communications chemistry

ARTICLE

https://doi.org/10.1038/s42004-023-01086-y OPEN

Automatic feature engineering for catalyst design using small data without prior knowledge of target catalysis

Toshiaki Taniike[1✉], Aya Fujiwara[1], Sunao Nakanowatari[1], Fernando García-Escobar[2] & Keisuke Takahashi[2]

The empirical aspect of descriptor design in catalyst informatics, particularly when confronted with limited data, necessitates adequate prior knowledge for delving into unknown territories, thus presenting a logical contradiction. This study introduces a technique for automatic feature engineering (AFE) that works on small catalyst datasets, without reliance on specific assumptions or pre-existing knowledge about the target catalysis when designing descriptors and building machine-learning models. This technique generates numerous features through mathematical operations on general physicochemical features of catalytic components and extracts relevant features for the desired catalysis, essentially screening numerous hypotheses on a machine. AFE yields 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 is swapped. Moreover, through the application of active learning that combines AFE and high-throughput experimentation for OCM, we successfully visualize the machine's process of acquiring precise recognition of the catalyst design. Thus, AFE is a versatile technique for data-driven catalysis research and a key step towards fully automated catalyst discoveries.

[1] Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan. [2] Department of Chemistry, Hokkaido University, North 10, West 8, Sapporo 060-0810, Japan. ✉email: taniike@jaist.ac.jp

Over the years, the trajectory of natural science has been conventionally steered by the intuition of individual researchers, guiding the formulation of hypotheses and their subsequent validation through experimentation. However, with the advent of a data-driven approach, this paradigm is now shifting, challenging established norms and registering significant success across diverse fields, including catalysis[1–4]. Within the realm of data-driven catalysis research, particularly in the context of experimental catalyst discoveries, the limited availability of data characterized by both sufficient quantity and quality for effective machine learning (ML) presents a major hurdle[5–8]. In this context, data typically assume the form of tabular datasets comprising observations (e.g., catalyst samples) and parameters describing these observations (properties of catalysts), commonly referred to as features or descriptors when employed to predict a specific target variable (performance of catalysts) within the framework of supervised ML. In the field of catalysis, data are predominantly categorized into small data, seldom surpassing a thousand observations. This characteristic renders the data unsuitable for the deployment of elaborate ML models with a multitude of adjustable parameters necessary to capture intricate trends. Thus, the design of descriptors that encapsulate the essence of catalysis is imperative for the efficient and accurate capturing of data trends using simple ML models. However, except in limited cases of crystal structures[9] and organic reactions[10], the data limitation has rendered the application of deep learning impractical, prompting researchers to address the fundamental issue of descriptor design in ML[1,11]. Indeed, descriptor design based on 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 in catalyst informatics[6,14–16]. However, such descriptor design is generally challenging and performed ad hoc, as it requires profound domain knowledge to identify all pertinent factors for the target catalysis[1,16,17]. In particular, practical solid catalysts constitute multiple components that are structured in an ill-defined manner, and the complex interplay of these components over multiple spatiotemporal scales results in the overall catalytic performance[18,19]. This intricacy, coupled with data scarcity, elevates the difficulty of crafting descriptors in catalysis, when compared to other fields.

To surmount these challenges, 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 structured pipeline of (i) assigning a series of features to materials of arbitrary compositions, (ii) synthesizing numerous higher-order features considering nonlinear and combinatorial effects, and (iii) selecting a feature subset in the context of supervised ML. This study explores the applicability of AFE across various heterogeneous catalysis scenarios, each characterized by distinct catalyst designs. Furthermore, an extension of AFE to active learning, coupled with high-throughput experimentation (HTE), is implemented to comprehend catalyst design rules and streamline catalyst discoveries.

