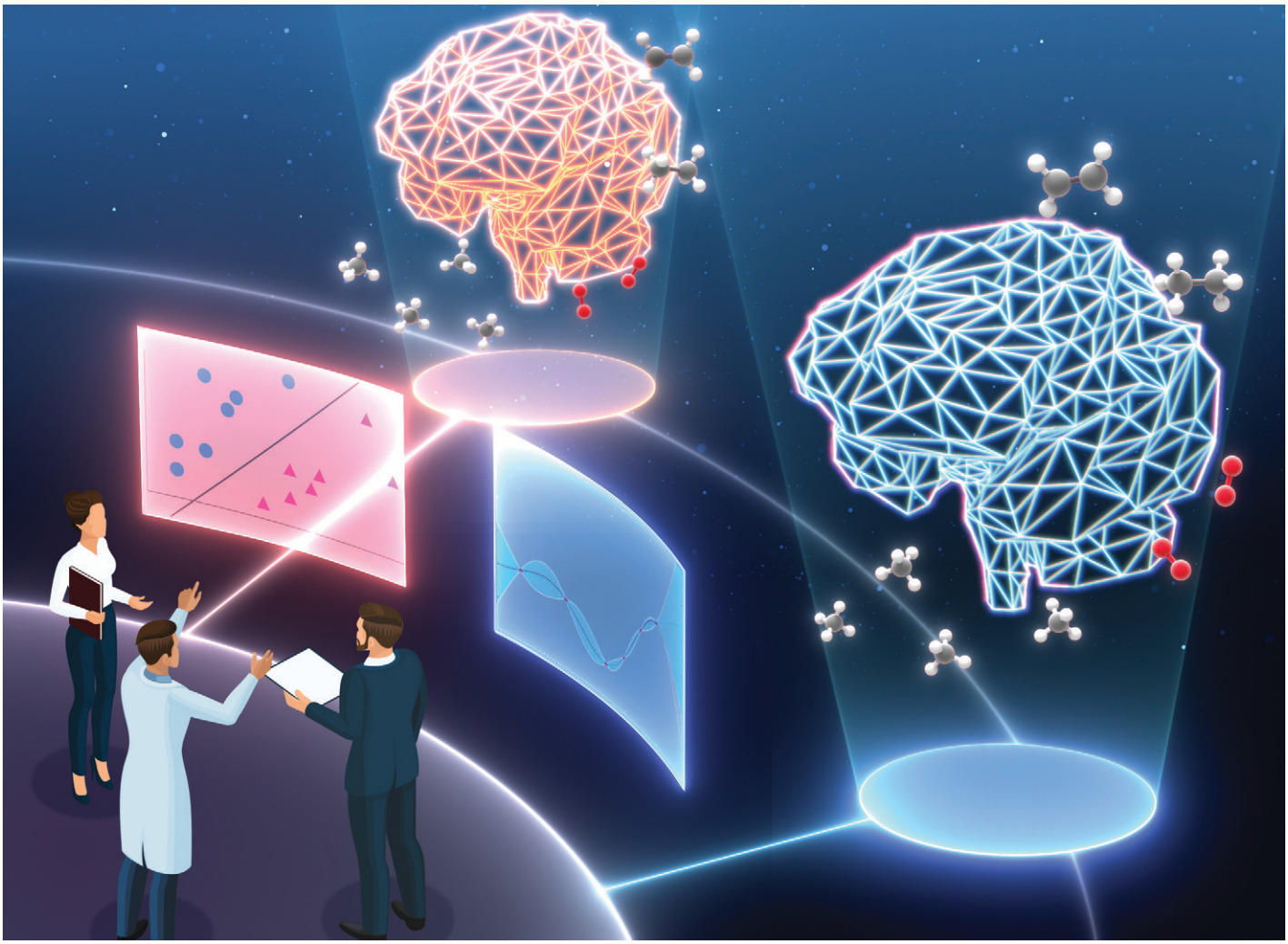


Showcasing research from Dr. Nishimura's laboratory, Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan.

