# Automatic Feature Engineering for Catalyst Design Using Small Data without Prior Knowledge of the Target Catalysis

Toshiaki Taniike*,†, Aya Fujiwara†, Sunao Nakanowatari†, Fernando G. Escobar‡, Keisuke Takahashi‡

† Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

‡ Department of Chemistry, Hokkaido University, North 10, West 8, Sapporo 060-0810, Japan

**ABSTRACT:** The empirical aspect of descriptor design with limited data in catalyst informatics entails a logical contradiction as it relies on sufficient prior knowledge for exploring the unknown. In this study, we developed a technique for automatic feature engineering (AFE) that works on small catalyst data without requiring any prior knowledge of the target catalysis. This technique generates a large number of features through mathematical operations on general physicochemical features of catalytic components, and extracts the relevant features for the desired catalysis, essentially screening a large number of hypotheses on a machine. AFE yielded reasonable regression results for three types of heterogeneous catalysis: oxidative coupling of methane (OCM), conversion of ethanol to butadiene, and three-way catalysis, where only the training set was swapped. Moreover, through the application of active learning that combines AFE and high-throughput experimentation for OCM, we successfully visualized the machine's process of acquiring precise recognition of catalyst design. AFE is a versatile technique for data-driven catalysis research and a key step towards fully automated catalyst discoveries.

## INTRODUCTION

Natural science has been essentially driven by the sense of individual researchers in designing hypotheses and validating them through experiments. However, the emerging data-driven approach is challenging this tradition by achieving significant success in many fields including catalysis.[1–4] The major bottleneck in data-driven catalysis research, particularly in the context of experimental catalyst discoveries, is the limited availability of data with sufficient quantity and quality for effective machine learning (ML).[5–8] Except in limited cases of crystal structures[9] and organic reactions,[10] this data limitation has rendered the application of deep learning impractical, and forced researchers to address the issue of descriptor design in ML.[1,11] Indeed, descriptor design based on the individual researchers' insights into structure-activity relationships, such as the *d*-band center in metal nanoalloys[12] and the buried volume in organometallic asymmetric catalysis,[13] constitutes a key aspect of the progress of catalyst informatics.[6,14–16] However, such descriptor design is generally challenging and *ad hoc*, as it requires deep domain knowledge to identify all the important factors for the target catalysis.[1,16,17] In particular, practical solid catalysts constitute multiple components that are structured in an ill-defined manner, whose complex interplay over multiple spatiotemporal scales results in the overall catalytic performance.[18,19]

To overcome these issues, in this study, we developed an automatic feature engineering (AFE) technique that works on small data for complex materials such as solid catalysts without requiring any prior knowledge of the target system. The AFE is a pipeline of (i) assigning a series of features to materials of arbitrary compositions, (ii) synthesizing a large number of higher-order features considering nonlinear and combinatorial effects, and (iii) selecting a feature subset in the context of supervised ML. We investigated the applicability of AFE for various heterogeneous catalysis with different catalyst designs. Moreover, its extension to active learning in combination with high-throughput experimentation (HTE) was carried out to refine a feature set and obtain a globally fit model.

## RESULTS AND DISCUSSION

Figure 1a depicts the workflow of AFE. Here, we consider supported multi-element catalysts as a typical example, whose dataset lists the elemental composition and performance of individual catalysts. Proposing a feature of these catalysts is equivalent to hypothesizing its importance. This AFE technique is based on the premise of our scarce knowledge of the system, which is not unusual in today's R&D with the continuously emerging demands over a short period of time. The first step of AFE involves the assignment of primary features to the catalysts by computing commutative operations of a feature library such as maximum and weighted average to account for the notational order invariance and elemental compositions of the catalysts.[20] The feature library collects all possible features of the catalyst constituents (such as the properties of elements and molecules) from all available sources, assuming that all features are equally probable. In the next step, higher-order features, also called compound features,[21–23] that are arbitrary functions of the primary features (first-order) and products of two or more of these functions (second or higher-order) are