Results and discussion

Automatic feature engineering. Figure 1a illustrates the workflow of AFE. Here, we consider supported multi-element catalysts as typical examples, wherein the dataset comprises elemental composition and performance data for individual catalysts. While the straightforward and commonly employed approach involves directly using elemental compositions as descriptors in constructing an ML model, this neglects the physical properties of elements, leading to drawbacks such as insufficient prediction accuracy and an inability to handle elements absent in the training data. However, crafting physically meaningful features of catalysts remains challenging, as proposing these features is equivalent to hypothesizing their relevance in the target catalysis. The proposed AFE technique is based on the premise of our scarce knowledge of a system, a common characteristic in today's research and development landscape with continually emerging demands over a short period. The first step in AFE involves assigning primary features to catalysts by computing commutative operations of a feature library, such as a maximum and weighted average. This accounts for notational order invariance (e.g., features of Li–W must be equal to those of W–Li) and the elemental compositions of catalysts (e.g., the features of Li–Li–W must be differentiated from those of Li–W–W)[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], are synthesized. These features are arbitrary functions of primary features (first order) and products of two or more of these functions (second or higher order), addressing the nonlinear and combinatorial aspects of the problem. This compensates for the limited expressive power of simple ML models suitable for small data. A detailed classification of different feature types is presented in Table S1. In the final step, the optimum feature combination that maximizes the performance of supervised ML is selected from a large pool of features (typically 10^3–10^6). Hence, AFE generates a vast number of features (hypotheses) and recommends the most plausible combination within the context of supervised ML. While previous studies have employed preselected physical properties of elements to describe multi-element catalysts[24–26], these properties have been hardly utilized to systematize feature engineering through the synthesis and screening of a large number of features. Herein, AFE was demonstrated using three HTE datasets of supported multi-element catalysts for different catalysis[27–32] (Fig. 1b–d; the datasets are given in Tables S2–4). In particular, 5568 first-order features were constructed by applying eight 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) using Huber regression. Note that Huber regression is a linear regression method that employs the Huber loss instead of ordinary least squares to enhance robustness against outliers[34]. This approach not only mitigates the risk of overfitting on small data owing to its simplicity but also provides resilience against experimental errors and singular catalysts. Note that many of the generated features are inherently ineffective in describing the desired catalysis. However, given the limited knowledge and the fact that algorithm-based filtrations necessarily deteriorate the regression scores, filtering these features prior to feature selection is discouraged. Further details on this aspect are presented in the Methods section. In all cases, reasonable regression results evidenced the versatility of the method in tailoring the features for individual catalysis without prior knowledge (Fig. 1b–d). The MAE values of the obtained models during training and CV were 1.69% and 1.73% in C_2 yields, 3.77% and 3.93% in butadiene yields, and 11.2 °C and 11.9 °C in T_{50} of NO conversion, respectively. Notably, these MAE values are significantly smaller than the span of each target variable and comparable to the respective experimental errors. The remarkable accuracy of the AFE-generated models in CV was unattainable when using catalyst elemental compositions as descriptors, regardless of the ML methods and hyperparameter sets (Fig. S1). In particular, relatively complex methods such as support vector

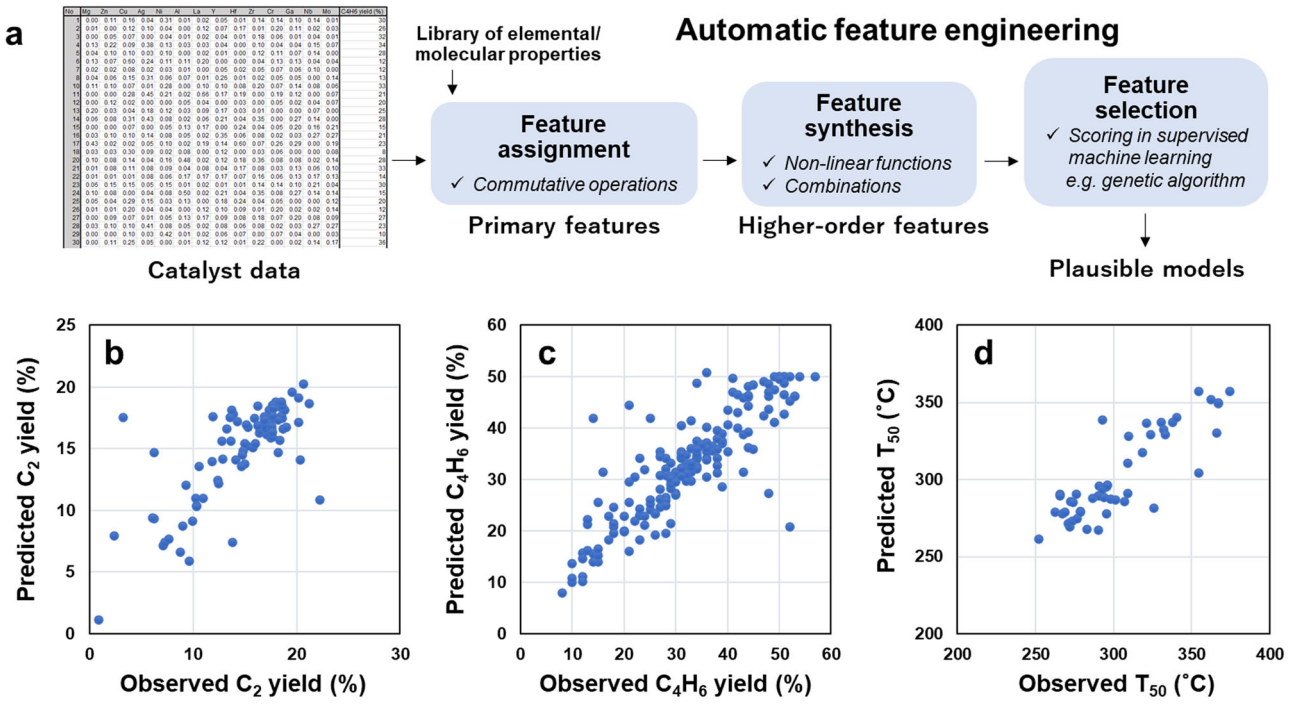

Fig. 1 Automatic feature engineering (AFE) and its demonstration. a Schematic of the AFE pipeline. Prediction of (**b**) C_2 yields in the oxidative coupling of methane (OCM), (**c**) butadiene yields in ethanol conversion, and (**d**) light-off temperatures for NO conversion in three-way catalysis. Eight features that minimized the mean absolute error (MAE) in leave-one-out cross-validation (LOOCV) with Huber regression were selected from 5568 first-order features.

regression and random forest regression exhibited a tendency of overfitting with MAE in training significantly lower than that in CV. By contrast, AFE led to consistently low MAE values in both training and CV, offering a minimal set of engineered features suitable for capturing complex trends with limited data.