Leveraging machine learning engineering to uncover insights into heterogeneous catalyst design for oxidative coupling of methane

Support vector regression and Bayesian optimization techniques were implemented for literature data-driven catalyst designs for the oxidative coupling of methane to clarify future challenging subjects for machine learning-assisted catalyst investigation.

Image credit: Art Action Inc., Takahiro Tamura

As featured in:

See Shun Nishimura, Keisuke Takahashi *et al.*, *Catal. Sci. Technol.*, 2023, **13**, 4646.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/catalysis

Registered charity number: 207890

Catalysis Science & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Catal. Sci. Technol.*, 2023, 13, 4646

Leveraging machine learning engineering to uncover insights into heterogeneous catalyst design for oxidative coupling of methane†

Shun Nishimura, *[a] Xinyue Li,[a] Junya Ohyama [b] and Keisuke Takahashi *[c]

Machine learning (ML)-assisted catalyst investigations for oxidative coupling of methane (OCM) are assessed using published datasets that include literature data reported by different research teams, along with systematic high-throughput screening (HTS) data. Support vector regression (SVR) is performed on the selected 2842 data points. The first SVR leads to eight catalysts with C_2 yields higher than 15.0% under the current reaction conditions, but the second attempt with the updated dataset including the first validation results does not improve the prediction because of spatial shrinkage. The Bayesian optimization processes also start with datasets of 3335 data points, and are considered for three cycles using the updated dataset. Repeating the Bayesian processes certainly improves the C_2 yields observed in the validation results, but the convergence of the elements presents another issue. Accordingly, data-driven catalyst investigations involve a different set of defect issues from the conventional style of catalyst investigations. The unveiling of issues in the highly active OCM catalyst investigation by ML engineering conducted for this study is intended to clarify future challenging subjects for ML-assisted research innovations. Actions to proactively discover the encounters with serendipity to broaden the scope of the material survey area using ML approaches and/or working with the researcher's intuition can increase the possibility of fortuitous discoveries and the achievement of desired outcomes.

Received 29th April 2023,
Accepted 26th May 2023

DOI: 10.1039/d3cy00596h

rsc.li/catalysis

1. Introduction

Informatics approaches have been eagerly pursued by catalyst scientists in recent years. Data-driven catalyst design and discovery of hidden trends using machine learning (ML) engineering and/or data management are hoped to guide direct access to the goal of achieving desired performance more effectively than using conventional approaches.[1–6] In fact, the combined use of informatics techniques can suggest "unreported" areas in some cases, and can suggest unexpected areas in catalyst research. The exploration of these areas can engender new motivations for revealing important hidden characteristics in components known to have catalyst performance. Literature data are commonly available classic big data. Analyses of these data from an informatics perspective have revealed common trends of catalyst components and the roles of the respective element, which has been helpful for subsequent proposals of catalyst design.[7–11] Nevertheless, the inherent biases arising from differences among research groups related to aspects such as preparation methodology, reactor design, and elucidation manner must be considered carefully because these biases can sometimes mislead the ML considerations.[12–14] A recent paper presents the argument that even identifying the best performance based on the literature remains challenging because of the variety of reaction situations.[15] Since the concept of high-throughput screening (HTS) was discussed in 1970,[16] the design and use of HTS to generate large datasets have received much attention. Various types of atmospheric fixed-bed reactor systems consisting of 6,[17,18] 16,[19] 20,[20] 48,[21] 49,[22] and 64 (ref. 23) parallel tubes have enabled the systematic screening of large data. The combined use of a "smart" laboratory with robotics[24] is expected to transform big data-driven informatics investigations of catalysts into a popular style of catalyst research in the near future. However, interestingly, the HTS approach reportedly generates much "garbage data" during experimental trials,[25] and another report presents a debate about the importance of such extra "negative data".[26] Apparently, emphasizing that the skills of conventional catalyst scientists are still needed is better to guide the story of such

