBO strategy was successfully produced and novel candidate catalysts were suggested, the possibility of identifying truly novel catalysts using the developed approach was limited, because it was based on the "proposed method (exploitative)" in which the elemental compositions of the catalysts were still represented directly by a vector of compositional ratios. In this study, we developed a new approach based on "proposed method (explorative)" in order to obtain more ambitious catalyst suggestions far from any of the previously tested catalysts in chemical space. For this purpose, only three

Figure 9. Temperature dependence of the effects of element descriptors ((a) Na, (b) La, and (c) Li) on the predicted C\(_2\) yields calculated based on SHAP using the entire OCM dataset.
descriptors, namely, electronegativity, density, and $\Delta H_{\text{fus}}$ were selected. The use of discrete values such as atomic number as descriptors was avoided, as the inclusion of such values would result in the same situation as inputting the elemental compositions directly (conventional method). In addition to promising catalyst compositions, suggested reaction conditions were also identified, as the reaction conditions affect the outcome significantly.

The core procedure of our SMBO strategy is described in Figure 12. This single procedure aims to identify $K$ points having locally maximum EIs based on an ML surrogate model fitted to $n$ given data points, where the number $K$ is a user parameter. In Step 1, the input representation for the proposed method (explorative), i.e., the SWED representation and experimental conditions, is prepared for $n$ given points as previously described in Figure 3, and an ML surrogate model is fitted to it. Then, in Step 2, those $n$ points are gradually updated toward the direction that can maximize the EI calculated by the ML surrogate in the local-search iterations. The perturbations used for $\delta_1$, $\delta_2$, and $\delta_3$ are made as follows:

- **The SWED part:** the variable range is normalized to [0,1] before the perturbation occurs by adding a Gaussian random

---

**Figure 10.** Potential $C_2$ yield obtained using catalysts that contain Mn and W (i.e., Na$_2$WO$_4$/SiO$_2$ type-catalyst), La (i.e., La$_2$O$_3$ type-catalyst), and Li and Mg (i.e., Li/MgO type-catalyst) as well as other catalysts as calculated using ETR based on the conventional method. Experimental conditions were averaged to evaluate the intrinsic effectiveness of the catalysts.

**Figure 11.** Description of expected improvement (EI). The EI score $E(\max(\mu(x) - y^*, 0))$ corresponds to the average of the improvement in $y$ weighted by the probability of the red-colored region, thus considering not only the predicted values but also their predicted variances. Because the ML prediction (blue curve) usually cannot exceed the best (maximum) value of the training data by definition, it is inadequate to use the predicted values directly to attempt to find data points with higher $C_2$ yields.
number with zero mean and 0.2 standard deviation, rejecting new values outside the interval [0,1].

- The experimental condition part: a variable is selected randomly from the part, randomly set to a possible value if it is discrete and perturbed in the same way as for SWED if it is numerical.

To generate the perturbed SWED value, we used a new method developed for recovering the composition ratios from a given SWED vector (Figure 13). The corresponding composition is then estimated from a given perturbed SWED, before a valid SWED is recalculated from the recovered composition and the elemental descriptors. This local-search update is continued until no further EI improvements are observed. In Step 3, redundant points (the same points from different starting points) and the original points are removed. In Step 4, clustering to K clusters is performed in order to group very similar candidates. Finally, in Step 5, representatives of candidates with the maximal EI are selected from each cluster and sorted in descending order of EI. It should be noted here that an ideal SMBO is a continual alternating process combined with experimental evaluation, where the C2 yields of top-ranking candidates are experimentally validated, and the results are fed into the procedure to design the next promising experiment. It is important to note that in the present study, we always obtained the same solution for a given input in this recovery procedure. However, it is possible that multiple solutions (catalyst compositions) might be obtained in some cases, especially when a smaller number of descriptors is used.

**Self-evaluation of EI maximization by the proposed method (explorative) to spot higher C2 yields**

The following simulation was carried out in order to evaluate how early EI maximization by ML surrogates could potentially identify catalysts and experimental conditions for higher C2 yields: 1) 10 catalysts were selected randomly from the dataset as initial samples. 2) An ML model was fitted to the samples, and the EI was calculated for the catalysts in the dataset that had not yet been selected. 3) The catalyst with the best EI value was added to the set of initial samples before step 2 was repeated. For ML surrogates, we used ETR for the conventional method or proposed methods, in addition to Gaussian process regression (GPR), widely used in Bayesian optimization. ETR was carried out using an n_estimators value of 500 and default