Integration with active learning. In scenarios where the available data are limited, researchers cannot disregard alternative hypotheses. Similarly, when the training data are either limited in size or constrained in the diversity of elemental compositions in catalysts, AFE proposes multiple models exhibiting similar scores, even though different feature sets are selected. Although these models demonstrate similar performance in explaining the training data, their predictive behaviors for unknown catalysts can vary significantly. In other words, many of these models are only locally fit, lacking the global characteristics necessary to explain the entire composition. An active learning strategy enables AFE to exclude locally fit models and identify a globally fit model, i.e., the true hypothesis set. Here, this was practised using the oxidative coupling of methane (OCM) dataset (Table S2). 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. Subsequently, 20 catalysts were prepared and evaluated through HTE, among which 18 catalysts 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. Note that FPS adds catalysts that are least similar to those in the training data within the selected feature space, which aids in efficiently excluding models lacking global characteristics. The obtained data were fed back to AFE to update the feature space (Fig. 2a). This process was repeated over four iterations, resulting in the addition of 80 new catalysts (Table S5). A more detailed procedure is presented in Figure S2. Figure 2b, c provides 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 various 0% C_2 yield data (any observed inactivity may be attributed to several reasons). 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}. These extrapolations correspond to the model attempting to explain catalysts entirely beyond its original consideration. 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. 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 feature engineering toward a global model.

Decoding machine's perception. Figure 2d visualizes the progress of feature engineering using t-distributed stochastic neighbor embedding (t-SNE)[35], where the eight features selected during each active learning cycle were reduced in two dimensions, maintaining the pairwise similarities of the catalysts. This approach allowed us to monitor the evolution of the machine's ability to perceive individual catalysts. The plot shows all 4060 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. Leveraging the advancements in active learning, the data were divided into a larger number of clusters, representing the machine's process of refining a feature space to distinguish the catalysts better through distinct

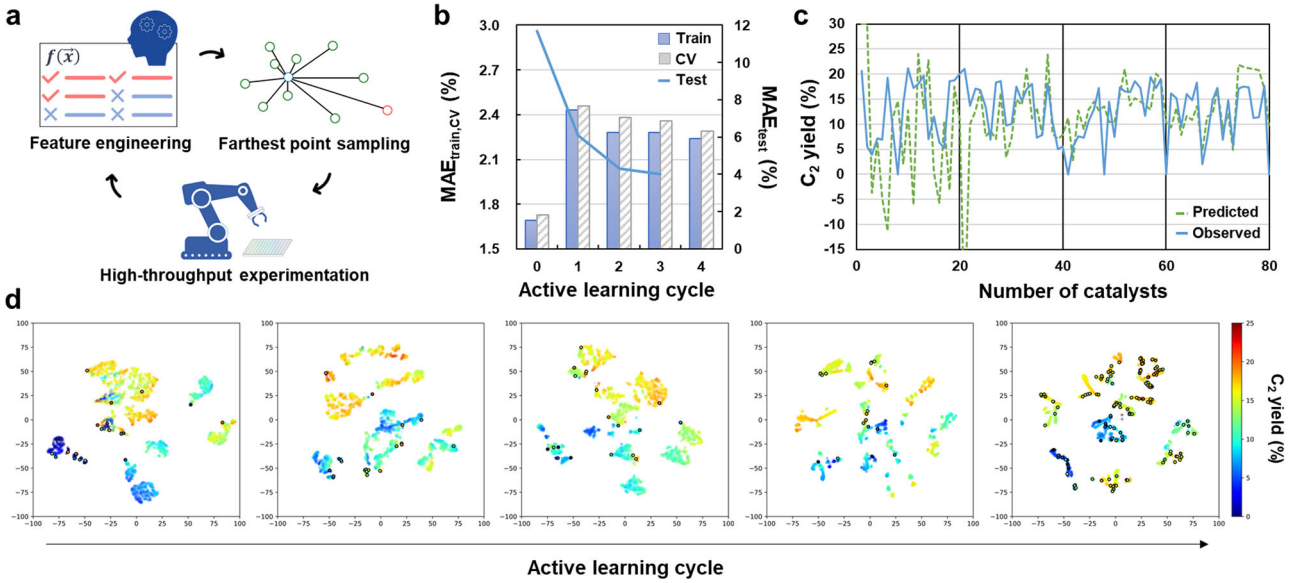

Fig. 2 Active learning implemented for the OCM catalyst design. a Schematic of the active learning loop. The feature engineering was repeated five times with the data of 20 catalysts added per update. The model scores and the testing results are shown in (**b**) and (**c**), respectively. The deviation between predicted and observed C_2 yields decreased monotonically throughout the active learning cycle. (**d**) Eight features were selected from 5568 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). 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. Each t-SNE image delineates how the machine perceives the composition and performance of individual catalysts in each active learning cycle. The increase in the number of clusters during active learning signifies the evolution of the machine's ability to discern diverse catalysts based on their distinct composition-performance relationships.