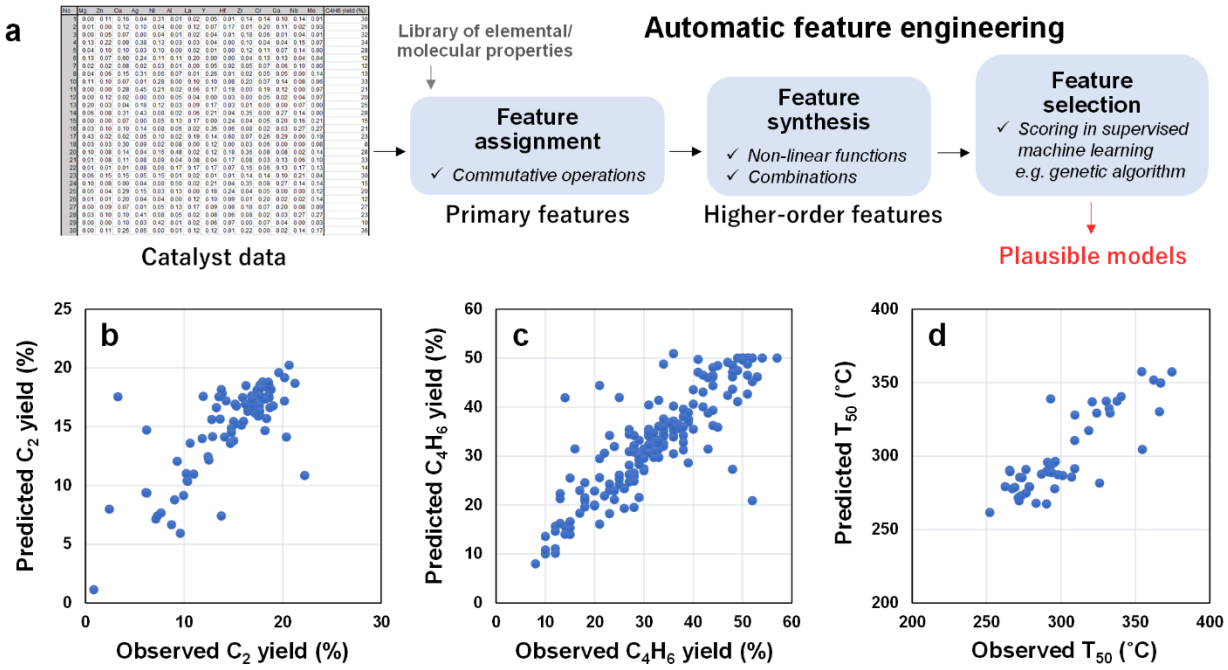


**Figure 1.** Automatic feature engineering (AFE) and its demonstration. (a) Schematic of the AFE pipeline. Prediction of (b) $C_2$ yields in the oxidative coupling of methane (OCM), (c) butadiene yields in ethanol conversion, and (d) light-off temperatures for NO conversion in three-way catalysis. Eight features that minimized the mean absolute error (MAE) in leave-one-out cross-validation (LOOCV) with Huber regression were selected from 5,568 first-order features.

synthesized. This considers the non-linear and combinatorial aspects of the problem and compensates for the shortage of the expressive power of simple ML models suitable for small data. In the final step, the optimum feature combination that maximizes the performance of supervised ML is selected from the large pool of the features (typically $10^3$–$10^6$). Hence, AFE generates a vast number of features (hypotheses) and recommends the most plausible combination in the context of supervised ML. In the literature, preselected physical properties of elements have been employed to describe multi-element catalysts.[24–26] However, these properties have been hardly utilized to systematize feature engineering through the synthesis and screening of a large number of features. AFE was demonstrated on three HTE datasets of supported multi-element catalysts for different catalysis[27–32] (Figures 1b–d, the datasets are given in Tables S1–3). 5,568 first-order features were created by applying 8 types of commutative operations and 12 types of functions to 58 features of elements stored in XenonPy.[33] Then, eight features were selected to minimize the mean absolute error (MAE) in leave-one-out cross-validation (LOOCV) with Huber regression. Huber regression[34] not only mitigates the risk of overfitting on small data but also provides robustness against experimental errors and singular catalysts. Refer to the Method section and Table S4 of Supporting Information (SI) for further details. Reasonable regression results in all cases evidenced the versatility of the method in tailoring the features for individual catalysis without prior knowledge.