[a] *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi 923-1292, Japan. E-mail: s_nishim@jaist.ac.jp*
[b] *Faculty of Advanced Science and Technology, Kumamoto University, 2-39-1 Kurokami, Chuo-ku, Kumamoto 860-8555, Japan*
[c] *Department of Chemistry, Hokkaido University, North-10, West-8, Sapporo 060-0810, Japan. E-mail: keisuke.takahashi@sci.hokudai.ac.jp*

† Electronic supplementary information (ESI) available: List of chemicals, sequence of reaction, performance of NaMnW/SiO_2 standard, and raw data of ML predictions. See DOI: https://doi.org/10.1039/d3cy00596h

innovative research technologies properly. Computational simulations such as density functional theory (DFT) calculations are the third type of big data. They have been used conventionally as a tool to prove the expected mechanisms for the characteristic properties of developed catalysts. With increasing attention devoted to data-driven catalyst investigation, the conventional DFT goal has shifted from understanding to discovering catalysts, leading to the next idea for catalyst design.[27,28] The cost of DFT processing remains high, but DFT for screening catalysts[29,30] is also proving to be a powerful tool for the next style of catalyst design. By virtue of the development of innovative technologies such as quantum computing, computational simulation can again be a game changer in the field of materials science. Big data of these three types can reveal emerging movements of the catalyst informatics: the literature, HTS experiments, and computational simulation data-driven catalyst investigations.[31]

The application of supervised ML to the discovery of catalysts with exceptional catalytic performance still entails many persistent issues related to its capabilities and accuracy for validation. A well-defined trained ML regression model, by its very nature, follows popular rules in original datasets, *i.e.*, relations between a material and its performance as determined by some selected descriptors. For this reason, it constitutes a successful approach when the common rule can represent the feature at the outer range. However, in most cases, because of the nature of ML regression, outlier performance data cannot be predicted directly from typical trends in the data. Furthermore, a persistent issue is the difficulty in ascertaining global descriptors that represent specific trends between catalyst materials and catalysis characteristics. This difficulty particularly arises for heterogeneous catalyst areas because the heterogeneous catalyst performance depends on multi-dimensional characteristics such as the catalyst components, loading level, size and morphology, crystallinity, oxidation state, density of active species, surface roughness, defects, and acid–base nature. Indeed, clarification of the causal analysis requires a long history, except in the cases of selected components and a simple target reaction. Some examples are CO oxidation over gold-based catalysts at ambient temperature, for which a long discussion has been had for the identification of active sites.[32] Elucidating such multiple networks between the nature of catalyst materials and the trends of catalysis features requires a great cost, which might never improve and which might yield only temporary results. Therefore, more important than the construction of a comprehensive ML model, unveiling of the current issues in supervised ML-aided heterogeneous catalyst investigation can clarify future topics for this innovative research field: how the "imperfect" supervised ML regression acts during steps in a catalyst trial-and-error process for catalyst investigation.

For this study, oxidative coupling of methane (OCM), which was discovered in the 1980s,[33,34] is chosen as a model reaction for ML-aided catalyst investigation. For OCM, there is a 40 year history of catalyst studies using conventional methods. It is noteworthy that, in 2014, one US start-up company established a pilot scale OCM process to provide ethylene to the US market,[35] but the cost-effective design of OCM plants remains a challenge compared to the process *via* naphtha cracking.[36,37] Consequently, the successful implementation of ML to aid OCM catalyst investigation is an attractive dream to propose an alternative path forward for this research area. The present study, based on big data generated from earlier literature data along with systematic HTS data,[11,20,38] investigates supervised ML using support vector regression (SVR) and Bayesian optimization with an expected improvement (EI) function for ternary element supported OCM catalysts (M1–M2–M3/support).

