Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tsta20

Acquiring and transferring comprehensive catalyst knowledge through integrated high-throughput experimentation and automatic feature engineering

Aya Fujiwara, Sunao Nakanowatari, Yohei Cho & Toshiaki Taniike

Publisher: Taylor & Francis & The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group.

Journal: *Science and Technology of Advanced Materials*

DOI: 10.1080/14686996.2025.2454219

Acquiring and transferring comprehensive catalyst knowledge through integrated high-throughput experimentation and automatic feature engineering

Aya Fujiwara[a], Sunao Nakanowatari[a], Yohei Cho[a], Toshiaki Taniike[a*]

[a] *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

*Corresponding author. E-mail: taniike@jaist.ac.jp

Acquiring and transferring comprehensive catalyst knowledge through integrated high-throughput experimentation and automatic feature engineering

ABSTRACT

Solid catalyst development has traditionally relied on trial-and-error approaches, limiting the broader application of valuable insights across different catalyst families. To overcome this fragmentation, we introduce a framework that integrates high-throughput experimentation (HTE) and automatic feature engineering (AFE) with active learning to acquire comprehensive catalyst knowledge. The framework is demonstrated for oxidative coupling of methane (OCM), where active learning is continued until the machine learning model achieves robustness for each of the BaO-, CaO-, La_2O_3-, TiO_2-, and ZrO_2-supported catalysts, with 333 catalysts newly tested. The resulting models are utilized to extract catalyst design rules, revealing key synergistic combinations in high-performing catalysts. Moreover, we propose a method for transferring knowledge between supports, showing that features refined on one support can improve predictions on others. This framework advances the understanding of catalyst design and promotes reliable machine learning.

KEYWORDS: Catalyst informatics; machine learning; high-throughput experimentation; descriptor; oxidative coupling of methane

1. Introduction

The intricacy of composition-function relationships has made trials and errors a major driver in the development of solid catalysts.[1,2] Individual researchers propose their own hypotheses and test various materials as potential catalysts. The research focus then evolves along with improved hypotheses and catalysts found in prior research for further refinement. Such a process in catalyst development typically proceeds in parallel, leading to the discovery of multiple catalyst families differing in their design concepts. This situation is exemplified by oxidative coupling of methane (OCM)—a reaction that converts methane, the main component of natural gas and biogas, into C_2 compounds in a single step.[3] OCM is considered potentially more efficient than the current two-step route involving steam reforming and the Fischer–Tropsch process.[4] However, due to the chemical inertness of CH_4 when compared to the C_2 products, achieving high yields is challenging. Currently, few catalysts are known to stably achieve techno-economical performance (e.g., C_2 yields over 30% otherwise 28% with 80% selectivity) in a conventional fixed-bed flow reactor.[5–7] Nonetheless, catalyst development, continued over 40 years since 1982,[3] has identified several high-performing catalyst families, represented by Li/MgO, Sr/La_2O_3, and $Mn–Na_2WO_4/SiO_2$.[8–10] Briefly, in Li/MgO, the addition of Li increases the number of active sites on the MgO surface.[11–15] However, strategies are needed to prevent Li sublimation at high reaction temperatures. In Sr/La_2O_3, the inclusion of Sr^{2+} in La_2O_3 introduces lattice distortion and/or electronic modulation, which enhances the C_2 selectivity of La_2O_3, likely due to the formation of superoxide species ($O_2^–$).[16,17] The development has been continued mainly to strengthen the advantage of low-temperature activity. $Mn–Na_2WO_4/SiO_2$ is known as one of the most promising catalysts for OCM due to its high C_2 selectivity and durability, which result from the synergistic interactions among the components. Na^+ stabilizes tetrahedral $WO_4^{2–}$

species, active in OCM, by interacting with it and promoting the formation of the cristobalite phase of SiO_2, which also stabilizes these species. Mn^{2+} aids OCM by facilitating the recovery of W^{6+}.[18,19] Efforts have been directed at optimizing the conversion-selectivity tradeoff through various methods, including catalyst preparation techniques, the addition of extra elements, and so on. Likewise, these catalyst families stand on very different design guidelines, and they have been developed nearly independently without explicit exchange of design guidelines across the families.[20] It is believed that more comprehensive understanding or more rational design of catalysts would be achieved if one can clarify and apply the commonalities and differences among different catalyst families.

Data-driven catalysis research, also known as catalyst informatics, aims to accelerate the development and understanding of catalysts by discovering applicable trends and patterns hidden in catalyst data with the aid of data science techniques such as machine learning (ML) and visualization. The bottlenecks in catalyst informatics are the scarce availability of catalyst data suitable for data science and the difficulty of hand-crafting descriptors that capture the essence of intricate composition-function relationships.[21] Our research group has employed high-throughput experimentation (HTE) in the preparation and evaluation of a large number of solid catalysts to generate sized, qualified, and consistent datasets for various heterogeneous catalysis, including OCM.[22–26] Moreover, we recently introduced an automatic feature engineering (AFE) technique, which programmatically designs descriptors that can capture the essence of target catalysis, starting from general physical properties of elements such as atomic radii and electronegativity.[27] AFE generates a predictive ML model with tailored descriptors without requiring researchers to make any assumptions or hypotheses about the target system. Using BaO-supported catalysts for OCM as an example, we demonstrated that the recursive refinement of the ML model and descriptors—referred to as the design hypothesis—through active learning with AFE ultimately resulted in a robust design hypothesis. This refined hypothesis applies to a broad range of catalysts and is powerful in pinpointing various high-performing catalysts as well as in comprehending the underlying design guidelines.[27]

In this study, the same active learning approach integrated with HTE and AFE was applied to five catalyst families for OCM. We started with OCM catalyst data previously acquired by us for five supports (BaO, CaO, La_2O_3, TiO_2, and ZrO_2), and conducted large-scale active learning by adding a total of 333 catalysts to establish a robust design hypothesis for each support. We then analyzed these design hypotheses to clarify the commonalities and differences among the five catalyst families. Importantly, we found that a design hypothesis established for one catalyst family could assist in building an ML model for another family based on their commonalities. This illustrates how knowledge gained from one catalyst family can be transferred to facilitate the design of another family.

2. Method

In this study, our goal is to establish a design hypothesis for each of the five catalyst families corresponding to different support materials, using previously acquired data. However, similar to how researchers cannot exclude alternative hypotheses when evidence is limited, AFE cannot disregard alternative design hypotheses—ML models with differently tailored descriptors that fit the training data similarly—when the diversity of catalysts in the training data is restricted. Therefore, we employed an active learning strategy as illustrated in Figure 1. This strategy uses farthest point sampling (FPS) within the descriptor space defined by AFE to propose catalysts that are maximally dissimilar to those included in the training data.[28] These catalysts serve as rigorous control experiments for validating the proposed design hypothesis. The performance of the proposed catalysts is then assessed using HTE to reinforce the training data and update the design hypothesis via AFE. This iterative process aims to eliminate design hypotheses that do not generalize well across catalysts, leading to a robust and experimentally validated design hypothesis. Further details are provided below.

2.1. Dataset

Our group has accumulated OCM data for quaternary catalysts represented as M1–M2–M3/Support using a consistent experimental protocol through HTE.[22,27,29,30] M1–3 represent the supported elements, which can be selected, with duplication allowed, 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 (none indicates no addition of elements). The support is selected from MgO, Al_2O_3, SiO_2, CaO, TiO_2, ZrO_2, BaO, La_2O_3, and CeO_2. M1–3 are chosen from 28 elements (including “none”), creating 4,060 candidate catalysts. The “none–none–none” combination is excluded as no features can be assigned within the AFE algorithm. This constitutes a parameter space containing 4,059 catalysts per support and a total of 36,531 catalysts. The loading amount of supported elements, except for none, is 0.37 mmol per gram of support per selection.

In this study, out of 636 quaternary catalysts that we have reported,[30–32] we extracted 381 catalysts relating to CaO, BaO, La_2O_3, TiO_2, and ZrO_2 supports, for establishing design hypotheses through active learning. CaO and La_2O_3 were selected as they are the most studied basic oxides in the history of OCM, and likely possessing distinct catalyst designs.[33] BaO was selected as a reference to these supports but with higher C_2 yields and selectivity. Dissimilar to these, the redox-active support TiO_2 and the non-redox-active support ZrO_2, both from Group 4, were incorporated to investigate potential correlations between their general physical properties and the proposed design hypotheses. Among the 381 catalysts, 175 were obtained via random sampling from the entire space, while the remaining 206 catalysts were obtained to validate various ML techniques.[30–32]

2.2. Automatic feature engineering

AFE is a technique that automates the design of physically meaningful features for a given catalyst dataset within the framework of supervised ML. It involves a structured pipeline of feature assignment, synthesis, and selection. First, it assigns physical quantities of elements to catalysts with their elemental compositions reflected through commutative operations. Then, from these assigned primitive features, it synthesizes higher-order features that involve non-linear and combinatorial effects through mathematical operations. Eventually, from the synthesized large array of features, a specified number of features, i.e., descriptors, that optimize the score of supervised ML, are selected.