Researchers cannot exclude alternative hypotheses when available data is limited. Similarly, AFE proposes different models with similar scores when the training data is limited in size or distribution. An active-learning strategy enables AFE to exclude locally fit models and identify a globally fit model, i.e., the true hypothesis set. This was practiced using the oxidative coupling of methane (OCM) dataset shown in Table S1. The dataset includes the $C_2$ yield of catalysts with up to three elements selected from an element library and supported on BaO, each at a fixed amount. [27] Initially, eight first-order features were selected based on LOOCV-MAE in Huber regression on a given training dataset. Then, twenty catalysts were prepared and evaluated through HTE, where eighteen were selected via farthest point sampling (FPS) in the selected feature space and two were chosen based on their highest absolute errors in the regression. The obtained data were fed back to AFE to update the feature space (Figure 2a). This process was repeated four times, resulting in the addition of 80 new catalysts (Table S5). Figures 2b and 2c provide a summary of the relevant scores and individual test results, respectively. In the first cycle, the largest diversification of catalyst composition driven by FPS moderately increased the $MAE_{train,CV}$ values, but subsequent cycles did not largely change these values. The final $MAE_{trainv,CV}$ values (2.2–2.3%) were higher than the typical experimental error (1.0–2.0%) partly because the linear model failed to capture 0% $C_2$ yield data. Excluding these data points reduced the $MAE_{CV}$ to ~1.9%. The changes in the test score were larger than those in the training and CV scores. Several extrapolations occurred during the first cycle, where the predicted yield was >30% or <0%, resulting in an extremely large $MAE_{test}$. As the cycle progressed and the catalysts in the training dataset diversified sufficiently, these extrapolations disappeared and the difference between the observations and predictions decreased monotonically. The Pearson's correlation coefficient between the regression models increased from 0.6 in Cycles 0 and 1 to 0.9 in Cycles 3 and 4, indicating the convergence of the feature engineering towards a global model.

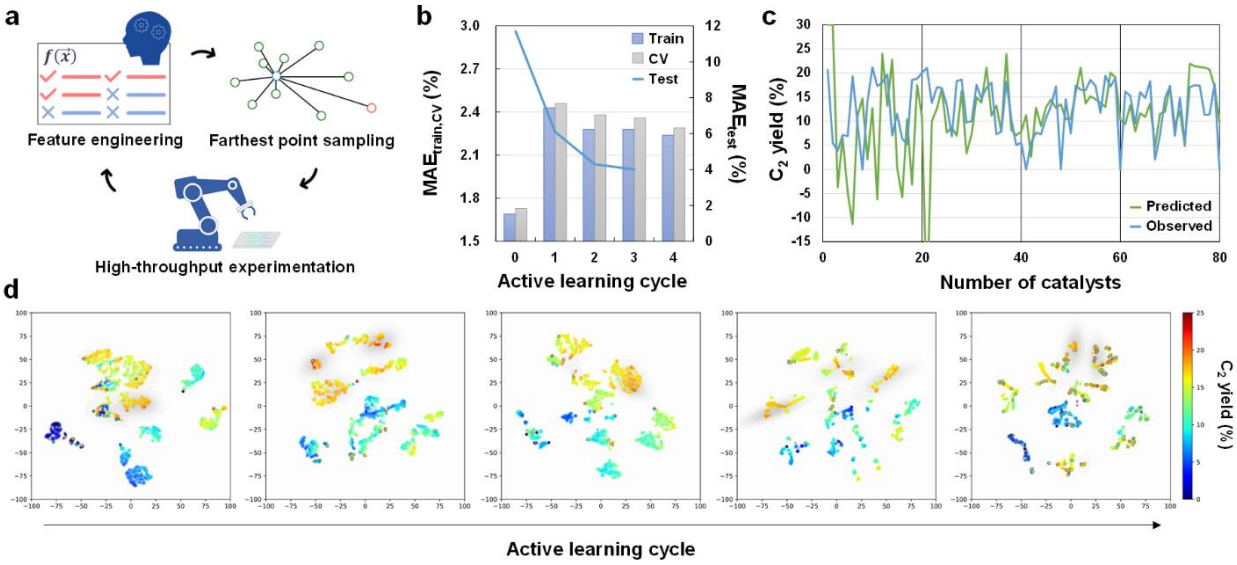


**Figure 2.** Active learning implemented for the OCM catalyst design. (a) Schematic of the active learning loop. The feature engineering was repeated five times with the data of 20 catalysts added per update. The model scores and the testing results are shown in (b) and (c), respectively. Eight features were selected from 5,568 first-order features to minimize the MAE in LOOCV with Huber regression. The development of the feature engineering and prediction is visualized based on t-distributed stochastic neighbor embedding (t-SNE) in (d). The circled data points are the test results except for the last cycle, which used the training data instead. The color reflects the predicted or observed $C_2$ yield, and the counters indicate the Gaussian kernel density estimation for the $C_2$ yield above 18%.