---

**Figure 12.** The core procedure of our SMBO strategy. This single procedure aims to find K points with locally maximum Els based on an ML surrogate model fitted to n given data points, where the number K is a user parameter. The details of perturbations 1, 2, and 3 are described in the main text. The recovery procedure of the composition from SWED is described in Figure 11. The plots on the right-hand side illustrate the local search of SMBO to find the peak points that have locally maximum Els. It should be noted that actual local searches are performed over very high-dimensional representations (SWED + Exp. Cond.), whereas these pictures are oversimplified to one-dimensional plots for an intuitive visual understanding.
values for all other hyperparameters. For GPR, the scikit-learn implementation was used with the Matern 5/2 kernel and the parameters \( \alpha = 0.01 \), \( n_{\text{restarts}} = 10 \), and normalize_y = True.

Figure 14 shows the averaged curves of the highest \( C_2 \) yields for the first 400 iterations when the simulation was run 10 times. The EI maximization by ETR was able to spot higher \( C_2 \) yields much earlier than GPR and random-selection baselines. This result indicates that testing catalysts and experimental conditions with high EIs calculated by ETR surrogates is promising for the sequential design of catalytic experiments. Additionally, the performance of the proposed method (exploitative) was comparable to that of the conventional method and proposed method (exploitative). It is also important to observe that the proposed method (explorative) performed relatively well with only three descriptors. We thus concluded that

---

**Figure 13.** Recovery of compositions estimated from an obtained SWED representation. \( d_e \) indicates the \( p \)-dimensional array of descriptors for element e. The basic idea is a greedy assignment by the smallest errors. Each i-th feature is compared to all possible elements and its ratio \( r \), and the assignment with the smallest error is generated individually from the 1\(^{st} \) to the \( K^{th} \) features. The final assigned ratios are renormalized to sum up to 100\% when the cumulative sum \( s \) is larger than 100\%.

---

**Figure 14.** Comparison of the performance of the first 400 catalysts identified using the EI maximization by the ‘conventional method’, ‘proposed method (exploitative)’, or ‘proposed method (explorative)’ based on ETR, in conjunction with GPR and random-selection baselines. Each curve represents the average of curves obtained from 10 simulations.
focusing on only three selected descriptors, i.e., electronegativity, density, and \( \Delta H_{\text{run}} \), can play a role to expand our SMBO strategy toward more extrapolative and ambitious directions, without significant loss in prediction performance (Table 2) as well as surrogate performance to calculate Es (Figure 14).

Promising catalysts and reaction conditions suggested by the entire OCM dataset

We aimed to identify the most promising catalysts and experimental conditions to worth further investigation according to our entire OCM dataset. For this purpose, we performed the SMBO procedure using the proposed method (explorative) with ETR and the three descriptors electronegativity, density, and \( \Delta H_{\text{run}} \) to suggest a list of distinct promising catalysts (Table 3) that have large Es inferred from our new dataset that covers catalysts reported to date. Elements having the atomic number 3 (Li) through 83 (Bi) except O, Tc, Hg, and noble gasses were used as catalyst component candidates for this study. The original data after preprocessing comprised 4387 points, while our SMBO delivered 4381 points (6 points were the same as in the original data). We also calculated the Es of 10,000 random catalysts, where the compositions are randomly set according to the frequencies of elements in the original data and the experimental conditions by one of the values in the data (Figure S8). The highest-ranking ones were Tb \( ^{10} \) (Tb 100; EI = 0.891), Tm \( ^{11} \) (Pr 100; EI = 0.855), and Zn \( ^{28} \) (In 50.8, Pd 35.1, and Ru 14.1; EI = 0.781), which demonstrates that the local search can identify higher EI points effectively. For assessing the optimal number K of clusters, we performed standard elbow \( ^{[91]} \) and silhouette analyses (Figure S9). \( ^{[92]} \) We concluded that K = 100 is optimal because the elbow of the curve was observed around K = 100–200 and K = 100 was the first point where no clusters with silhouette scores below average were observed when K was varied from K = 200 to K = 100 (number of perturbations: N = 10).