2. Experimental

2.1. OCM reaction

The OCM reaction was conducted in a conventional fixed-bed reactor system with a tube furnace (L = 270 mm; ARF-30KC, Asahi Rika Co. Ltd., Japan). The well-ground powder catalyst (50 mg) was sandwiched between quartz glass wool (<10 mg as sum) at the neck position of a step-jointed quartz tube reactor (4 mm ID, 235 mm length (from top) – 2 mm ID, 150 mm length (from bottom)). The reactor temperature was monitored using an R-type thermocouple, with the tip placed near the outer quartz wall of the catalyst bed location. The *in situ* catalyst pre-treatment was applied at 500 °C for 30 min under an O_2 flow (30.0 ml min^{-1}); then the OCM performance was investigated at 500–850 °C at 25 °C intervals under a $CH_4/O_2/N_2$ flow (21.0/7.0/3.0 ml min^{-1}) (Fig. S1 in the ESI†). The reaction mixture was evaluated after 3–4 min at each reaction temperature using a device (Micro GC FusionTM; INFICON Co., Ltd.) equipped with a dual-column system consisting of an Rt-Molsieve 5A column (0.25 mm × 10 m, Ar carrier, backflush) and an Rt-U-Bond column (0.25 mm × 8 m, He carrier). The amounts of the target gases H_2, O_2, N_2, CH_4, CO, CO_2, C_2H_4, and C_2H_6 were estimated according to our earlier reports using N_2 as an internal standard.[39,40]

2.2. Catalyst preparation

All chemicals used for this study are presented in Table S1 in the ESI.† Multicomponent M1–M2–M3 supported catalysts were prepared with co-impregnation using a parallel synthesis method.[39,40] All elemental resources (0.20 mmol for each) and support materials (1.0 g) were placed in a glass tube (φ18) with 6 mL of highly purified water (18.2 MΩ × cm), and were mixed at 50 °C for 6 h under vigorous stirring with a magnetic stirrer. The slurry was centrifuged under vacuum at 80 °C and was dried overnight at 110 °C. The resulting precipitate was well-ground using an alumina mortar and was placed in an alumina crucible (φ54); then it was calcined at 900 °C for 3 h in a furnace (KDF 300-Plus; Denken Highdental Co. Ltd.). Water-sensitive metal resources such as $Ti(OiPr)_4$ and $BiCl_3$ were impregnated successively with ethanol solvent (two-step) before calcination. For the ethanol solvent, a set at 60 °C was used for centrifugation

and the vacuum step. The reference catalysts none(M1)–none(M2)–none(M3)/support (denoted as "bare") were prepared using the same protocol with no metal resources.

Conventional NaMnW/SiO_2 was prepared using co-impregnation and was used as a standard catalyst to determine the potential of the catalysts, as in our earlier studies.[13,29,39,40] Both 0.93 mmol of $Mn(NO_3)_2{\cdot}6H_2O$ and 0.37 mmol of $Na_2WO_4{\cdot}2H_2O$ were dissolved in 300 mL of deionized water in a round-bottom flask. Then, after 2.5 g of SiO_2 was added to the flask with vigorous stirring using a magnetic stirrer, it was mixed for 24 h at 50 °C. The water solvent was removed gradually using a rotary evaporator system heated to 65 °C. The resultant product was dried overnight at 110 °C. The resulting powder was well ground using an alumina mortar, placed in an alumina crucible (φ73), and was calcined at 1000 °C for 3 h in a furnace (KDF 300-Plus). As shown in Fig. S2 in the ESI,† the as-prepared NaMnW/SiO_2 has good performance in the OCM reaction, with the best C_2 yield value of 17.6% at CH_4/O_2 = 3.0 and 19.9% at CH_4/O_2 = 1.8 under the present experiment conditions: 31.0 ml min^{-1} of total flow including N_2 balance (3.0 ml min^{-1}, const.) at a furnace length of 270 mm.