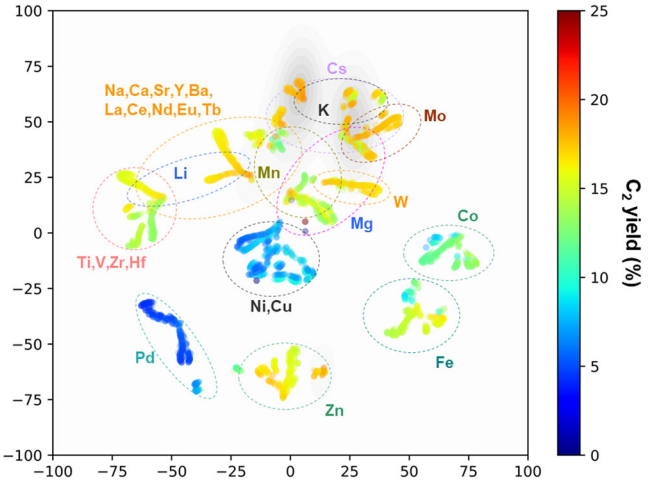

Fig. 3 Machine perception of the OCM catalyst design. The feature space of the latest model is visualized by t-SNE, along with the Gaussian kernel density estimation for the C_2 yield above 18%. The dotted lines indicate the regions where catalysts containing individual elements are concentrated. This visualization illustrates the machine's perception in identifying the composition and performance of catalysts based on specific elements. It showcases common elements found in high and low-performing catalysts, similarities among elements within the feature space, and other pertinent insights.

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 Fig. S3. Early transition metals such as Mo and Zr and heavy alkali metals such as K and Cs are attributed high performance (Fig. S3a, b). This is because early transition metals can form oxometalate anions active for OCM when they are 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]. By contrast, late transition metals (excluding Zn with completely filled 3d orbitals) tend to decrease the C_2 yield with increasing group number (Fig. S3a, 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 (Fig. S4). Figure 3 summarizes the regions where individual elements are concentrated after active learning, which decodes the machine perception. 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. By contrast, elements with a wide spatial distribution either have limited data points (e.g., La) or exhibit significantly different performance 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, high-performing K and Cs have overlapping distributions, whereas the less-effective Li and Na are separated. These observations align with the researchers' understanding acquired from Fig. S3. An application of the same analysis to the unselected feature set and the feature set selected before active learning (Fig. S5) revealed the essentiality of both feature engineering and active learning in achieving such 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. The visualization of the feature space is also valuable for uncovering combinatorial rules (Fig. S6). For example, catalysts containing both high-performing Mo and low-performing Pd are found within the cluster of Pd-based catalysts,

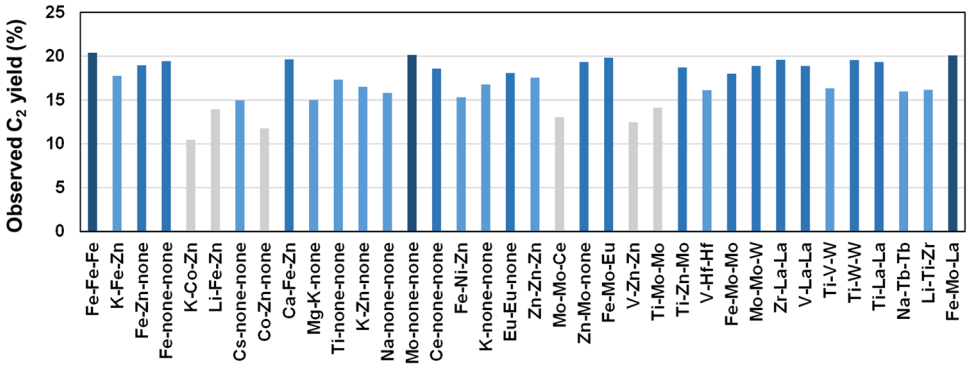

Fig. 4 Discovery of high-performing OCM catalysts using the developed ML model. Herein, 36 catalysts were selected from a subset of catalysts with predicted C_2 yields ≥ 15% using FPS. The bars represent experimentally obtained C_2 yields, with colors indicating the yield levels.

suggesting that Pd has a more dominant influence than Mo in OCM. Strongly interacting combinations, such as those of Cs with Ti, Zr, and Mo, that are frequently observed in high-performing catalysts, are distributed in small clusters separated from the main cluster for Cs-based catalysts. Additionally, Fe-Zn, while not prominently featured in the training data, is isolated in a very narrow region with relatively high predicted C_2 yields, an aspect to be explored further.

Validation, limitations, and future prospects. The primary advantage of AFE, particularly when combined with active learning, lies in its high predictive accuracy and applicability across a wide range of catalysts. To showcase this, we applied FPS to a subset of catalysts with predicted C_2 yields ≥ 15% using the model obtained after active learning; this resulted in the recommendation of 36 catalysts. Subsequent experimental evaluation revealed that 30 out of the 36 catalysts actually exhibited C_2 yields ≥ 15%, with 16 of them surpassing a yield of 18% (Fig. 4, Table S6). This is compared to only 37 cases exceeding a yield of 18% among 175 catalysts in the training data. These catalysts predominantly comprise elements whose oxides possess high basicity, such as alkaline, alkaline earth, and rare earth metal elements, along with early transition metal elements from groups 4 to 6. By contrast, many of the high-performing catalysts identified in Fig. 4 do not conform to this pattern, with a notable presence of elements like Fe and Zn. These elements are largely underexplored in the history of OCM research[42]. A unique advantage of our methodology lies in utilizing the integration of AFE and HTE to systematize the model's education, rather than solely focusing on catalyst discoveries. As a result, the model, enhanced through active learning, significantly streamlined the discovery of high-performing catalysts.