Figure 2d visualizes the progress of the feature engineering using t-distributed stochastic neighbor embedding (t-SNE).[35] The plot shows all 4,060 catalysts in the library (including both tested and untested ones) with the color indicating the predicted $C_2$ yield, and circled data points representing the test results. With the advancement of active learning, the data were divided into a larger number of clusters, which represents the machine's process of refining a feature space to distinguish the catalysts better through distinct composition-performance relationships. Then, the question is how does the machine perceive the composition-performance relationships? This was addressed in two steps. First, the dataset was subjected to manual statistical analysis, as shown in Figure S1. Early transition metals such as Mo and Zr and heavy alkali metals such as K and Cs are attributed high performance (Figures S1a,b). This is because early transition metals can form oxometalate anions active for OCM when they were combined with Ba in the support or other supported elements with low electron affinity.[28,36,37] Alkali metals can enhance the $C_2$ selectivity by strengthening the basicity of alkali earth metal oxides.[38–40] In contrast, late transition metals (excluding Zn with completely filled 3d orbitals) tended to decrease the $C_2$ yield with increasing group number (Figures S1a,c), as they act as combustion catalysts.[41] Next, keeping the abovementioned researcher's observations in mind, the machine's perception was interpreted by analyzing the distribution of individual elements in the feature space (Figure S2). Figure 3 summarizes the regions where individual elements are concentrated after active learning, which decodes the machine perception. It can be seen that late transition metals form separate clusters, whereas Mo and W are concentrated in narrow regions, indicating that the machine recognizes these elements as having differently significant impacts on the performance. In contrast, elements with a wide spatial distribution either have limited data points (e.g. La) or exhibit significantly different performances depending on their combination (e.g. Mg and Mn). Elements with overlapping distributions are not only similar in their physicochemical properties but also in their impact on the catalytic performance. For example, the high-performing K and Cs have overlapping distributions, whereas the less effective Li and Na are separated from them. These observations align with the researchers' understanding acquired from Figure S1.

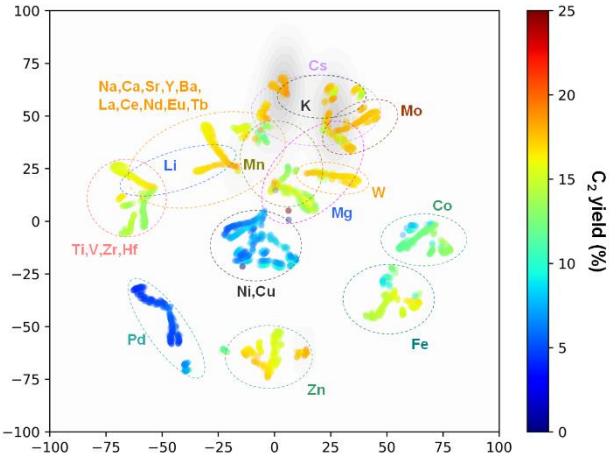


**Figure 3.** Machine perception of the OCM catalyst design. The feature space of the latest model is visualized by t-SNE, along with the Gaussian kernel density estimation for the $C_2$ yield above 18%. The dotted lines indicate the regions where catalysts containing individual elements are concentrated.

Application of the same analysis to the unselected feature set and the feature set selected before active learning (Figure S3) revealed the essentiality of both feature engineering and active learning in achieving that level of discrimination. Eventually, AFE transformed general physicochemical knowledge of elements into an OCM-specific one, while active learning enhanced the machine's accuracy in discriminating elements.

## CONCLUSIONS

In summary, we have developed and demonstrated AFE as a versatile technique to enable effective ML for small data of solid catalysts with diverse compositions. It designs highly expressive features specific to a given catalyst system without requiring prior knowledge of the system. The availability of process-consistent datasets obtained through HTE was crucial in the development of AFE. Active learning that integrated AFE, FPS, and HTE in a loop helped eliminate alternative hypotheses and identified a true hypothesis set that applies to diverse catalysts. This is attributed to the ability of the machine to develop a feature or knowledge space for recognizing composition-performance relationships of catalysts. Incorporating AFE into automated experiments[42] would enable highly efficient autonomous catalyst design. Moreover, the knowledge acquired for a specific system will not only help predict the performance of unknown compositions in the same system but also assist in acquiring knowledge for different systems through transfer learning. As the machine accumulates knowledge across many catalytic systems, it will ultimately develop comprehensive catalytic knowledge and achieve catalyst development freed from researchers' experiences and knowledge.