2.3. Machine learning methodology

For the OCM reaction, open-source systematic high-throughput screening (HTS) datasets of 300 random catalysts[38] and 59 NaMnW-based catalysts[20] were provided by Taniike *et al.*, whereas 4759 experiment data points reported in the literature and patents were collected by Shimizu *et al.*[11] For the 300 + 59 HTS datasets, only "the best C_2 yield value" of each catalyst and its reaction conditions were selected to reduce the data effects of HTS in the data source. Pre-treatment of the datasets was conducted according to the following points. Particularly, the literature data include incomplete data such as i) the sum of the partial pressures of the reactants is higher than 1.0, ii) information related to the flow rate and/or the support is missing, and iii) the label of the support is inadequate as metal catalyst. In some cases, the authors can manually expect the type of support when the cation % is higher than 60%. In addition, cations of 4–5 kinds, anions of 1–2 kinds, and mixed support usage information of 2–3 kinds were removed because they were outliers of the target of this study. From the HTS datasets, none(M1)–none(M2)–none(M3) was also removed. Then, 2842 data points were used as the dataset at the beginning of this study. They are presented as **List0.csv** in the ESI.† Accordingly, the element survey area used for the present study was found with 52 loading element variations from Ag, Al, Au, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Ge, Hf, In, K, La, Li, Lu, Mg, Mn, Mo, Na, Nb, Nd, Ni, Pb, Pd, Pr, Pt, Rb, Re, Sb, Si, Sm, Sn, Sr, Tb, Th, V, W, Y, Yb, Zn, Zr, and "none", and 37 support variations from Al_2O_3, BEA zeolite, BN, BaO, BeO, Bi_2O_3, CaO, CeO_2, Dy_2O_3, Eu_2O_3, Fe_2O_3, Gd_2O_3, K_2O, La_2O_3, Li_2O_3, MgO, Mn_2O, Na_2O_3, Nb_2O_5, Nd_2O_3, Pb_2O_3, Pr_2O_3, Sb_2O_3, Sc_2O_3, SiC, SiC nanofibers, SiO_2, Sm_2O_3, SnO, SrO, Tb_2O_3, ThO_2, Y_2O_3, Yb_2O_3, ZSM-5 zeolite, ZnO, and ZrO_2.

To represent the catalyst material and reaction condition information, the following methods are implemented in the descriptor setting: i) catalyst component represented by a one-hot encoding manner, where binary numbers 0 and 1 are assigned into the box in the survey table; and ii) reaction temperature, which was divided by 10 for input data, and partial pressure of iii) CH_4 gas (p_{CH_4}), iv) O_2 gas (p_{O_2}), and v) balance gas (p_{Inert}). The use of one-hot encoding is particularly helpful to reduce the space expansion of the sort order of ternary components (permutations, ${}_3P_3$ = 6 ways). The division process at temperature is for space control for space awareness by ML. The ML output data under the diluted condition of $p_{Inert} > 0.2$ were excluded according to the validation condition at $p_{N_2} \fallingdotseq 0.1$. In addition, the predicted components with the element thorium (Th) are skipped for validation because of the very low availability of Th salts.

Support Vector Regression (SVR) was implemented with a radial basis function kernel of C = 14 and gamma = 0.25. Cross-validation was examined with a train and test split of 80% and 20%, evaluated using the mean R^2 values on 10 random data splits. The Gaussian process regression was implemented using Scikit Learn.[41] The kernel of the Gaussian process regression was optimized, where the kernel consists of WhiteKernel, ConstantKernel × Radial Basis Function (RBF), and ConstantKernel × DotProduct as described in our recent report.[42] The standard deviation (SD) of the predicted variable distribution at a data point is also calculated during Gaussian process regression. Bayesian optimization was applied to find data points with large SDs and with high C_2 yield by Gaussian process regression based on the acquisition function of updated Expected Improvement (EI) calculated using the following eqn (1) and (2) as:

$$U = (y_{max} + \mu - \xi)/\sigma \quad (1)$$

$$EI = \sigma \times U \times \Phi(U) + \sigma \times \varphi(U) \quad (2)$$

where y_{max} stands for the highest C_2 yield in the trained dataset, ξ is the optimization parameter calculated from the SD of the predicted variables multiplied by 0.01, μ and σ respectively denote the predicted C_2 yield and SD calculated using Gaussian process regression, and $\Phi(U)$ and $\varphi(U)$ respectively represent the cumulative distribution function and probability density function of U in eqn (1).[42]