In this study, we utilized 58 parameters of elements stored in XenonPy with normalization,[34] following previous literature. These parameters were assigned to each catalyst using five types of commutative operations (maximum, minimum, average, product, standard deviation), resulting in 290 primary features. These primary features were further synthesized into 3,480 features using twelve function forms (x, sqrt(x), x^2, x^3, ln(x), exp(x), and their reciprocals, where x denotes each primary feature). We employed a genetic algorithm-based approach to select eight features that minimize the mean absolute error (MAE) in leave-one-out cross-validation (LOOCV) with Huber regression.[35] This process involved evaluating approximately 4,000,000 models per dataset with different feature combinations and selecting the combination of features (X) and model (f(X)) with the lowest cross-validation (CV) score as the most plausible design hypothesis. The adoption of Huber regression, a type of multiple linear regression, aimed not only to prevent overfitting due to its reduced number of parameters but also to ensure robustness against outliers (such as experimental failures). The number of selected features was empirically determined to balance the CV score and the cost of feature selection. Further details on the AFE itself and parameter selection are referred to in our previous literature.[27]

2.3. Farthest point sampling

To validate and refine the design hypothesis obtained in Section 2.2, we added 10 or 20 catalysts in each cycle of active learning for experimental testing. Among these, 90% were selected using FPS within the normalized eight-dimensional feature space determined by AFE. This approach recommends evaluating the most extrapolative data to the current design hypothesis. Further details on FPS, including its methodology and implementation, are provided in Figure S1 and the accompanying description. The remaining 10% corresponded to the re-evaluation of catalysts that showed the largest deviations between observation and prediction in the last cycle. This step helped to eliminate experimental errors.

2.4. High-throughput experiment

The preparation and evaluation of the catalysts proposed in 2.3 were conducted under the same experimental methods and conditions as those used when acquiring the original training data, which are briefly described as follows.

Materials

Metal precursors: $LiNO_3$, $NaNO_3$, $Mg(NO_3)_2$, KNO_3, $Ca(NO_3)_2 \cdot 4H_2O$, $Ti(OiPr)_4$, $VOSO_4 \cdot xH_2O$ (x = 3–5), $Mn(NO_3)_2 \cdot 6H_2O$, $Fe(NO_3)_3 \cdot 9H_2O$, $Co(NO_3)_2 \cdot 6H_2O$, $Ni(NO_3)_2 \cdot 6H_2O$, $Cu(NO_3)_2 \cdot 3H_2O$, $Zn(NO_3)_2 \cdot 6H_2O$, $Sr(NO_3)_2$, $Y(NO_3)_3 \cdot 6H_2O$, $ZrO(NO_3)_2 \cdot xH_2O$ (x = 2), $(NH_4)_6Mo_7O_{24} \cdot 4H_2O$, $Pd(OAc)_2$, $CsNO_3$, $Ba(NO_3)_2$, $La(NO_3)_3 \cdot 6H_2O$, $Ce(NO_3)_3 \cdot 6H_2O$, $Nd(NO_3)_3 \cdot 6H_2O$, $Eu(NO_3)_3 \cdot 5H_2O$, $Tb(NO_3)_3 \cdot 5H_2O$, $Hf(OEt)_4$, and $(NH_4)_{10}H_2(W_2O_7)_6$. These were purchased from one of the following suppliers: Sigma-Aldrich, Kanto Chemical, Wako Pure Chemical Industries, Alfa-Aesar, or Sumitomo Chemical.

Oxide supports and their precursors: $Ca(OH)_2$ (3.0 m^2/g, Wako Pure Chemical Industries), $Ba(OH)_2 \cdot 8H_2O$ (1.1 m^2/g, Wako Pure Chemical Industries), La_2O_3 (8.3 m^2/g, Wako Pure Chemical Industries), TiO_2 (17.4 m^2/g, anatase type, Kanto Chemical), and ZrO_2 (3.2 m^2/g, Kanto Chemical).

Catalysts preparation

The catalyst preparation was conducted using a parallelized wet impregnation method, where support powder was impregnated with an aqueous solution of specified metal precursors at 50 °C for 6 hours. The loading amount of elements was fixed at 0.37 mmol/g-support per selection within M1–3. After impregnation, the powder was vacuum-dried and calcined in air at 1000 °C for 3 hours. When the precursors involved metal alkoxides, the impregnation was performed in two steps: First with an aqueous solution of the other precursors, followed by impregnation with an ethanol solution of metal alkoxides.

Catalysts evaluation

The catalysts were evaluated for OCM using a custom-built HTE system.[22] The system functions through a combination of a gas mixer, flow distributor, electric furnace bearing reactors, autosampler, and quadrupole mass spectrometer (QMS), enabling the automated evaluation of the performance of 20 catalysts under a programmed sequence of reaction conditions. The catalysts were fixed into 1 cm-high beds in reaction tubes (made of fused quartz tubes with 4 mm and 2 mm inner diameters), with the aid of quartz wool. After inline activation at 1000 °C for 3 hours in an oxygen stream, the 20 catalysts were tested under 135 conditions varying in temperatures (700, 750, 800, 850, and 900 °C), total gas flow rates (10, 15, 20 mL/min/channel), CH_4/O_2 ratios (2, 4, 6 mol/mol), and Ar partial pressures as a balance gas (0.15, 0.40, 0.70 atm). Each catalyst was labeled based on its highest C_2 yield among the 135 conditions.

3. Results and discussion

3.1. Active learning for obtaining robust design hypotheses

In conventional hypothesis validation, researchers add data not only to confirm the validity of the main hypothesis but also to eliminate alternative hypotheses that could similarly explain the original data. Likewise, when training data is limited, AFE may provide multiple design hypotheses with similar scores, regarded as alternative design hypotheses. These arise due to multicollinearity among features within the given data,[36] and eliminating incorrect or less robust hypotheses requires active learning. In this study, catalysts were selected by FPS to expand the diversity of the data, thereby testing the robustness of the proposed design hypotheses. By repeating active learning cycles, design hypotheses that do not apply to diverse data are filtered out, resulting in more robust and reliable design hypotheses. According to our previous report, the active learning cycle was conducted five times, and in some cases, six times.[27]

Taking CaO-based catalysts as an example, the development of design hypotheses along with the active learning cycles is demonstrated in Table 1. Notably, across all cycles, the MAE_{CV} values were similar to the MAE_{train} values, indicating an absence of overfitting. Both the MAE_{train} and MAE_{CV} values remained within a narrow range of 1.6–1.9%, comparable to the experimental error (1.5–2% in C_2 yields). This suggests that AFE found design hypotheses that similarly fit to the training data, even though its diversity increased along with the cycles. On the other hand, the elimination of less robust design hypotheses that did not apply to the newly added data led to variation in the selected features. In Table 1, the selected features are ordered by their permutation feature importance. Three features—minimum covalent radius Pyykko double, minimum atomic radius Rahm, and minimum covalent radius Cordero—consistently appeared in most cycles with high importance, despite slight differences in functional forms. These features are thus considered as irreplaceable descriptors for the performance of CaO-based catalysts. Other less important features varied from cycle to cycle, corresponding to the elimination of less robust design hypotheses.

Less robust design hypotheses are eliminated through active learning, but this does not necessarily lead to convergence onto a single design hypothesis. Some feature combinations may be thoroughly collinear across the entire space (i.e., the 4059 catalysts), making them indistinguishable regardless of data addition. Such global collinearity implies that design hypotheses, despite bearing differing features, exhibit similar predictive behavior over the entire space and are therefore considered similarly robust. To investigate this, we compared the predictive behavior of the design hypothesis obtained in one cycle with that obtained in the last cycle. Specifically, we predicted the C_2 yields for the 4059 catalysts using two design hypotheses and calculated the MAE between the two prediction sets. Figure 2 illustrates the development of MAE values across active learning cycles. For all supports, the MAE value tended to decrease significantly in the early stages of active learning. This reflects that less robust design hypotheses proposed by AFE due to limited data were quickly corrected in these early stages. In the later stages, changes in MAE became much smaller, indicating that the predictive behavior of the design hypotheses stabilized with data augmentation. In the final cycle, MAE values for all supports were below 2%, comparable to or lower than the experimental error (further reduction in MAE is considered meaningless due to limitations in experimental precision). Based on this, we concluded that the design hypotheses had become sufficiently robust for use in subsequent analyses. The predicted values derived from these design hypotheses are considered to be sufficiently reliable as measured values, and are used in the subsequent analysis. Note that design hypotheses are based on the correlation between the physical characteristics of catalyst compositions (particularly features derived from the properties of constituent metals) and their performance.

Active learning for the five supports added a total of 333 catalysts to the original dataset, which contained 381 catalysts. Their measured OCM performances are summarized in Figure S1 and digitally available in a database platform (https://cads.eng.hokudai.ac.jp/).[37] Also, all predicted C_2 yields from robust design hypothesis are summarized in csv file as supporting information.

3.2. Extracting catalyst design guidelines

The design hypotheses obtained in 3.1 for the five catalyst families were utilized to extract catalyst design guidelines. Specifically, we analyzed the relationships between the catalyst elemental compositions and their predicted C_2 yields to identify rules such as high-performing designs for each support, as well as similarities and differences across supports.

First, we analyzed the distribution of predicted C_2 yields for the 4059 catalysts across each support (see Figure S3). The distributions varied by support, with medians ranked as BaO > La_2O_3 > CaO > TiO_2 > ZrO_2. This suggests that basic oxides generally achieved better yields, consistent with common practices in OCM. To investigate the overall similarity between the supports irrespective of absolute yield differences, the Pearson's correlation coefficient was derived between the predicted C_2 yields for the 4059 catalysts across different supports (Figure 3). A high positive correlation coefficient between supports indicates that their design hypotheses are generally similar. It can be seen that the correlation coefficients among CaO, BaO, and La_2O_3 were all above +0.8, indicating that not only do these basic supports generally achieve higher yields, but their design hypotheses are also similar. On the other hand, the design hypotheses for the group 4 oxides, TiO_2 and ZrO_2, showed lower similarity. This difference is believed to arise from whether redox activity is present or absent. Interestingly, La_2O_3 held the highest correlation with the other supports, suggesting the generality of the design hypothesis for La_2O_3. The fact that the correlation coefficients are not equal to 1 suggest that globally similar design hypotheses are not identical when examined locally. This local variation highlights the impact of individual chemical properties reflected on the resulting design rules for each support.