The preceding discussions have elucidated the usefulness of the model involving the engineered features in understanding catalyst design rules and identifying various high-performing catalysts. Conversely, directly extracting physical insights from the engineered features themselves is currently not practical. The engineered features, either individually or in combination, exhibit statistical correlations with catalytic performance. However, statistical correlations do not guarantee causality in catalysis. Moreover, the physical properties of single elements used to generate catalyst features are logically too distant from causal relationships. For instance, the model obtained after active learning is presented as a combination of features: $22.0\,(\text{first\_ion\_en\_max})^3 + 3.32\,\ln(\text{gs\_mag\_moment\_min})^{-1} - 8.63\,(\text{Polarizability\_min})^{-0.5} + 4.59\,(\text{dipole\_polarizability\_min})^{-0.5} - 4.22\,\text{lattice\_constant\_min} - 6.44\,\exp(\text{electron\_affinity\_pro})^{-1} + 10.0\,(\text{gs\_mag\_moment\_std})^2 + 3.26\,\text{hhi\_r\_max} + 27.8$, among which Polarizability_min, dipole_polarizability_min, and first_ion_en_max are identified to be particularly impactful. These features serve to discriminate between elements whose oxides exhibit strong basicity, those that are useful for O_2 activation, and other elements, particularly late transition metal elements that catalyze unselective combustion. However, such interpretations are not insights gained directly from the features themselves but rather post hoc explanations assigned to their roles with reference to existing knowledge. Indeed, attempts to extract physical insights based on elemental features have been hardly successful in literature[24,26]. To extract physical insights from the engineered features without relying on prior knowledge, a diverse and comprehensive collection of catalytically relevant properties of elements, so-called a catalysis feature library, is essential (e.g., formation energies of oxides, redox properties, acidity/basicity, and interaction with various molecules). Such a library, albeit currently unavailable, would leverage the advantage of AFE's compatibility with simple and interpretable ML models. This catalysis feature library, in addition to transparent ML models[43], is another indispensable piece for achieving fully interpretable catalyst informatics, where density functional theory calculations are expected to play a significant role[44].

Conclusion

In summary, we developed and demonstrated AFE as a versatile technique, facilitating effective ML for small datasets of solid catalysts characterized by diverse compositions. AFE exceled in designing highly expressive features tailored to a specific 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. The integration of AFE, FPS, and HTE in an iterative loop through active learning systematized the process to educate the machine, promoting the elimination of alternative hypotheses and the identification of a true hypothesis set that applies to a wide array of catalysts. This success can be attributed to the ability of the machine to develop a feature or knowledge space for recognizing the composition–performance relationships of catalysts. Our systematic approach led the enhanced machine to equip remarkable efficiency in pinpointing various high-performing catalysts. However, the extraction of direct insights from engineered features remains a future challenge, necessitating a comprehensive collection of catalytically relevant properties. The integration of AFE into automated experiments[45] would enable highly efficient autonomous catalyst designs. Furthermore, the knowledge acquired for a specific system is not only beneficial for predicting the performance of unknown compositions within the same system but also for facilitating knowledge acquisition for different systems through transfer learning. As the machine accumulates knowledge across diverse catalytic systems, it is poised to develop comprehensive catalytic knowledge. This advancement promises a future in catalyst development that transcends reliance on researchers' experiences and knowledge.

Methods

Automatic feature engineering. Feature engineering is an essential part of catalyst informatics, as constructing predictive ML models necessitates features that capture the essence of catalysts. Although deep learning can automate feature engineering, the accompanying training requires big data and is often not suitable in the catalysis field where small data are prevalent. Consequently, current feature engineering heavily relies on researchers' intuition, but this empirical approach is insufficient for exploring diverse designs of catalysts. To address this challenge, we developed an AFE technique capable of handling small data on a variety of materials, including catalysts, without prior knowledge. This technique involves assigning features, synthesizing higher-order features, and selecting important features in the context of supervised ML (Fig. 1). Each step is detailed below, using multi-element solid catalysts as a representative example.

Feature assignment. A feature library is created by collecting all possible properties of elements from public databases. It can be appropriately normalized and shifted to prevent the divergence of first-order features. Commutative operations are applied to this feature library to assign primary features (denoted as $\mathbf{X_0}$) that consider the notational order invariance and elemental composition of individual catalysts[20]. We adopted 58 features of elements stored in XenonPy[33] and applied eight types of commutative operations (maximum, minimum, weighted sum, weighted average, weighted sum of squared distance, weighted average squared distance, weighted product, and weighted geometric mean), resulting in 464 primary features.

Feature synthesis. Expressive ML models generally require larger training datasets. Simpler models are suitable for small data, but the reduced expressiveness must be compensated through feature engineering. Therefore, first-order features ($f(\mathbf{X_0})$) that consider nonlinearity and second- or higher-order features ($f(\mathbf{X_0})\cdot g(\mathbf{X_0})$, etc.) that combines two or more first-order features are synthesized[21–23]. We adopted 12 types of functions (x, $x^{1/2}$, x^2, x^3, $\exp(x)$, $\ln(x)$, and their reciprocals), resulting in 5568 first-order features.