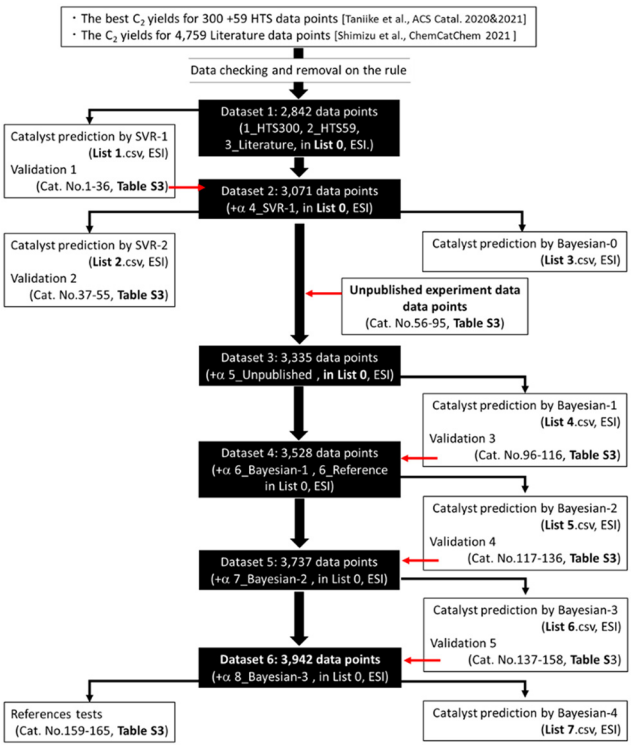
2.4. Workflow

Catalyst validations were performed under conditions not specified by ML, but under identical conditions to the fixed compositions of the $CH_4/O_2/N_2$ flow (21.0/7.0/3.0 mL min^{-1}). Because the reactivity in OCM is influenced by both reaction conditions[43,44] and reactor design,[45,46] clarification is based not

on absolute, but on relative OCM performance in comparison to the well-known NaMnW/SiO_2 catalyst as a standard catalyst under identical conditions, which has the best C_2 yield of 17.6% under the standard $CH_4/O_2/N_2$ flow conditions (21.0/7.0/3.0 mL min^{-1}) at a furnace length of 270 mm (*vide infra*). The strict reaction situation with high $CH_4 + O_2$ concentration (*ca.* 90.3%) is selected in this study for validation. Once a markedly active OCM catalyst has been found by this protocol, optimization of the reaction conditions and reactor design can be specifically examined. To update the input data, experimental validation results with C_2 yield higher than 5.0% are added to the dataset for the next ML prediction. All input data used for this study are presented in **List0.csv** in the ESI.† As shown in Scheme 1, two-times trials of SVR prediction and three times-trials (+one) of Bayesian-optimization processing by EI were investigated. All experimentally obtained data are presented in Table S3, in the ESI.† Experimentally obtained data for SVR-2 were not included in the dataset for additional prediction and validation studies, but these components, Cat. No. 37–55 in Table S3 in the ESI,† were excluded for Bayesian validations to reduce the experiment cost because the catalyst potentials have already been investigated under the current conditions.

3. Results

The first SVR (denoted as SVR-1) implemented in the selected 2842 data points based on the systematic HTS and literature



Scheme 1 Schematic illustration of the dataset history for prediction and validation. At the steps indicated by the red arrows, the validation data points with C_2 yield greater than 5.0% were added to the datasets for the subsequent prediction.

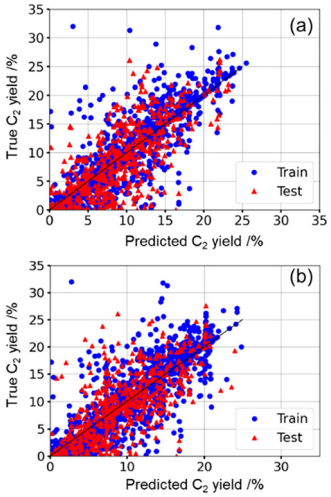


Fig. 1 Cross-validation plots for the (a) first and (b) second trials by SVR on the best scoring case out of 10 examinations.