Next, to analyze design rules for high-performing catalysts, we created a subset of catalysts with predicted C_2 yields in the top 10% for each support and analyzed the frequency of occurrences of individual elements within this subset. The results are visualized in Figure 4a in a periodic table format, where elements with higher frequencies are considered more effectively paired with the corresponding supports. Additionally, focusing on the three most frequent elements for each support, we analyzed the frequency of secondary elements appearing in association with these main elements. Figure 4b presents the results in pie charts. Insights obtained from Figures 4a and 4b for each support are summarized below.

- CaO: Elements most frequently appearing in high-performing catalysts are Sr > Mg > Ca > Cs. Alkaline earth metal elements are predominant, followed by rare earth elements and group 4 elements. Notably, Cs is the only alkali metal element that appeared frequently. The effectiveness of pairing alkaline earth metal elements, especially Sr, with CaO is well-known in the literature, suggesting the validity of the established design hypothesis.[38] The secondary elements frequently associated with the top three elements—Sr, Mg, and Ca—are shown in Figure 4b. The patterns of secondary elements are reasonably similar among these elements, predominantly including alkali metal, alkaline earth metal, rare earth, and group 4 elements excluding Ti. Particularly, Cs is the most frequently appearing secondary element, although alternative alkaline earth metal elements were expected to be more frequent from Figure 4a. This suggests a synergistic combination between alkaline earth metal elements and Cs on CaO, which can also be confirmed in Figure S4.
- BaO: The trends of elements frequently appearing on BaO are quite different from those on CaO, the same alkaline earth metal oxide (Figure 4a). There is no overall preference for alkaline earth and rare earth elements; rather, specific elements are frequently found on BaO. In particular, La has a significantly higher frequency of appearances, followed by Mo, Cs,

Zn, and Sr in this order. From Figure 4b, the elements associated with La are quite diverse, suggesting that the La–BaO combination alone can achieve high performance. For Mo, Zn and W in addition to La are frequently paired. It was reported that Zn–La combination is effective in enhancing catalyst stability and preventing carbon deposition in OCM.[39] The association of Mo and W suggests that increasing the amount of group 6 elements can improve performance, as evidenced by the observed high yield for W–W–W/BaO. Cs is frequently associated with the main elements, La, Mo, and Sr. The specificity of Cs among the alkali metal elements was also observed on CaO.

- La_2O_3: The elements frequently appearing on this basic oxide are similar to those on CaO, with a prevalence of alkaline earth metal and rare earth elements excluding Ce, as well as group 4 elements excluding Ti. It was reported that doping La_2O_3 with elements having an oxidation state equal to or lower than that of La (3+) enhances the activity,[16] which aligns with the frequent occurrence of alkaline earth metal (2+) and rare earth elements (3+) excluding Ce, which usually bears a 4+ oxidation state. Notably, Ba appears overwhelmingly frequently, similar to the La–BaO combination for BaO. The diversity of elements associated with Ba further supports strong synergy between La and Ba. Our previous work demonstrated that using BaO and La_2O_3 as a mixed support results in a complementary effect, where BaO's high C_2 selectivity at high temperatures and La_2O_3's high activity at low temperatures together increase C_2 yields.[29]
- TiO_2: Alkaline metal elements, particularly Cs, are significantly prevalent (Cs >> K > Li). Although the Cs–TiO_2 combination has not been reported for OCM, Cs is known to enhance the photocatalytic activity of TiO_2 by promoting its reduction.[40,41] This suggests that Cs might also facilitate the redox action of TiO_2 in OCM. Cs is frequently associated with

elements that prefer a 4+ oxidation state, such as Hf and Ce, followed by other rare earth elements. [42,43]

- ZrO_2: Similar to TiO_2, alkaline metal elements are frequently observed on ZrO_2, though the strong emphasis on Cs seen on TiO_2 is largely diminished. Additionally, alkaline earth metal and early transition metal elements are more prevalent overall. From Figure 4b, it is evident that alkaline metal and alkaline earth metal elements often appear in conjunction with early transition metal elements. This suggests that ZrO_2, with minimal intrinsic OCM activity on its own, benefits from elemental combinations that can form oxometalate anions, which are active for OCM.[44,45]

We have also analyzed the design rules for catalysts in the bottom 10% based on predicted C_2 yields (Figure S5). In contrast to high-performing catalysts, the design rules for low-performing catalysts are less dependent on the supports, with late transition metals being predominant. These elements are known to catalyze non-selective combustion, implying their overall effectiveness in methane combustion, regardless of the support.

3.3. Transferability of design hypotheses

Experienced catalyst researchers often apply knowledge and experience gained from other catalytic systems when developing catalysts for new systems. This transfer requires a certain degree of intuition, and replicating this leap through data science is a key aspect of catalyst informatics. Here, to demonstrate this concept, we attempt to transfer design hypotheses across different catalyst families.

The underlying idea is to exploit the similarities of design hypotheses across different catalyst supports. Specifically, a design hypothesis obtained for one support was converted to features to assist in acquiring a design hypothesis for another support through AFE. The most straightforward approach for this is to directly use the predicted C_2 yields for one support as a feature, as suggested by the correspondence of the predicted yields across supports (cf. Figure 3). Additionally, we considered using the descriptors obtained for one support as features through dimensional reduction. This approach is based on an assumption that the similarities observed between elements or elemental combinations on one support could hold for another support, regardless of the C_2 yields. Here, we reduced the original eight features to two features using two different methods. One method utilizes principal component analysis (PCA), which maximizes variance in the higher-dimensional space, effectively representing the distribution of elemental combinations in the original feature space. The other method is based on t-distributed stochastic neighbor embedding (t-SNE), which prioritizes preserving the local relationship between elemental combinations in the original space, effectively capturing the similarity of the combinations.[46] These features derived from design hypotheses on different catalyst families are referred to as "design hypothesis features" to distinguish them from the standard features derived from the XenonPy elemental physical properties.

Figure 5 compares the MAE_{CV} scores of ML models that incorporate at least one design hypothesis feature from other supports with those that use no such features. The comparison covers three catalyst families: CaO, La_2O_3, and ZrO_2. The figure displays the design hypothesis features selected via AFE, along with their ranking in permutation feature importance. For CaO and La_2O_3, incorporating design hypotheses from other supports significantly reduced MAE_{CV} scores, surpassing the fluctuations (error bars in Figure 5) observed with genetic algorithm-based feature selection. In both cases, the predicted C_2 yield on BaO was identified as the most important descriptor, aligning with the high correlation of predicted C_2 yields between these supports shown in Figure 3. Additionally, while less critical, the t-SNE feature from TiO_2 was selected for both supports. For ZrO_2, transferring design hypotheses from other supports did not lower MAE_{CV} scores, consistent with its distinct design hypothesis compared to the other supports, as discussed in section 3.2. Interestingly, only t-SNE features were selected from the dimensionally reduced features, with none of the PCA features being chosen. This suggests that local relationships or similarities between elemental combinations are more effective for transferring knowledge than their distribution in the original feature space. This finding is promising for catalyst researchers, as local similarities are more accessible than distributions in the feature space.

In summary, we demonstrated that within the context of AFE, knowledge can be effectively transferred from one catalyst family to another. This is considered beneficial for catalyst development in two aspects: first, it improves the prediction accuracy of ML models, allowing more precise pinpointing of high-performing catalysts; second, it could accelerate the active learning as providing validated design hypotheses as features is equivalent to providing established knowledge to the ML process.

4. Conclusion

The empirical aspect of catalyst development often leads to a focus or bias toward specific compositions, resulting in fragmented pieces of knowledge that are difficult to integrate. However, in the realm of catalyst informatics, iterative updating of both descriptors and machine learning (ML) models, accompanied by strategic data additions, can yield more comprehensive and applicable catalyst knowledge. In this study, by employing high-throughput experimentation (HTE) and automatic feature engineering (AFE), we implemented active learning to acquire comprehensive catalyst knowledge for each of BaO-, CaO-, La_2O_3-, TiO_2-, and ZrO_2-supported catalysts applied to oxidative coupling of methane (OCM). The acquired knowledge was then used to extract catalyst design rules and evaluate the potential for knowledge transfer across them.

Active learning was performed for each catalyst family, starting with previously acquired training data. A design hypothesis, comprising an ML model and descriptors, was generated by AFE, which produced numerous catalyst features and selected the most plausible ones in the context of supervised ML. Challenging catalysts were identified using farthest point sampling (FPS) and were experimentally tested to enhance the training data and refine the design hypothesis. This process was repeated five to six times, resulting in a robust design hypothesis applicable to all compositions, providing comprehensive catalyst knowledge for each support.

The established design hypotheses were then used to identify catalyst design rules and evaluate their similarities and differences across supports. We identified synergistic combinations frequently found in high-performing catalysts, such as alkaline earth metals for CaO, La for BaO, Ba for La_2O_3, Cs for TiO_2, and alkali metals for ZrO_2, along with some novel findings.