Feature selection. Identifying a feature subset is crucial for constructing predictive models, as it is not feasible to use all synthesized higher-order features for model fitting. Despite the availability of several feature selection techniques, an exhaustive approach is typically recommended. We employed a genetic algorithm mainly to minimize the MAE value in LOOCV with a specified number of selected features. Huber regression[34] was adopted owing to its superior performance in handling experimental noise and singular catalysts compared to that of its non-robust counterpart.

AFE was implemented using Python 3.8 and common libraries such as Pandas, NumPy, and scikit-learn, executed in parallel on a PC cluster. The significance of each step is outlined in Table S7, wherein AFE was applied to the OCM dataset, with certain steps intentionally omitted. The analysis revealed that both feature assignment and feature selection were critical for producing a meaningful model, emphasizing the importance of selecting appropriate physicochemical descriptions of catalysts. Higher-order features resulted in a systematic improvement in the score by providing more direct features to the target variable. For small datasets like the OCM dataset, controlling the overfitting in complex models such as random forest regression was difficult. A genetic algorithm (an exhaustive approach) yielded better feature sets than sequential feature selection (a greedy approach)[46].

Dataset. We used three HTE datasets for different heterogeneous catalytic systems to demonstrate AFE (Tables S2–S4). These datasets were obtained using a single protocol, rendering them process-consistent, a crucial feature for reliable ML[47]. A brief overview of the datasets is provided below, with additional details available in published papers[27–32].

Dataset for oxidative coupling of methane. The C_2 yields of 95 M1–M2–M3/BaO catalysts during OCM were collected[27–30]. M1–M3 were selected from Li, Na, Mg, K, Ca, Ti, V, Mn, Fe, Co, Ni, Cu, Zn, Sr, Y, Zr, Mo, Pd, Cs, Ba, La, Ce, Nd, Eu, Tb, Hf, W, and none (blank), with repetitive selection allowed. The amount of each element was fixed at 0.371 mmol per gram support. Although most catalysts were obtained through random selection of elements, certain catalysts were recommended by different ML methods. The experimental protocol used to obtain this dataset is identical to that of the high-throughput experiment described later in this section.

Dataset for conversion of ethanol to butadiene. The butadiene (C_4H_6) yields in ethanol conversion were collected for 177 catalysts[31]. The catalysts were prepared by co-supporting up to 14 elements (Mg, Zn, Cu, Ag, Ni, Al, La, Y, Hf, Zr, Cr, Ga, Nb, and Mo) on SBA-15 through wet impregnation. The loadings of individual elements were optimized within a total loading of 3.00 mmol per gram support to maximize the C_4H_6 yield using a genetic algorithm. The C_4H_6 yield was measured using a catalyst bed packed in a fused quartz reactor (bed height: 2.0 cm; inner diameter: 4 mm on the influent side and 2 mm on the effluent side) at 400 °C and 21.8 mL min^{-1} of 8.4% ethanol diluted in Ar.

Dataset for three-way catalysis. The light-off temperatures of 51 nanoparticle-supported catalysts for NO reduction in three-way catalysis were collected[32]. The light-off temperature is defined as the temperature at 50% NO conversion. Bimetallic to pentametallic nanoparticles with equimolar compositions and containing at least one Pt-group element were prepared using a hot-injection method and deposited onto a γ-Al_2O_3 support at 0.3 wt%. Temperature ramping experiments were performed using a catalyst bed packed in a fused quartz reactor (bed weight: 60 mg; inner diameter: 4 mm on the influent side and 2 mm on the effluent side) with a 10 mL min^{-1} gas flow of a stoichiometric mixture of CO (13000 ppm), C_3H_6 (2000 ppm), NO (3000 ppm), CO_2 (100000 ppm), O_2 (14000 ppm), and He (balance).

High-throughput experiment. To demonstrate active learning, selected catalysts were actually prepared and evaluated using the same experimental method that was used to obtain the training data[27–30]. The catalysts were sampled from a pool of 4060 candidates, generally expressed as M1–M2–M3/BaO. M1–M3 were chosen from Li, Na, Mg, K, Ca, Ti, V, Mn, Fe, Co, Ni, Cu, Zn, Sr, Y, Zr, Mo, Pd, Cs, Ba, La, Ce, Nd, Eu, Tb, Hf, W, or none, with repetitive selection allowed. They were prepared using a parallelized impregnation method using $LiNO_3$, $NaNO_3$, $Mg(NO_3)_2$, KNO_3, $Ca(NO_3)_2{\cdot}4H_2O$, $Ti(OiPr)_4$, $VOSO_4{\cdot}xH_2O$ ($x = 4$), $Mn(NO_3)_2{\cdot}6H_2O$, $Fe(NO_3)_3{\cdot}9H_2O$, $Co(NO_3)_2{\cdot}6H_2O$, $Ni(NO_3)_2{\cdot}6H_2O$, $Cu(NO_3)_2{\cdot}3H_2O$, $Zn(NO_3)_2{\cdot}6H_2O$, $Sr(NO_3)_2$, $Y(NO_3)_3{\cdot}6H_2O$, $ZrO(NO_3)_2{\cdot}2H_2O$, $(NH_4)_6Mo_7O_{24}{\cdot}4H_2O$, $Pd(OAc)_2$, $CsNO_3$, $Ba(NO_3)_2$, $La(NO_3)_3{\cdot}6H_2O$, $Ce(NO_3)_3{\cdot}6H_2O$, $Nd(NO_3)_3{\cdot}6H_2O$, $Eu(OAc)_3{\cdot}4H_2O$, $Tb(NO_3)_3{\cdot}5H_2O$, $Hf(OEt)_4$, and $(NH_4)_{10}H_2(W_2O_7)_6$ as precursors. These precursors were obtained from Sigma-Aldrich, Kanto Chemical, Wako Pure Chemical Industries, and Alfa Aesar. $Ba(OH)_2{\cdot}8H_2O$ purchased from Wako Pure Chemical Industries was used as the precursor for the BaO support. The support powder (1.0 g) was suspended