Table 3 lists the 20 OCM system candidates identified as being most promising for future testing, which were suggested using the present method and the entire OCM dataset. A

| Elemental composition | Promoter Preparation method | T [K] | P(CH\(_{2}\)/P(O\(_{2}\)) | \( P_{\text{total}} \) [bar] | Contact time [s] | EI mean sd | Predicted C\(_{\text{yield}}\) [%] 95 %CI lower | 95 %CI upper |
|-----------------------|-----------------------------|-------|----------------------|----------------|----------------|-----------|--------------------------|----------------|
| Mn:72.3 Li:27.7       | B Solid-phase technique     | 1023  | 1.67                 | 1.01           | 1.20           | 2.29       | 21.29 13.52 -0.80 52.19 |
| Sr:50.0 Ge:45.0 Yb:5.0| − Solid-phase technique    | 1023  | 1.99                 | 1.01           | 0.79           | 2.29       | 25.88 13.35 -0.28 52.04 |
| Si:60.9 Na:19.3 Cl:17.2Mn:1.6 W:1.0| − Hydrothermal treatment | 1023  | 1.60                 | 1.01           | 0.02           | 1.67       | 25.55 11.71 2.60 48.50 |
| Mn:80.0 Li:20.0       | Cl Solid-phase technique   | 1023  | 1.96                 | 1.00           | 0.60           | 1.37       | 23.63 12.12 -0.13 47.40 |
| Mg:82.8 Li:17.2       | − Solid-phase technique    | 1043  | 3.00                 | 1.01           | 5.50           | 1.31       | 22.97 12.38 -1.29 47.23 |
| C:42.7 Sc:23.6 Ge:17.8K:10.1 I:5.8| − Physical mixing          | 1023  | 1.60                 | 1.01           | 0.00           | 1.29       | 24.00 11.56 1.34 46.66 |
| Si:45.9 Mg:22.0 Ru:20.5Sc:6.0 Ge:5.7| − Physical mixing          | 1023  | 1.62                 | 1.39           | 0.03           | 1.22       | 23.48 11.69 0.57 46.39 |
| Ge:30.9 Sc:30.7 As:25.3 Be:6.7 I:6.3| − Physical mixing          | 1023  | 1.41                 | 2.75           | 0.13           | 1.05       | 22.75 11.56 0.11 45.40 |
| Si:35.5 Br:32.4 Mg:14.4Al:9.0 Ho:6.5 Y:2.1| − Physical mixing          | 1023  | 1.28                 | 1.01           | 0.00           | 1.01       | 25.54 9.49 6.94 44.13 |
| Mg:36.0 Ge:33.6 Mo:30.3| − Precipitation Ceramic method| 1073  | 2.50                 | 1.01           | 3.60           | 0.95       | 23.97 10.34 3.71 44.23 |
| La:46.4 Ge:27.9 Cu:25.7| − Ceramic method           | 1023  | 1.94                 | 0.68           | 2.40           | 0.95       | 23.16 10.86 1.88 44.44 |
| Sc:36.9 Ca:32.9 Mo:30.2| − Ceramic method           | 1040  | 0.90                 | 0.70           | 1.79           | 0.92       | 23.84 10.31 3.63 44.05 |
| Nd:83.6 Ge:16.4       | − Hydrothermal treatment   | 1100  | 3.95                 | 2.24           | 0.35           | 0.92       | 22.20 11.38 -0.10 44.50 |
| Si:34.4 Ca:29.1 Ge:23.2 Nb:9.7 As:3.6| − Physical mixing          | 1002  | 1.39                 | 1.90           | 0.02           | 0.91       | 20.33 12.55 -4.26 44.92 |
| C:45.8 Sc:20.8 Ge:20.3Mo:7.7 Nb:5.3| − Physical mixing          | 1015  | 0.89                 | 2.96           | 0.04           | 0.91       | 20.79 12.22 -3.16 44.75 |
| Sc:49.4 Au:34.1 Ge:16.5| − Ceramic method           | 1023  | 2.00                 | 1.01           | 2.00           | 0.90       | 22.62 11.00 1.06 44.19 |
| C:38.8 Sc:31.5 Ge:16.0As:7.9 Rh:5.8| − Physical mixing          | 1017  | 0.63                 | 3.02           | 0.07           | 0.89       | 21.32 11.83 -1.88 44.51 |
| Mo:38.9 V:37.9 Ge:23.2| − Ceramic method           | 1048  | 2.00                 | 0.85           | 2.03           | 0.89       | 22.39 11.14 0.57 44.22 |
| Sr:45.7 Ge:33.7 As:20.6| − Ceramic method           | 1023  | 1.99                 | 0.69           | 1.20           | 0.89       | 22.39 11.14 0.56 44.22 |
| Sc:38.7 Mo:37.5 Ca:23.9| − Ceramic method           | 1023  | 1.96                 | 1.01           | 4.10           | 0.88       | 21.29 11.80 -1.85 44.42 |
system consisting of a catalyst composed of 72.3 mol% Mn and 17.7 mol% Li with a promoter B, a reaction temperature of 1023 K, a \( P(\text{CH}_4)/P(\text{O}_2) \) value of 1.67, total pressure of 1.01 bar and a contact time of 1.20 s had the greatest EI value. The second-best system was 50.0 mol% Sr, 45.0 mol% Ce, and 5.0 mol% Yb, with a reaction temperature of 1023 K, a \( P(\text{CH}_4)/P(\text{O}_2) \) value of 1.99, total pressure of 1.01 bar and a contact time of 0.79 s. Note that oxygen was not included in the elemental compositions. In order to investigate the factors responsible for the promising EI values of these catalysts, SHAP analysis was employed to analyze the suggested catalyst systems (Figure S10). For the most promising catalytic system, which had a catalyst composition of 72.3 mol% Mn and 17.7 mol% Li, the \( P(\text{CH}_4)/P(\text{O}_2) \) value of 1.67 was identified as the most influential factor in increasing the \( C_4 \) yield. The addition of boron (B) as a promoter was also found to positively influence the \( C_4 \) yield. Note that there is also a “Promoter” column for the pre-2010 dataset, and the “Promoter” seems to have a relatively high impact on the predictions. On the other hand, the use of the solid-phase technique as the preparation method was found to decrease the SHAP value by 1.79, that is, it negatively influenced the \( C_4 \) yield. Among the elemental descriptors (i.e., catalyst composition), the density of Li and \( \Delta H_m \) of Mn were found to be important in increasing the \( C_4 \) yield. For the second-best catalytic system, which had a catalyst composition of 50.0 mol% Sr, 45.0 mol% Ce, and 5.0 mol% Yb, \( P(\text{CH}_4)/P(\text{O}_2) \) was also identified as the most influential factor in increasing the \( C_4 \) yield, indicating the importance of this descriptor. Among the elemental descriptors, the electronegativity of Sr, density of Yb, and \( \Delta H_m \) of Sr were found to have an important positive influence on the \( C_4 \) yield. In contrast, similarly to in the case of the Mn and Li catalytic system, the use of the solid-phase technique as the preparation method decreased the SHAP value (\( C_4 \) yield).