Finally, we introduced a method for transferring catalyst knowledge between different supports using AFE, leveraging the similarities between the design hypotheses. It was demonstrated that the descriptors or the model’s output refined through active learning on one support could be adapted into features that enhanced the predictive accuracy of the ML models on other supports.

In summary, this paper established a data scientific framework for acquiring comprehensive catalyst knowledge, which can aid in designing novel and diverse high-performance catalysts. Additionally, transferring validated design hypotheses across different catalyst systems enhances model reliability and may reduce the need for extensive active learning.

Supplemental material

Scatter plot of CH_4 conversion vs. C_2 selectivity for 726 catalysts, box plot of predicted C_2 yields for each support, frequency of appearances of secondary elements associated with Cs–CaO, element appearance frequency analysis for the bottom 10% in C_2 yield (PDF). Predicted C_2 yields from robust design hypothesis (CSV).

The dataset obtained in the active learning cycle is available on the web platform (https://cads.eng.hokudai.ac.jp/).

Author contributions

A.F. wrote the paper. T.T. designed the study. T.T., S.N., and A.F. analyzed the data. S.N., Y.C., and T.T. reviewed and edited the paper. All authors approved the final paper.

Disclosure statement

The authors declare no competing financial interest.

Acknowledgments

This work is supported by JST-Mirai Program (Grant Number JPMJMI22G4). The work of A.F. is supported by JST SPRING (Grant number JPMJSP2102). The work of Y.C. is supported by JSPS KAKENHI (Grant Number 24KJ1201).

References

[1] Martín AJ, Mitchell S, Mondelli C, et al. Unifying views on catalyst deactivation. Nat Catal. 2022;5(10):854–866.

[2] Heard AW, Suárez JM, Goldup SM. Controlling catalyst activity, chemoselectivity and stereoselectivity with the mechanical bond. Nat Rev Chem. 2022;6(3):182–196.

[3] Keller G. Synthesis of ethylene via oxidative coupling of methane I. Determination of active catalysts. J Catal. 1982;73(1):9–19.

[4] Fischer F, Tropsch H. The preparation of synthetic oil mixtures (synthol) from carbon monoxide and hydrogen. Brennst-Chem. 1923;4:276–285.

[5] Cruellas A, Bakker JJ, Van Sint Annaland M, et al. Techno-economic analysis of oxidative coupling of methane: Current state of the art and future perspectives. Energy Convers Manag. 2019;198:111789.

[6] Reyes SC, Iglesia E, Kelkar CP. Kinetic-transport models of bimodal reaction sequences—I. Homogeneous and heterogeneous pathways in oxidative coupling of methane. Chem Eng Sci. 1993;48(14):2643–2661.

[7] Su Y. Upper bound on the yield for oxidative coupling of methane. J Catal. 2003;218(2):321–333.

[8] Lin CH, Ito T, Wang J, et al. Oxidative Dimerization of Methane over Magnesium and Calcium Oxide Catalysts Promoted with Group IA Ions: The Role of $[M^+O^-]$ Centers. J Am Chem Soc. 1987;109:4808–4810.

[9] DeBoy JM, Hicks RF. The oxidative coupling of methane over alkali, alkaline earth, and rare earth oxides. Ind Eng Chem Res. 1988;27(9):1577–1582.

[10] Shahri SMK, Alavi SM. Kinetic studies of the oxidative coupling of methane over the $Mn/Na_2WO_4/SiO_2$ catalyst. J Nat Gas Chem. 2009;18(1):25–34.

[11] Qian K, You R, Guan Y, et al. Single-Site Catalysis of Li-MgO Catalysts for Oxidative Coupling of Methane Reaction. ACS Catal. 2020;10(24):15142–15148.

[12] Luo L, Jin Y, Pan H, et al. Distribution and role of Li in Li-doped MgO catalysts for oxidative coupling of methane. J Catal. 2017;346:57–61.

[13] Myrach P, Nilius N, Levchenko SV, et al. Temperature-Dependent Morphology, Magnetic and Optical Properties of Li-Doped MgO. ChemCatChem. 2010;2(7):854–862.

[14] Zavyalova U, Geske M, Horn R, et al. Morphology and Microstructure of Li/MgO Catalysts for the Oxidative Coupling of Methane. ChemCatChem. 2011;3(6):949–959.

[15] Richter NA, Stavale F, Levchenko SV, et al. Defect complexes in Li-doped MgO. Phys Rev B. 2015;91(19):195305.

[16] Kiatsaengthong D, Jaroenpanon K, Somchuea P, et al. Effects of Mg, Ca, Sr, and Ba Dopants on the Performance of La_2O_3 Catalysts for the Oxidative Coupling of Methane. ACS Omega. 2022;7(2):1785–1793.

[17] Schucker RC, J. Derrickson K, K. Ali A, et al. The Effect of Strontium Content on the Activity and Selectivity of Sr-Doped La_2O_3 Catalysts in Oxidative Coupling of Methane. Appl Catal Gen. 2020;607:117827.

[18] Kidamorn P, Tiyatha W, Chukeaw T, et al. Synthesis of Value-Added Chemicals via Oxidative Coupling of Methanes over $Na_2WO_4TiO_2$ $–MnO_x/SiO_2$ Catalysts with Alkali or Alkali Earth Oxide Additives. ACS Omega. 2020;5(23):13612–13620.

[19] Farrell BL, Igenegbai VO, Linic S. A Viewpoint on Direct Methane Conversion to Ethane and Ethylene Using Oxidative Coupling on Solid Catalysts. ACS Catal. 2016;6(7):4340–4346.

[20] Kiani D, Sourav S, Baltrusaitis J, et al. Oxidative Coupling of Methane (OCM) by SiO_2-Supported Tungsten Oxide Catalysts Promoted with Mn and Na. ACS Catal. 2019;9(7):5912–5928.

[21] Ramprasad R, Batra R, Pilania G, et al. Machine learning in materials informatics: recent applications and prospects. Npj Comput Mater. 2017;3(1):54.

[22] Nguyen TN, Nhat TTP, Takimoto K, et al. High-Throughput Experimentation and Catalyst Informatics for Oxidative Coupling of Methane. ACS Catal. 2020;10(2):921–932.

[23] Taniike T, Cannavacciuolo FD, Khoshsefat M, et al. End-to-End High-Throughput Approach for Data-Driven Internal Donor Development in Heterogeneous Ziegler–Natta Propylene Polymerization. ACS Catal. 2024;14(10):7589–7599.

[24] Yanagiyama K, Takimoto K, Dinh Le S, et al. High-throughput experimentation for photocatalytic water purification in practical environments. Environ Pollut. 2024;342:122974.

[25] Le SD, Ton NNT, Seenivasan K, et al. High-throughput screening of multimetallic catalysts for three-way catalysis. Sci Technol Adv Mater Methods. 2024;4(1):2284130.

[26] Jayakumar TP, Suresh Babu SP, Nguyen TN, et al. Exploration of ethanol-to-butadiene catalysts by high-throughput experimentation and machine learning. Appl Catal Gen. 2023;666:119427.

[27] Taniike T, Fujiwara A, Nakanowatari S, et al. Automatic feature engineering for catalyst design using small data without prior knowledge of target catalysis. Commun Chem. 2024;7(1):11.

[28] Gonzalez TF. Clustering to minimize the maximum intercluster distance. Theor Comput Sci. 1985;38:293–306.

[29] Nishimura S, Ohyama J, Li X, et al. Machine Learning-Aided Catalyst Modification in Oxidative Coupling of Methane via Manganese Promoter. Ind Eng Chem Res. 2022;61(24):8462–8469.

[30] Nakanowatari S, Nguyen TN, Chikuma H, et al. Extraction of Catalyst Design Heuristics from Random Catalyst Dataset and their Utilization in Catalyst Development for Oxidative Coupling of Methane. ChemCatChem. 2021;13(14):3262–3269.

[31] Nguyen TN, Nakanowatari S, Nhat Tran TP, et al. Learning Catalyst Design Based on Bias-Free Data Set for Oxidative Coupling of Methane. ACS Catal. 2021;11(3):1797–1809.

[32] Takahashi L, Nguyen TN, Nakanowatari S, et al. Constructing catalyst knowledge networks from catalyst big data in oxidative coupling of methane for designing catalysts. Chem Sci. 2021;12(38):12546–12555.

[33] Yang TL, Feng LB, Shen SK. Oxygen Species on the Surface of La_2O_3/CaO and Its Role in the Oxidative Coupling of Methane. J Catal. 1994;145(2):384–389.

[34] Yoshida R. XenonPy is a Python software for materials informatics. 2018;

[35] Huber PJ. Robust Estimation of a Location Parameter. Ann Math Stat. 1964;35(1):73–101.

[36] Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2013;36(1):27–46.

[37] Fujima J, Tanaka Y, Miyazato I, et al. Catalyst Acquisition by Data Science (CADS): a web-based catalyst informatics platform for discovering catalysts. React Chem Eng. 2020;5(5):903–911.

[38] Matras D, Jacques SDM, Poulston S, et al. Operando and Postreaction Diffraction Imaging of the La–Sr/CaO Catalyst in the Oxidative Coupling of Methane Reaction. J Phys Chem C. 2019;123(3):1751–1760.

[39] Li Y, Niu P, Wang Q, et al. Performance of Zn-Al co-doped La_2O_3 catalysts in the oxidative coupling of methane. J Fuel Chem Technol. 2021;49(10):1458–1467.