in 4–5 mL of a precursor solution under stirring at 50 °C for 6 h. The concentration of the solution was adjusted to 0.371 mmol per gram support for each of the selected elements. After drying, the catalyst was calcined in air at 1000 °C for 3 h and thoroughly ground using a mortar and pestle before use. When using metal alkoxides, impregnation was performed in two steps, starting with an aqueous solution and followed by an ethanol solution of the metal alkoxides.

The performance of the catalysts in OCM was evaluated using an in-house high-throughput screening instrument[47]. The instrument comprises a gas mixer for generating the reaction gas mixture (MU-3504, HORIBA STEC), a gas distributor for splitting the reaction gas equally into 20 reactor tubes (fused quartz tubes with an inner diameter of 4 mm on the influent side and 2 mm on the effluent side) loaded with catalyst powder and symmetrically placed in a hollow electric furnace, and an auto-sampler for supplying the effluent gas from individual tubes to a quadruple mass spectrometer (Transpector CPM 3, INFICON). Mass signals were converted into the relative pressures of individual gases based on external calibration. Cooperative action among the programmed gas generation, temperature, and auto-sampling enabled an automatic evaluation of the performance of 20 catalysts under a predetermined set of reaction conditions.

The catalyst powder was packed at a height of 10 mm in the neck of the reactor tube using quartz wool and was in-line calcined at 1000 °C under an O_2 atmosphere for 3 h. A reaction gas mixture of CH_4 and O_2 balanced with Ar was flowed through the 20 tubes, and the temperature was decreased stepwise from 900 to 700 °C in 50 °C increments. The total gas flow volume (10, 15, and 20 mL min^{-1}), CH_4/O_2 ratio (2, 4, and 6 mol mol^{-1}), and Ar concentration (P_{Ar} = 0.15, 0.40, and 0.70 atm) were respectively varied at each temperature, resulting in a total of 135 reaction conditions. The C_2 yield, defined as the percentage of the doubled sum of the partial pressures of C_2H_6 and C_2H_4 relative to that of CH_4 in the influent, was obtained at each of the 135 conditions, and the maximum C_2 yield was recorded for further analysis.

Data availability

The three datasets used to demonstrate automatic feature engineering in Fig. 1 are curated from published papers and listed in the Supplementary Information. The authors declare that all data supporting the findings and those used for reproducing the figures in this paper are available within the paper and its Supplementary Information. Source data are provided with this paper.

Code availability

Codes are available at https://github.com/TaniikeLaboratory/Automatic-feature-engineering-for-catalyst-small-data.

Received: 8 August 2023; Accepted: 8 December 2023;
Published online: 12 January 2024