data obtained a mean R^2 score of 0.54 in the cross-validation (Fig. 1(a)). The data referred from the literature include much information to describe both the catalyst components in terms of the preparation method, metal loading, metal rate, and activation protocol, in addition to the catalyst performance by reaction conditions, reactor design, elucidation methodology, *etc.* When these categories are set individually as a descriptor index, the amount of data is insufficient for regression by ML because each report in the literature includes different descriptor index information collected according to the different experiences of each research team. In addition, even if common information exists for several descriptors collected from data in the literature, excessive variation of descriptors engenders overfitting problems. In brief, the existence of poor relations of descriptor indices with the target function will misguide the literature data-driven ML. From the viewpoint of extracting loose correlation from literature data, for this study, the catalyst component by element variation (M1–M2–M3/support) and reaction conditions by p_{CH_4}, p_{O_2}, and p_{Inert} are selected as the descriptors. The obtained R^2 score value is not higher than 0.6, which is one threshold for a good regression model. However, the authors believe that the list of ML predictions for the first validation presents nice ideas for our goal: the development of OCM catalysts for M1–M2–M3/support with a high C_2 yield performance.

The list of catalyst components and corresponding reaction conditions proposed by SVR-1 is presented in **List1.csv** in the ESI.† When the export range is set to the predicted C_2 yield higher than 18.00% under $p_{Inert} \leq 0.2$ conditions, there are 92 lines of predictions, including M1–M2–M3/support components, reaction temperature, p_{CH_4}, p_{O_2}, p_{Inert}, and the corresponding C_2 yield value. There are 36 components ranging from 18.00 to 22.61% of the C_2 yield. In the first validation, all 36 catalysts are prepared and evaluated for these reactivities for OCM. At this time, 11 out of 36 (30.6%) catalysts are binary element supported

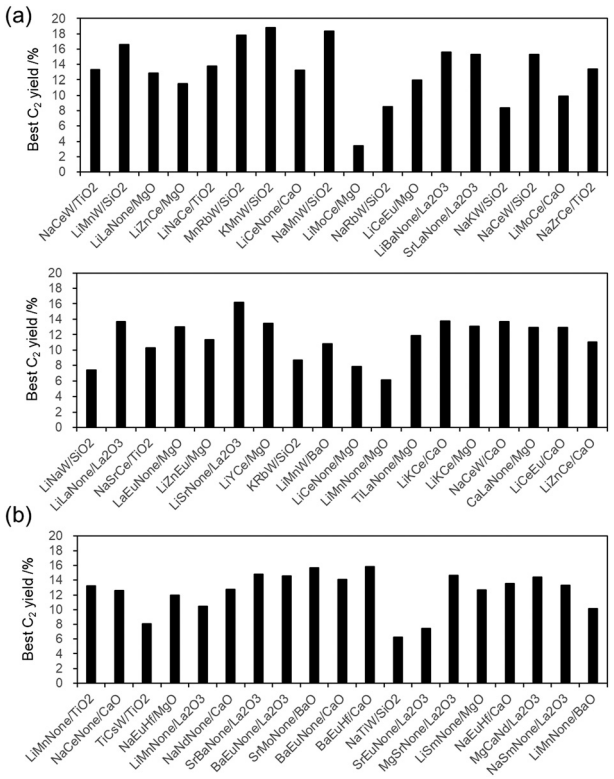


Fig. 2 Best C_2 yield plots of the (a) first and (b) second validation based on the SVR method.