[40] Javed HMA, Ahmad MI, Que W, et al. Encapsulation of TiO_2 nanotubes with Cs nanoparticles to enhance electron injection and thermal stability of perovskite solar cells. Surf Interfaces. 2021;23:101033.

[41] Brause M, Skordas S, Kempter V. Study of the electronic structure of TiO_2(110) and Cs/TiO_2(110) with metastable impact electron spectroscopy and ultraviolet photoemission spectroscopy (HeI). Surf Sci. 2000;445(2–3):224–234.

[42] Zhang J, Sun N, Ling L, et al. Effect of different valence metals doping on methane

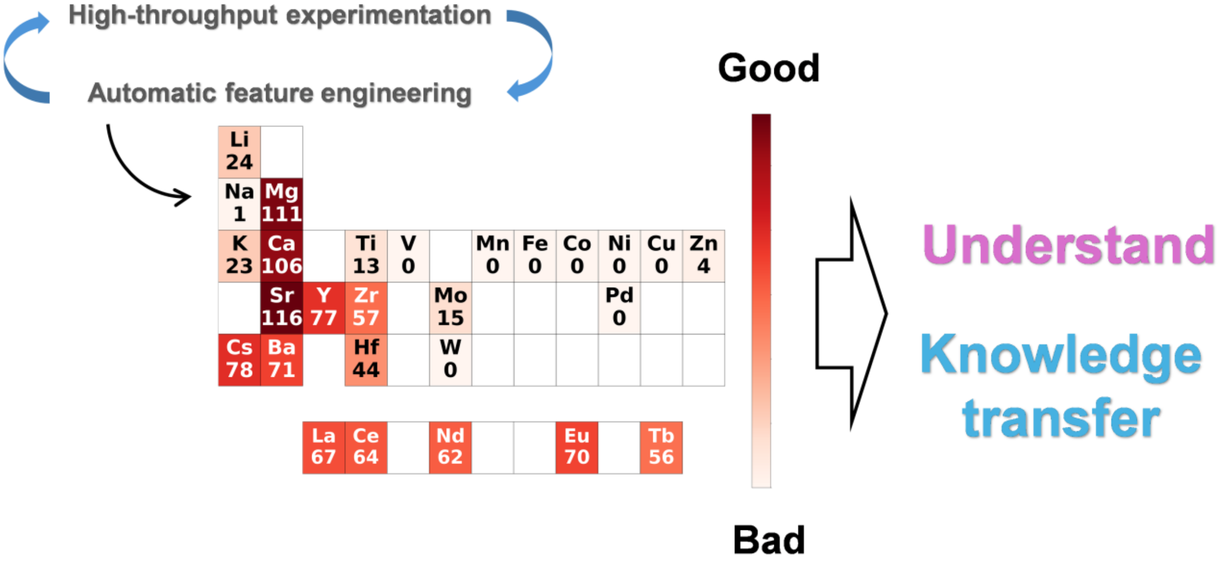
activation over La_2O_3(001) surface. J Fuel Chem Technol. 2023;51(5):673–682.

[43] Elkins TW, Roberts SJ, Hagelin-Weaver HE. Effects of alkali and alkaline-earth metal dopants on magnesium oxide supported rare-earth oxide catalysts in the oxidative coupling of methane. Appl Catal Gen. 2016;528:175–190.

[44] Ji S. Surface WO_4 tetrahedron: the essence of the oxidative coupling of methane over M_W_Mn/SiO_2 catalysts. J Catal. 2003;220(1):47–56.

[45] Kiani D, Sourav S, Baltrusaitis J, et al. Elucidating the Effects of Mn Promotion on SiO_2 -Supported Na-Promoted Tungsten Oxide Catalysts for Oxidative Coupling of Methane (OCM). ACS Catal. 2021;11(16):10131–10137.

[46] van der Maaten L, Hinton GE. Visualizing data using t=SNE. J Mach Learn Res. 2008;9(11):2579–2605.



Impact statement

Integrating high-throughput experimentation and automatic feature engineering in a loop enables the development of robust machine learning models, leading to comprehensive and transferable catalyst knowledge.

Figures and captions

Figure 1. Active learning cycle employed in this study. Automatic feature engineering (AFE) is applied to a given catalyst dataset to derive a design hypothesis—a machine learning (ML) model with tailored descriptors. Catalysts recommended by farthest point sampling (FPS) are evaluated using high-throughput experimentation (HTE). The resulting data are integrated back into the dataset, and this iterative process continues until a robust design hypothesis is established, elucidating the relationship between catalyst compositions and performances.

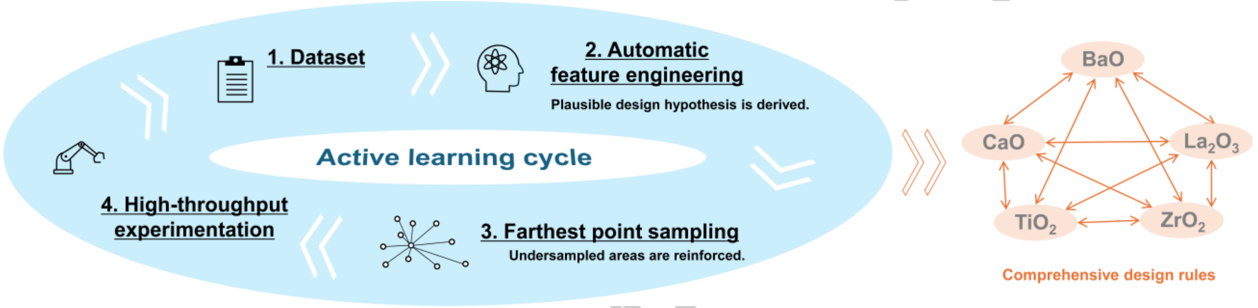

1. Dataset
2. Automatic feature engineering
Plausible design hypothesis is derived.
Active learning cycle
3. Farthest point sampling
Undersampled areas are reinforced.
4. High-throughput experimentation
BaO
CaO
La2O3
TiO2
ZrO2
Comprehensive design rules

Figure 2. Development of the predictive behavior of design hypotheses through active learning. The MAE was calculated by comparing the predicted C_2 yields for the 4059 catalysts between the design hypothesis obtained in one cycle and that from the previous cycle. Lower MAE values indicate greater similarity between the predictive behaviors of the two design hypotheses. In the early stages of active learning, MAE values tended to decrease significantly, reflecting the elimination of less robust design hypotheses due to data addition. In the later stages, further data addition led to minor changes in MAE, suggesting that the design hypotheses became similarly robust. Note that any accidental increase in MAE values during some cycles was due to the random nature of the genetic algorithm during feature selection; In such a case, the sixth cycle was added.

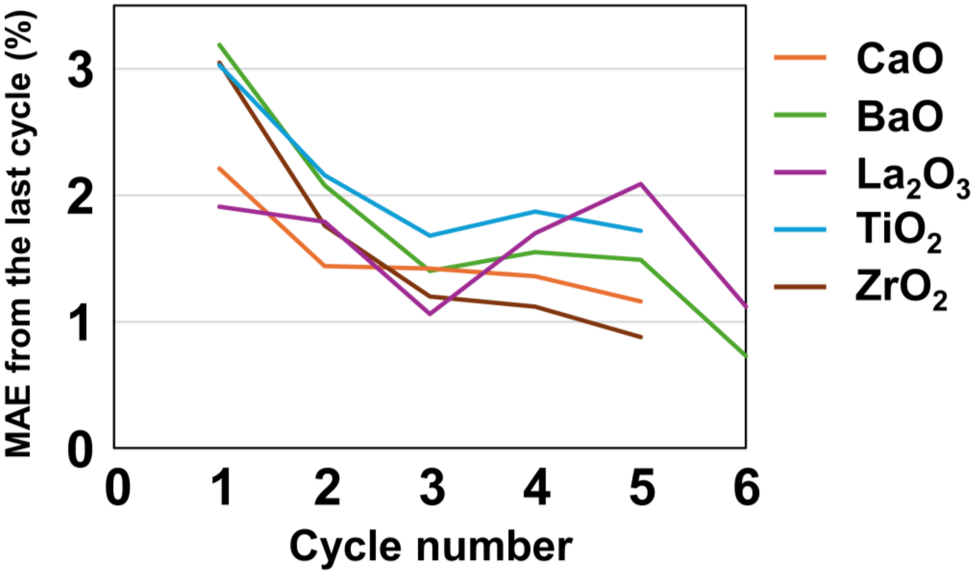
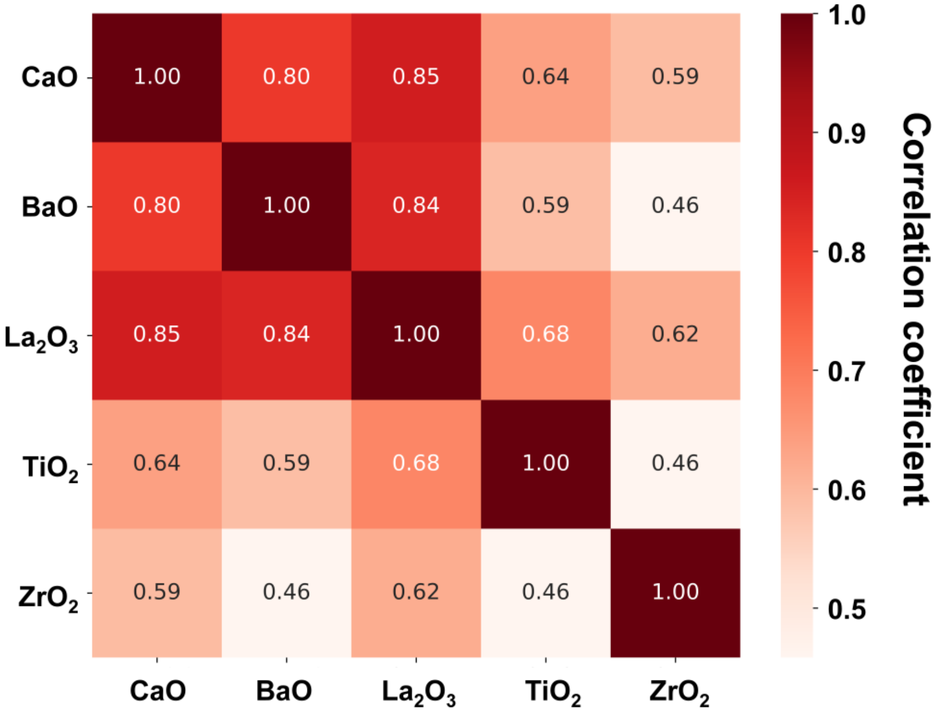