## ASSOCIATED CONTENT

**Supporting Information**. HTE datasets; Methods; Results of additional analysis

## AUTHOR INFORMATION


### Corresponding Author

* taniike@jaist.ac.jp

### Author Contributions

T.T., A.F, S.N., F.G.E., and K.T. contributed equally to this paper.



### Funding Sources

The authors acknowledge funding from the Japan Science and Technology Agency (JST) CREST (Grant number JPMJCR17P2) and JST Mirai Program (Grant Number JPMJMI22G4).


### Notes

The authors declare no competing financial interest.


## ACKNOWLEDGMENT

The work of A.F. and S.N. was supported by JST SPRING (Grant Number JPMJSP2102) and JSPS KAKENHI (Grant Number JP22J15549), respectively.


## REFERENCES


(1) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54.

(2) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.

(3) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260–2297.

(4) Takahashi, K.; Ohyama, J.; Nishimura, S.; Fujima, J.; Takahashi, L.; Uno, T.; Taniike, T. Catalysts Informatics: Paradigm Shift Towards Data-Driven Catalyst Design. *Chem. Commun.* **2023**, *59*, 2222–2238.

(5) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144*, 4819–4827.

(6) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5*, 83.

(7) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647.

(8) Taniike, T.; Takahashi, K. The Value of Negative Results in Data-Driven Catalysis Research. *Nat. Catal.* **2023**, *6*, 108–111.

(9) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158–10168.

(10) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370–377.

(11) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph Neural Networks for Materials Science and Chemistry. *Commun. Mater.* **2022**, *3*, 93.

(12) Hammer, B.; Nørskov, J. K. Theoretical Surface Science and Catalysis—Calculations and Concepts. *Adv. Catal.* **2000**, *45*, 71–129.

(13) Clavier, H.; Nolan, S. P. Percent Buried Volume for Phosphine and N-Heterocyclic Carbeneligands: Steric Properties in Organometallic Chemistry. *Chem. Commun.* **2010**, *46*, 841–861.

(14) Ringe, S. The Importance of a Charge Transfer Descriptor for Screening Potential $CO_2$ Reduction Electrocatalysts. *Nat. Commun.* **2023**, *14*, 2598.

(15) Santiago, C. B.; Guo, J. Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398–2412.

(16) Liu, J.; Luo, W.; Wang, L.; Zhang, J.; Fu, X.-Z.; Luo, J.-L. Toward Excellence of Electrocatalyst Design by Emerging Descriptor-Oriented Machine Learning. *Adv. Funct. Mater.* **2022**, *32*, 2110748.

(17) Zhang, Y.; Peck, T. C.; Reddy, G. K.; Banerjee, D.; Jia, H.; Roberts, C. A.; Ling, C. Descriptor-Free Design of Multicomponent Catalysts. *ACS Catal.* **2022**, *12*, 10562–10571.

(18) Urakawa, A.; Baiker, A. Space-Resolved Profiling Relevant in Heterogeneous Catalysis. *Top. Catal.* **2009**, *52*, 1312–1322.

(19) Wada, T.; Funako, T.; Chammingkwan, P.; Thakur, A.; Matta, A.; Terano, M.; Taniike, T. Structure-Performance Relationship of $Mg(OEt)_2$-Based Ziegler-Natta Catalysts. *J. Catal.* **2020**, *389*, 525–532.

(20) Liu, C.; Fujita, E.; Katsura, Y.; Inada, Y.; Ishikawa, A.; Tamura, R.; Kimura, K.; Yoshida, R. Machine Learning to Predict Quasicrystals from Chemical Compositions. *Adv. Mater.* **2021**, *33*, 2102507.

(21) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

(22) Kim, C.; Pilania, G.; Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory Using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* **2016**, *28*, 1304–1311.

(23) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Sci. Rep.* **2016**, *6*, 19375.

(24) Suzuki, K.; Toyao, T.; Maeno, Z.; Takakusagi, S.; Shimizu, K.; Takigawa, I. Statistical Analysis and Discovery of Heterogeneous Catalysts Based on Machine Learning from Diverse Published Data. *ChemCatChem* **2019**, *11*, 4537–4547.

(25) Williams, T.; McCullough, K.; Lauterbach, J. A. Enabling Catalyst Discovery through Machine Learning and High-Throughput Experimentation. *Chem. Mater.* **2020**, *32*, 157–165.

(26) Ishioka, S.; Fujiwara, A.; Nakanowatari, S.; Takahashi, L.; Taniike, T.; Takahashi, K. Designing Catalyst Descriptors for Machine Learning in Oxidative Coupling of Methane. *ACS Catal.* **2022**, *12*, 11541–11546.