It is important to mention that the predictions made in this study include features that do not appear frequently in the literature. In addition, As, which is not included in the original entire OCM dataset, newly appeared in the list of suggested catalyst compositions (Table 3). Moreover, other originally unseen elements such as Hf, Se, Os, and Pm were also observed in the 4381 catalysts suggested by the SMBO procedure before the clustering procedure (Figure S11). Although the actual use of these toxic and/or unstable elements would most likely not be ideal, this extrapolative prediction result clearly highlights high potential of our proposed method (explorative) to enable ambitious predictions for ML-aided future catalysis research.

For comparison purposes, the conventional method and proposed method (exploitative) were also used to generate a list of the 20 most promising candidate catalyst systems, as shown in Tables S2 and S3. The ML predictions obtained using the proposed method (exploitative) and the three descriptors are also given in Table S4. Note that this proposed (exploitative) method with the three descriptors showed almost the same prediction accuracy (RMSE value of 0.76% for training data and 3.51 for test data) as the proposed method (exploitative) using the full set of descriptors. Although these prediction results have some similarities, they are essentially different. The proposed method (explorative) provided more ambitious predictions than the others, especially the conventional method, as the lower bounds of the 95% confidence intervals (CIs) of the predicted yields for the proposed method (explorative) tend to be low. This suggests that the proposed method (exploitative) attempts to propose catalytic systems in the relatively unexplored areas in the given chemical space, while the conventional method makes conservative predictions. In real catalysis research, these different ML methods should be used in the appropriate phase of research to achieve particular objectives and to control the ‘ambitiousness’ of the predictions. It is also important to note that at the current stage of catalysis informatics, expert intuition should be utilized rather than fully relying on ML predictions. It would be highly valuable to select some catalytic systems from the proposed lists using expert intuition, experimentally evaluate them, and add the experimental outcomes to the OCM dataset, and make additional predictions. Repeating this procedure many times should eventually lead to the discovery and development of novel OCM catalytic processes with excellent performance.