MAE from the last cycle (%)
0
1
2
3
0
1
2
3
4
5
6
Cycle number
CaO
BaO
La_2O_3
TiO_2
ZrO_2

Figure 3. Similarity of design hypotheses across supports. The Pearson's correlation coefficient between the predicted C_2 yields for the 4059 catalysts across different supports was calculated and visualized as a heatmap. A correlation coefficient closer to 1 indicates that two design hypotheses are similar. It can be seen that basic oxides such as CaO, BaO, and La_2O_3 have similar design hypotheses, which differ from those of TiO_2 and ZrO_2.



	CaO	BaO	La_2O_3	TiO_2	ZrO_2
CaO	1.00	0.80	0.85	0.64	0.59
BaO	0.80	1.00	0.84	0.59	0.46
La_2O_3	0.85	0.84	1.00	0.68	0.62
TiO_2	0.64	0.59	0.68	1.00	0.46
ZrO_2	0.59	0.46	0.62	0.46	1.00

Figure 4. Design rules in high-performing catalysts. These summarize the characteristics of catalysts with predicted C_2 yields in the top 10% for each support. (a) A heatmap in a periodic table format visualizes the frequency of appearances of individual elements in the high-performing catalysts. Elements with higher frequencies (shown in darker red) are more likely to contribute to high performance when paired with the respective support. (b) For the top three most frequent elements in (a), pie charts illustrate the frequency of appearances of secondary elements associated with these main elements. If an element is selectively associated with a specific secondary element, it suggests a synergy between these elements.

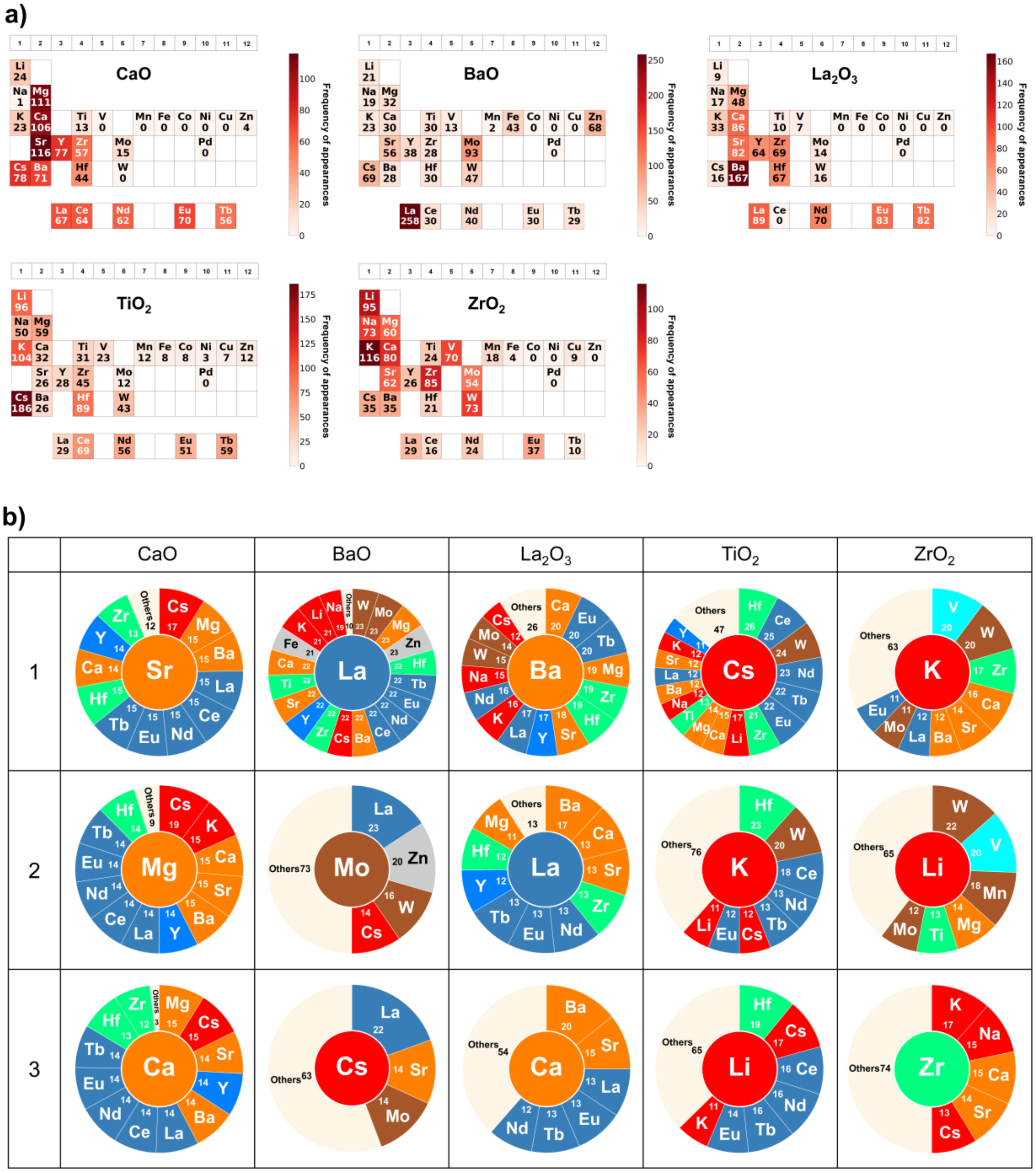
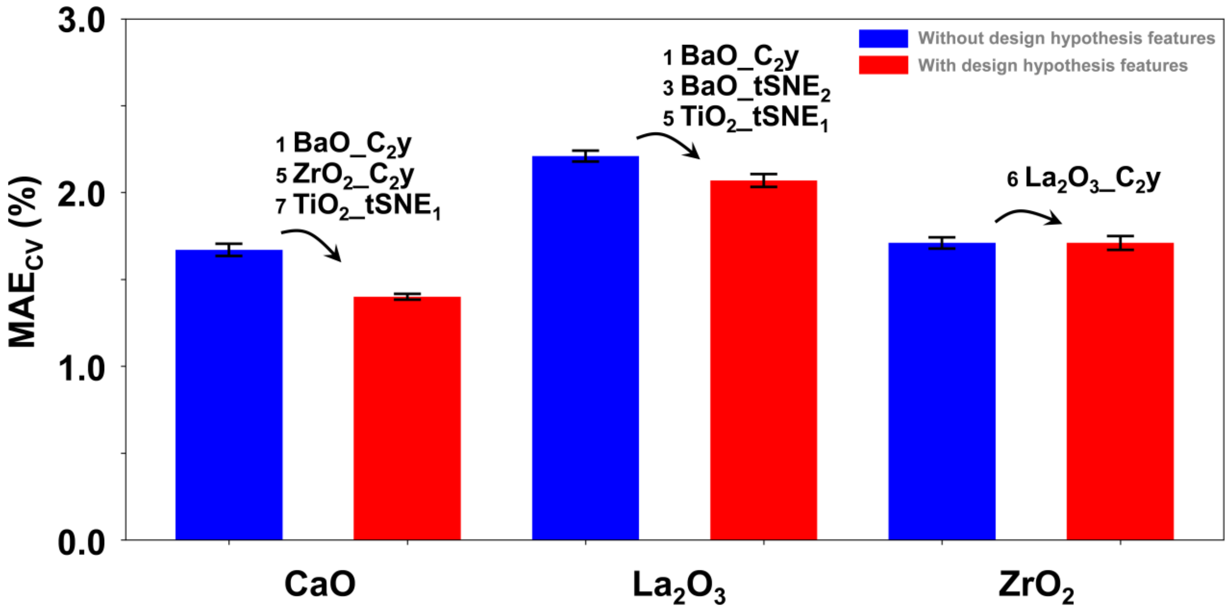