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.* **3**, 54 (2017).
2. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
3. Toyao, T. et al. Machine learning for catalysis informatics: recent applications and prospects. *ACS Catal.* **10**, 2260–2297 (2020).
4. Takahashi, K. et al. Catalysts informatics: paradigm shift towards data-driven catalyst design. *Chem. Commun.* **59**, 2222–2238 (2023).
5. Beker, W. et al. Machine learning may sometimes simply capture literature popularity trends: a case study of heterocyclic suzuki–miyaura coupling. *J. Am. Chem. Soc.* **144**, 4819–4827 (2022).
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.* **5**, 83 (2019).
7. Strieth-Kalthoff, F. et al. Machine learning for chemical reactivity: the importance of failed experiments. *Angew. Chem. Int. Ed.* **61**, e202204647 (2022).
8. Taniike, T. & Takahashi, K. The value of negative results in data-driven catalysis research. *Nat. Catal.* **6**, 108–111 (2023).
9. Ryan, K., Lengyel, J. & Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
10. Coley, C. W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
11. Reiser, P. et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **3**, 93 (2022).
12. Hammer, B. & Nørskov, J. K. Theoretical surface science and catalysis—calculations and concepts. *Adv. Catal.* **45**, 71–129 (2000).
13. Clavier, H. & Nolan, S. P. Percent buried volume for phosphine and N-eterocyclic carbeneligands: steric properties in organometallic chemistry. *Chem. Commun.* **46**, 841–861 (2010).
14. Ringe, S. The importance of a charge transfer descriptor for screening potential CO_2 reduction electrocatalysts. *Nat. Commun.* **14**, 2598 (2023).
15. Santiago, C. B., Guo, J. Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **9**, 2398–2412 (2018).
16. Liu, J. et al. Toward excellence of electrocatalyst design by emerging descriptor-oriented machine learning. *Adv. Funct. Mater.* **32**, 2110748 (2022).
17. Zhang, Y. et al. Descriptor-free design of multicomponent catalysts. *ACS Catal.* **12**, 10562–10571 (2022).
18. Urakawa, A. & Baiker, A. Space-resolved profiling relevant in heterogeneous catalysis. *Top. Catal.* **52**, 1312–1322 (2009).
19. Wada, T. et al. Structure-performance relationship of $Mg(OEt)_2$-based Ziegler-Natta catalysts. *J. Catal.* **389**, 525–532 (2020).
20. Liu, C. et al. Machine learning to predict quasicrystals from chemical compositions. *Adv. Mater.* **33**, 2102507 (2021).
21. Ghiringhelli, L. M. et al. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
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.* **28**, 1304–1311 (2016).
23. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
24. Suzuki, K. et al. Statistical analysis and discovery of heterogeneous catalysts based on machine learning from diverse published data. *ChemCatChem.* **11**, 4537–4547 (2019).
25. Williams, T., McCullough, K. & Lauterbach, J. A. Enabling catalyst discovery through machine learning and high-throughput experimentation. *Chem. Mater.* **32**, 157–165 (2020).
26. Ishioka, S. et al. Designing catalyst descriptors for machine learning in oxidative coupling of methane. *ACS Catal.* **12**, 11541–11546 (2022).
27. Nguyen, T. N. et al. Learning catalyst design based on bias-free data set for oxidative coupling of methane. *ACS Catal.* **11**, 1797–1809 (2021).
28. Nakanowatari, S. et al. Extraction of catalyst design heuristics from random catalyst dataset and their utilization in catalyst development for oxidative coupling of methane. *ChemCatChem.* **13**, 3262–3269 (2021).
29. Takahashi, L. et al. Constructing catalyst knowledge networks from catalyst big data in oxidative coupling of methane for designing catalysts. *Chem. Sci.* **12**, 12546–12555 (2021).
30. Takahashi, K. et al. Catalysis gene expression profiling: sequencing and designing catalysts. *J. Phys. Chem. Lett.* **12**, 7335–7341 (2021).
31. Jayakumar, T. P. et al. Exploration of ethanol-to-butadiene catalysts by high-throughput experimentation and machine learning. *Appl. Catal. A Gen.* **666**, 119427 (2023).
32. Son, S. D. et al. High-throughput screening of multimetallic catalysts for three-way catalysis. *Sci. Technol. Adv. Mater. Methods* https://doi.org/10.1080/27660400.2023.2284130 (2023).
33. Yoshida, R. XenonPy is a Python software for materials informatics. https://github.com/yoshida-lab/XenonPy (2018).
34. Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
35. Maaten, L. V. D. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
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.* **99**, 4566–4568 (1995).
37. Ji, S. et al. Surface WO_4 tetrahedron: the essence of the oxidative coupling of methane over M–W–Mn/SiO_2 catalysts. *J. Catal.* **220**, 47–56 (2003).

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.* **107**, 5062–5068 (1985).
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.* **35**, 215–231 (1995).
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.* **96**, 104254 (2021).
41. Choudhary, T. V., Banerjee, S. & Choudhary, V. R. Catalysts for combustion of methane and lower alkanes. *Appl. Catal. A Gen.* **234**, 1–23 (2002).
42. Mine, S. et al. Analysis of updated literature data up to 2019 on the oxidative coupling of methane using an extrapolative machine-learning method to identify novel catalysts. *ChemCatChem.* **13**, 3636–3655 (2021).
43. Esterhuizen, J. A., Goldsmith, B. R. & Linic, S. Interpretable machine learning for knowledge generation in heterogeneous catalysis. *Nat. Catal.* **5**, 175–184 (2022).
44. Mamun, O., Winther, K. T., Boes, J. R. & Bligaard, T. High-throughput calculations of catalytic properties of bimetallic alloy surfaces. *Sci. Data* **6**, 76 (2019).
45. Trunschke, A. Prospects and challenges for autonomous catalyst discovery viewed from an experimental perspective. *Catal. Sci. Technol.* **12**, 3650–3669 (2022).
46. Ferri, F. J., Pudil, P., Hatef, M. & Kittler, J. Comparative study of techniques for large-scale feature selection. In: *Pattern Recognition in Practice Iv: Multiple Paradigms, Comparative Studies, and Hybrid Systems: Proceedings of an International Workshop held on Vlieland, The Netherlands, 1–3 June 1994* (eds. Gelsema, E. S. & Kanal, L. S.) 403–416 (Elsevier, 1994).
47. Nguyen, T. N. et al. High-throughput experimentation and catalyst informatics for oxidative coupling of methane. *ACS Catal.* **10**, 921–932 (2020).

Acknowledgements

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

Author contributions

T.T. designed the study and wrote the paper. T.T. and A.F. performed the research. T.T., S.N., F.G.E., and K.T. analyzed the data. All authors approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-023-01086-y.

Correspondence and requests for materials should be addressed to Toshiaki Taniike.

Peer review information *Communications Chemistry* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.