Finally, we attempted to optimize the catalyst composition of Mn/NaWO\(_4\)/SiO\(_2\)-type catalysts. Although the new ML model was developed to explore new catalytic systems, it is still difficult to actually identify truly high-performance catalytic systems, because so much effort has already been made for the OCM reaction in the last decades. Moreover, in general, ML is representative of the given data, and it remains difficult to identify novel catalysts far away from the input data. Conversely, ML is generally good at optimizing values such as catalyst compositions and catalysis conditions within a limited chemical space. We thus focused on the popular Mn/NaWO\(_4\)/SiO\(_2\)-type catalysts, which are regarded as the most effective OCM catalysts to date, using combinatorial optimization based on ETR with proposed method (explorative) with all the eight descriptors. The total number of Mn/NaWO\(_4\)/SiO\(_2\) type catalysts in our dataset was 993. For the ML procedure, the composition ratios of Mn and W were changed in 0.05 mol% steps, and the EI value for each composition was obtained. During this procedure, the mol% of Na was set to be “100-(Mn + Na + W)”, as we assumed Mn/NaWO\(_4\)/SiO\(_2\)-type catalysts. That is, we explored the following compositions: (Mn, W, Na, Si) = (0, 0, 0, 100), (0.05, 0, 0, 99.95), (0.10, 0, 0, 99.9), ……(0, 0.05, 0.10, 99.85), (0.05, 0.05, 0.10, 99.8)……(20, 10, 10, 60). To simplify the search, the reaction conditions were set as \( T = 800^\circ\text{C} \), \( P(\text{CH}_4)/P(\text{O}_2) = 4 \), \( P_{\text{total}} = 1\) bar, and contact time = 0.1 s. The obtained results are shown in Figure 15. A catalyst composed of Na 2.0 mol%, W 1.0 mol%, Mn 2.2 mol%, and Si 94.8 mol% exhibited the best EI score of 2.338. Importantly, catalysts with Mn ratios from 2.15 mol% to 2.65 mol% and W ratios from 0.80 mol% to 1.0 mol% provided high EI values of >2.0. Note that the EI values for this analysis are not comparative with those provided in Table 3 because we used different datasets (entire dataset for Table 3 and only Mn/NaWO\(_4\)/SiO\(_2\)-type catalysts for Figure 15).
Conclusions

An OCM dataset consisting of nearly 5000 experimental data-points was analyzed using machine learning (ML) methods. A new ML approach that considers elemental features as input representations (named as Sorted Weighted Elemental Descriptor (SWED) representation) rather than inputting the catalyst compositions directly was developed, and its value was demonstrated. Among the ML methods tested, ETR exhibited the best performance in predicting the \( \text{C}_2 \) yield in the OCM reaction. Feature importance scores and SHAP values were calculated to obtain insights into which input variables had the greatest influence on the \( \text{C}_2 \) yield and the relationships between their values and the \( \text{C}_2 \) yield. Our developed ML model, denoted as ‘proposed method (explorative)’, was demonstrated to be effective for predicting novel promising catalyst candidates that include elements unseen in the original dataset for future studies. In future catalysis research, such data analysis approaches should be integrated seamlessly with conventional experimental and theoretical investigations.

Supporting Information

Additional data are given in the Supporting Information. All our scripts and data are available on github (https://github.com/mts-uw/OCM).

Acknowledgements

This work was supported by the KAKENHI grants 17H01783, 17K19953, 19K05556, 20H02518, 20H02775, 20H05962, and 20K0111 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” (JPMXP0112101003), as well as by the JST-CREST projects JPMJCR15P4 and JPMJCR17J3.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: Machine learning · Catalysis informatics · Oxidative coupling of methane (OCM) · SHapley Additive exPlanations (SHAP) · Sequential model-based optimization (SMBO)

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Nature 2018, 559, 547–555.
[2] P. S. Gromski, A. B. Henson, J. M. Granda, L. Cronin, Nat. Chem. Rev. 2019, 3, 119–128.
[3] A. Zunger, Nat. Chem. Rev. 2018, 2, 0121.
[4] B. Sanchez-Lengeling, A. Aspuru-Guzik, Science 2018, 361, 360–365.
[5] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, Y. Jung, Matter 2019, 1, 1370–1384.
[6] T. Le, V. C. EPA, F. R. Burden, D. A. Winkler, Chem. Rev. 2012, 112, 2889–2919.
[87] R. H. Myers, D. C. Montgomery, C. M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 1995.

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

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

[90] C. E. Rasmussen, C. K. I. Williams, *MIT Press* 2005.

[91] R. L. Thorndike, *Psychometrika* 1953, 18, 267–276.

[92] P. J. Rousseeuw, *J. Comput. Appl. Math.* 1987, 20, 53–65.

Manuscript received: April 5, 2021
Revised manuscript received: May 30, 2021
Accepted manuscript online: May 31, 2021
Version of record online: June 21, 2021