a)
1 2 3 4 5 6 7 8 9 10 11 12
CaO
Li 24
Na 1 Mg 111
K 23 Ca 106 Ti 13 V 0 Mn 0 Fe 0 Co 0 Ni 0 Cu 0 Zn 4
Sr 116 Y 77 Zr 57 Mo 15 Pd 0
Cs 78 Ba 71 Hf 44 W 0
La 67 Ce 64 Nd 62 Eu 70 Tb 56
0 20 40 60 80 100
Frequency of appearances
BaO
Li 21
Na 19 Mg 32
K 23 Ca 30 Ti 30 V 13 Mn 2 Fe 43 Co 0 Ni 0 Cu 0 Zn 68
Sr 56 Y 38 Zr 28 Mo 93 Pd 0
Cs 69 Ba 28 Hf 30 W 47
La 258 Ce 30 Nd 40 Eu 30 Tb 29
0 50 100 150 200 250
Frequency of appearances
La2O3
Li 9
Na 17 Mg 48
K 33 Ca 86 Ti 10 V 7 Mn 0 Fe 0 Co 0 Ni 0 Cu 0 Zn 0
Sr 82 Y 64 Zr 69 Mo 14 Pd 0
Cs 16 Ba 167 Hf 67 W 16
La 89 Ce 0 Nd 70 Eu 83 Tb 82
0 20 40 60 80 100 120 140 160
Frequency of appearances
TiO2
Li 96
Na 50 Mg 59
K 104 Ca 32 Ti 31 V 23 Mn 12 Fe 8 Co 8 Ni 3 Cu 7 Zn 12
Sr 26 Y 28 Zr 45 Mo 12 Pd 0
Cs 186 Ba 26 Hf 89 W 43
La 29 Ce 69 Nd 56 Eu 51 Tb 59
0 25 50 75 100 125 150 175
Frequency of appearances
ZrO2
Li 95
Na 73 Mg 60
K 116 Ca 80 Ti 24 V 70 Mn 18 Fe 4 Co 0 Ni 0 Cu 9 Zn 0
Sr 62 Y 26 Zr 85 Mo 54 Pd 0
Cs 35 Ba 35 Hf 21 W 73
La 29 Ce 16 Nd 24 Eu 37 Tb 10
0 20 40 60 80 100
Frequency of appearances
b)
CaO BaO La2O3 TiO2 ZrO2
1
Sr: Cs 17, Mg 15, Ba 15, La 15, Ce 15, Nd 15, Eu 15, Tb 15, Hf 15, Ca 14, Y 14, Zr 13, Others 12
La: W 23, Mo 23, Mg 23, Zn 23, Hf 22, Tb 22, Eu 22, Nd 22, Ce 22, Ba 22, Cs 22, Zr 22, Y 22, Sr 22, Ti 22, Ca 22, Fe 21, K 21, Li 21, Na 19, Others 10
Ba: Ca 20, Eu 20, Tb 20, Mg 19, Zr 19, Hf 19, Sr 18, Y 17, La 17, K 16, Nd 16, Na 15, W 15, Mo 14, Cs 12, Others 26
Cs: Hf 26, Ce 25, W 24, Nd 23, Tb 22, Eu 22, Zr 21, Li 17, Ca 15, Mg 14, Ti 14, Na 13, Ba 12, La 12, Sr 12, K 12, Y 11, Others 47
K: V 20, W 20, Zr 17, Ca 16, Sr 14, Ba 12, La 12, Mo 11, Eu 11, Others 63
2
Mg: Cs 19, K 15, Ca 15, Sr 15, Ba 15, Y 14, La 14, Ce 14, Nd 14, Eu 14, Tb 14, Hf 14, Others 9
Mo: La 23, Zn 20, W 16, Cs 14, Others 73
La: Ba 17, Ca 13, Sr 13, Zr 13, Nd 13, Eu 13, Tb 13, Y 12, Hf 12, Mg 11, Others 13
K: Hf 23, W 20, Ce 18, Nd 13, Tb 13, Cs 12, Eu 12, Li 11, Others 76
Li: W 22, V 20, Mn 18, Mg 14, Ti 13, Mo 12, Others 65
3
Ca: Mg 15, Cs 15, Sr 14, Y 14, Ba 14, La 14, Ce 14, Nd 14, Eu 14, Tb 14, Hf 13, Zr 12, Others 3
Cs: La 22, Sr 14, Mo 14, Others 63
Ca: Ba 20, Sr 15, La 13, Eu 13, Tb 13, Nd 12, Others 54
Li: Hf 19, Cs 17, Ce 16, Nd 16, Tb 16, Eu 14, K 11, Others 65
Zr: K 17, Na 15, Ca 15, Sr 14, Cs 13, Others 74

Figure 5. Results of transferring knowledge from one catalyst family to another in the context of data science. Design hypotheses obtained from other supports were converted into various features (referred to as design hypothesis features) and used in ML model building with AFE. Models incorporating design hypothesis features (red bars) exhibit a reduction in MAE_{CV} scores compared to models without these features (blue bars) for CaO and La_2O_3, indicating effective knowledge transfer. In contrast, ZrO_2, which lacks similarity to other supports, does not show a reduction in MAE_{CV}. Error bars represent the standard deviation of MAE_{CV} values from three independent feature selection runs. The figure also includes the selected design hypothesis features along with their permutation feature importance rankings.



Tables and captions

Table 1. Development of design hypotheses for CaO-based catalysts throughout the active learning cycles. This table presents the scores and selected features of the design hypothesis obtained in each individual active learning cycle.

Cycle	MAE$_{\text{CV (train)}}$[a]	Selected features[b]
0	1.65% (1.70%)	1. 1/(*covalent_radius_pyykko_double_min*) 2. ln(*covalent_radius_cordero_min*) 3. (*atomic_radius_rahm_min*)$^{1/2}$ 4. (*density_max*)^{2} 5. 1/ln(*thermal_conductivity_ave*) 6. ln(*density_ave*) 7. (*atomic_weight_min*)^{3} 8. 1/exp(*gs_est_fcc_latcnt_max*)
1	1.75% (1.67%)	1. 1/(*covalent_radius_cordero_min*) 2. 1/(*covalent_radius_pyykko_min*) 3. (*atomic_radius_rahm_min*)$^{1/2}$ 4. 1/(*covalent_radius_pyykko_double_min*)$^{1/2}$ 5. ln(*atomic_radius_rahm_min*) 6. exp(*num_f_unfilled_max*) 7. (*gs_mag_moment_max*)^{3} 8. ln(*period_ave*)
2	1.85% (1.79%)	1. (*first_ion_en_max*)^{2} 2. 1/(*polarizability_min*)$^{1/2}$ 3. (*atomic_radius_rahm_min*)$^{1/2}$ 4. 1/(*covalent_radius_pyykko_double_min*)$^{1/2}$ 5. (*num_f_unfilled_std*)$^{1/2}$ 6. (*num_d_valence_max*)^{2} 7. (*hhi_p_max*)^{3} 8. *num_f_valence_ave*
3	1.63% (1.56%)	1. 1/(*covalent_radius_pyykko_double_min*) 2. 1/(*covalent_radius_cordero_min*) 3. (*atomic_radius_rahm_min*)$^{1/2}$ 4. (*num_d_valence_max*)^{2} 5. 1/(*covalent_radius_pyykko_triple_min*) 6. ln(*vdw_radius_uff_max*) 7. 1/(*hhi_p_ave*) 8. 1/exp(*num_f_unfilled_std*)
4	1.67% (1.62%)	1. 1/(*covalent_radius_pyykko_double_min*)$^{1/2}$ 2. (*atomic_radius_rahm_min*)$^{1/2}$ 3. (*covalent_radius_cordero_min*)$^{1/2}$ 4. (*covalent_radius_pyykko_triple_min*)$^{1/2}$ 5. *covalent_radius_cordero_ave* 6. (*num_d_valence_std*)^{2} 7. exp(*fusion_enthalpy_max*)

		8. *heat_capacity_mass_std*
5	1.67% (1.65%)	1. (*atomic_radius_rahm_min*)$^{1/2}$ 2. 1/(*covalent_radius_pyykko_double_min*)$^{1/2}$ 3. ln(*covalent_radius_cordero_min*) 4. (*bulk_modulus_max*)^{2} 5. 1/(*sound_velocity_max*)$^{1/2}$ 6. 1/ln(*covalent_radius_pyykko_double_std*) 7. 1/exp(*heat_capacity_mass_min*) 8. exp(*gs_mag_moment_max*)

[a] Mean absolute error (MAE) in C_2 yields during cross validation (CV) and the same during training in the parentheses.

[b] Eight features were selected that minimized the $\mathrm{MAE}_{\mathrm{CV}}$ value. They are listed in order of permutation feature importance.

Supporting Information of

Acquiring and Transferring Comprehensive Catalyst Knowledge through Integrated High-Throughput Experimentation and Automatic Feature Engineering

Aya Fujiwara[a]*, Sunao Nakanowatari*[a]*, Yohei Cho*[a]*, Toshiaki Taniike*[a]***

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

*Corresponding author. E-mail: taniike@jaist.ac.jp

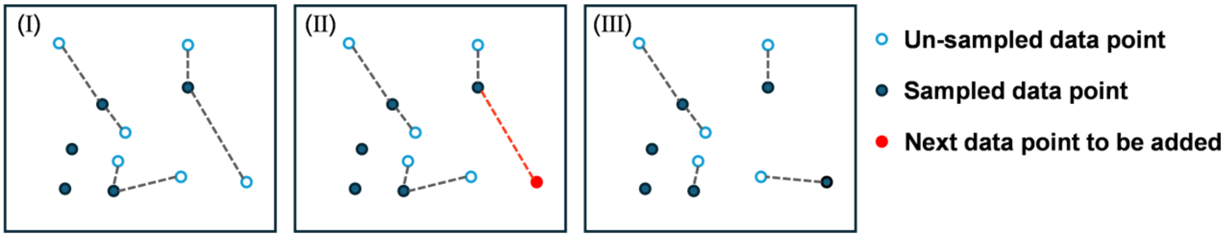


Figure S1. Illustration of selecting data using FPS. To determine the next data point to be sampled, the following operations were repeated. I. In the selected feature space, calculate the euclidean distance from each un-sampled data point to all sampled data points. Select the minimum distance to each un-sampled data point. II. Among the un-sampled data points, select the one with the largest minimum distance (farthest from the sampled data points). III. The selected data point is then removed from the un-sampled data points and added to the sampled data points.