(27) Nguyen, T. N.; Nakanowatari, S.; Tran, T. P. N.; Thakur, A.; Takahashi, L.; Takahashi, K.; Taniike, T. Learning Catalyst Design Based on Bias-Free Data Set for Oxidative Coupling of Methane. *ACS Catal.* **2021**, *11*, 1797–1809.

(28) Nakanowatari, S.; Nguyen, T. N.; Chikuma, H.; Fujiwara, A.; Seenivasan, K.; Thakur, A.; Takahashi, L.; Takahashi, K.; Taniike, T. Extraction of Catalyst Design Heuristics from Random Catalyst Dataset and their Utilization in Catalyst Development for Oxidative Coupling of Methane. *ChemCatChem* **2021**, *13*, 3262–3269.

(29) Takahashi, L.; Nguyen, T. N.; Nakanowatari, S.; Fujiwara, A.; Taniike, T.; Takahashi, K. Constructing Catalyst Knowledge Networks from Catalyst Big Data in Oxidative Coupling of Methane for Designing Catalysts. *Chem. Sci.* **2021**, *12*, 12546–12555.

(30) Takahashi, K.; Fujima, J.; Miyazato, I.; Nakanowatari, S.; Fujiwara, A.; Nguyen, T. N.; Taniike, T.; Takahashi, L. Catalysis Gene Expression Profiling: Sequencing and Designing Catalysts. *J. Phys. Chem. Lett.* **2021**, *12*, 7335–7341.

(31) Jayakumar, T. P.; Babu, S. P. S.; Nguyen, T. N.; Le, S. D.; Taniike, T. Exploration of Ethanol-to-Butadiene Catalysts by High-Throughput Experimentation and Machine Learning. *ChemRxiv (Catalysis)*, May 16, 2023. DOI: 10.26434/chemrxiv-2023-0lhl8 (accessed July 1, 2023).

(32) Le, D. S.; Ton, N. N. T.; Seenivasan, K.; Chammingkwan, P.; Praserthdam, S.; Taniike, T. High-Throughput Screening of Multimetallic Catalysts for Three-Way Catalysis. *Research Square*, May 10, 2023. DOI: 10.21203/rs.3.rs-2601219/v1 (accessed July 1, 2023).

(33) Yoshida, R. XenonPy is a Python Software for Materials Informatics. https://github.com/yoshida-lab/XenonPy (accessed July 1, 2023).

(34) Huber, P. J. Robust Estimation of a Location Parameter. *Ann. Math. Statist.* **1964**, *35*, 73–101.

(35) Maaten, L. V. D.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(36) Wu, J.; Li, S. The Role of Distorted $WO_4$ in the Oxidative Coupling of Methane on Supported Tungsten Oxide Catalysts. *J. Phys. Chem.* **1995**, *99*, 4566–4568.

(37) Ji, S.; Xiao, T.; Li, S.; Chou, L.; Zhang, B.; Xu, C.; Hou, R.; York, A. P. E.; Green, M. L. H. Surface $WO_4$ Tetrahedron: The Essence of the Oxidative Coupling of Methane over M–W–Mn/$SiO_2$ Catalysts. *J. Catal.* **2003**, *220*, 47–56.

(38) Ito, T.; Wang, J.; Lin, C. H.; Lunsford, J. H. Oxidative Dimerization of Methane over a Lithium-Promoted Magnesium Oxide Catalyst. *J. Am. Chem. Soc.* **1985**, *107*, 5062–5068.

(39) Xu, Y.; Yu, L.; Cai, C.; Huang, J.; Guo, X. A Study of the Oxidative Coupling of Methane over SrO-$La_2O_3$/CaO Catalysts by Using $CO_2$ as a Probe. *Catal. Lett.* **1995**, *35*, 215–231.

(40) Ortiz-Bravo, C. A.; Chagas, C. A.; Toniolo, F. S. Oxidative Coupling of Methane (OCM): An Overview of the Challenges and Opportunities for Developing New Technologies. *J. Nat. Gas. Sci. Eng.* **2021**, *96*, 104254.

(41) Choudhary, T. V.; Banerjee, S.; Choudhary, V. R. Catalysts for Combustion of Methane and Lower Alkanes. *Appl. Catal. A Gen.* **2002**, *234*, 1–23.

(42) Trunschke, A. Prospects and Challenges for Autonomous Catalyst Discovery Viewed from an Experimental Perspective. *Catal. Sci. Technol.* **2022**, *12*, 3650–3669.