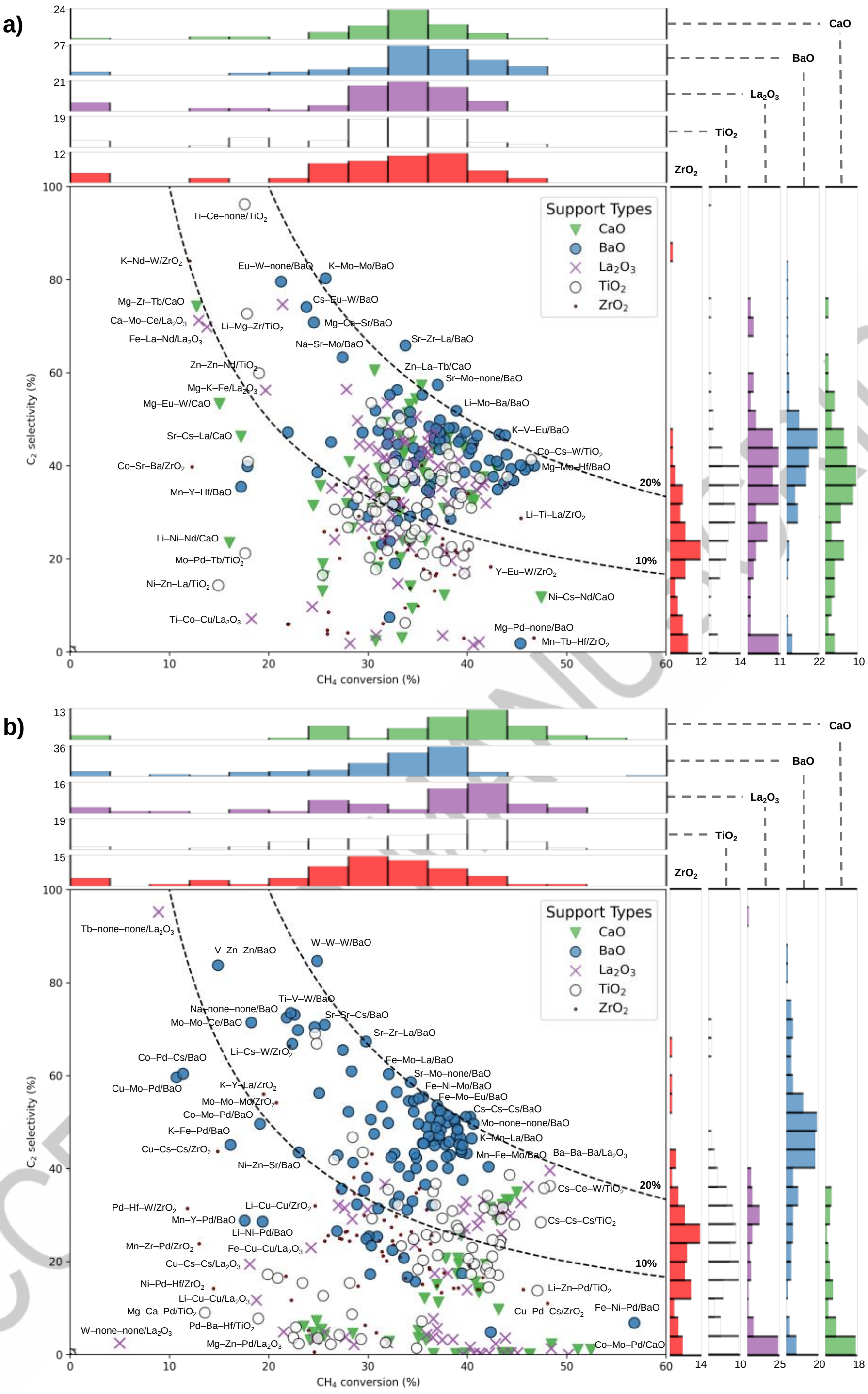


Figure S2. Scatter plot shows 726 catalysts experimentally tested in our previous and current studies, with a) 381 catalysts previously evaluated and b) 333 catalysts evaluated in this study plotted separately. The x-axis represents CH_4 conversion, while the y-axis represents C_2 selectivity. Different supports are represented by symbols with varying colors and shapes. Histograms outside the plot area show the data distribution for each support.

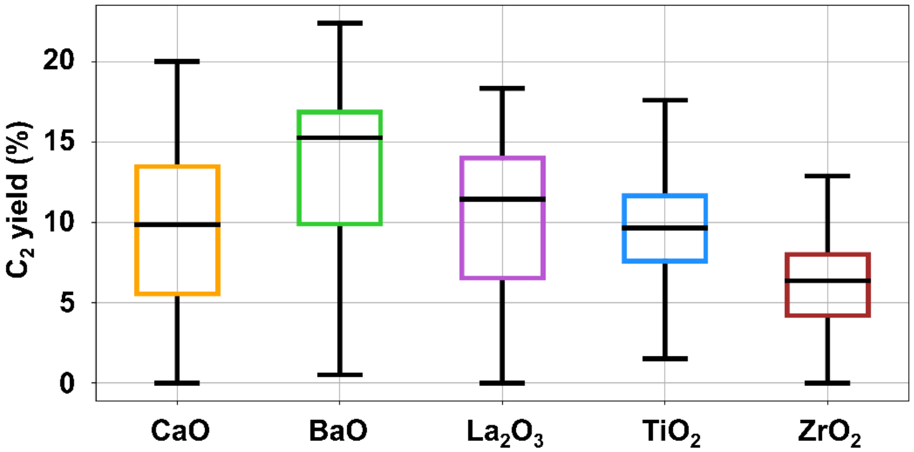


Figure S3. Box plot of predicted C_2 yields for each support, where the line: the median value, the box: the interquartile range (IQR), and the whisker: 1.5 times the IQR with any data points beyond the whisker regarded as outliers.

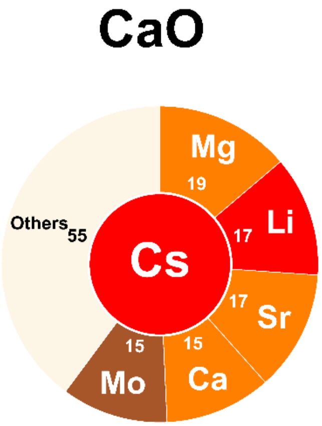


Figure S4. Pie chart illustrating the frequency of appearances of secondary elements associated with Cs–CaO.

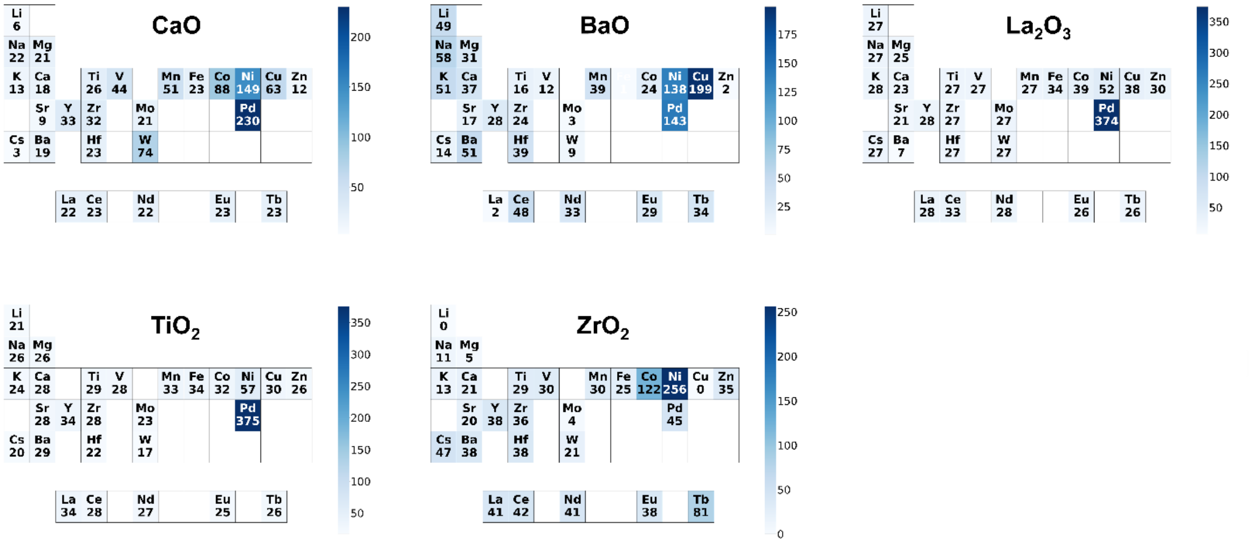


Figure S5. Design rules in low-performing catalysts. These summarize the characteristics of catalysts with predicted C_2 yields in the bottom 10% for each support. A heatmap in a periodic table format visualizes the frequency of appearances of individual elements in the low-performing catalysts. Elements with higher frequencies (shown in darker blue) are more likely to contribute to low performance when paired with the respective support.