catalysts. As shown in Fig. 2(a), the results include eight catalysts with C_2 yield higher than 15.0% under the present reaction conditions: $LiMnW/SiO_2$ (16.6%), $MnRbW/SiO_2$ (17.8%), $KMnW/SiO_2$ (18.8%), $NaMnW/SiO_2$ (18.3%), $LiBaNone/La_2O_3$ (15.6%), $SrLaNone/La_2O_3$ (15.3%), $NaCeW/SiO_2$ (15.3%), and $LiSrNone/La_2O_3$ (16.2%). The two categories of catalyst with high C_2 values are Na–Mn(or Ce)–W/SiO_2-derivatives and M1–M2–none/La_2O_3 categories. To investigate novel catalysts for OCM with high C_2 yields further, the second SVR prediction (SVR-2) is investigated based on the updated dataset of 3071 data points. One of our observations about this strategy is that ML prediction might allow revisitation of its predictions after validation trials, much as conventional scientists do during catalyst development. To increase the influence of the exact reaction situation and to reduce the influence of the literature and the HTS experimental situation, all data points of C_2 yield higher than 5.0% from the first validation results were added for the next validation. The cross-validation is shown in Fig. 1(b). Its mean R^2 score is 0.54, which is the same as that of SVR-1. The results of the second prediction by SVR are listed in **List2.csv** in the ESI.† The maximum value of the predicted C_2 yield (21.73%) was similar to 22.61% at SVR-1. It contains 11 lines with the predicted C_2 yield higher than 18.00%, with three types of $NaCeW/TiO_2$, $NaMnW/SiO_2$, and LiMnNone/MgO catalysts under different reaction conditions. However, these three have already been examined in the first validation based on SVR-1. They are duplicate components. In other words, after importing the first validation data, the SVR was not able to suggest other potential OCM catalysts based on C_2 yields higher than 18.00%: the space for predicting catalyst components was reduced in high C_2 yield ranges (>18.00%). Reaction conditions might still strongly influence these OCM performances (*viz.* C_2 yield). However, this point is not the subject of this study. When the screening area for data extraction was extended to the predicted C_2 yield above 16.00%, an additional 255 lines were suggested. Therefore, the second prediction list includes catalyst components of 91 types, including 14 duplicates with first validation in the range of predicted C_2 yields from 16.00% to 21.73%. It is noteworthy that 50 out of 91 types (54.9%) are made by binary element-supported categories (M1–M2–none/support). Moreover, it is apparent that a lower diversity of OCM catalysts has appeared in the continuous use of the SVR way for the second regression. Tentatively, catalysts with a predicted C_2 yield value higher than 16.75% were selected for the second validation, excluding duplicate components from the first validation. Accordingly, the 19 catalysts are examined. As presented in Fig. 2(b), the two C_2 yield values of 15.9% for BaEuHf/CaO and 15.7% for the SrMoNone/BaO catalyst are observed as values higher than 15.0%. In other words, SVR-2, which included real data after the validation of SVR-1, failed to improve the experimentally obtained results in the validation.

To explore other possible approaches for ML prediction, the authors specifically examine Bayesian optimization based on the EI index, the score of which guides the experiment to a higher potential value at the data missing pieces in the C_2 yield space.[42] Preliminarily, Bayesian optimization was implemented using the 3071 data points of Dataset 2 (denoted as Bayesian-0, in Scheme 1). The results for predicted C_2 yields higher than 16.00% are presented in **List3.csv** in the ESI.† At this stage, 263 out of 637 lines (41.3%) are binary element supported catalysts. When a component survey is conducted from the high EI values, excluding both duplicate components with validation first and second and Th containing elements, the following 25 catalysts can be selected as candidates for the subsequent validation: NaCeW/BaO, $KSmNone/La_2O_3$, $NaCeW/La_2O_3$, $NaKNone/La_2O_3$, $NaSrNone/La_2O_3$, LiKSm/CaO, $KEuNone/La_2O_3$, NaSrNone/BaO, $LiKNone/La_2O_3$, NaCeNone/BaO, LiSrNone/BaO, LiBaNone/ZnO, $KMoNone/La_2O_3$, LiNaNone/BaO, $LaBaNone/La_2O_3$, $KBaNone/La_2O_3$, LiCeSm/CaO, $NaKNone/La_2O_3$, LiKMo/CaO, $KLaNone/La_2O_3$, $KSrNone/La_2O_3$, KSmNone/CaO, NaBaNone/BaO, LaBaNone/ZnO, and KCeSm/CaO. Despite the application of Bayesian optimization based on the EI index, a high occurrence of binary element supported catalysts (19/25 catalysts) was achieved.

To expand the study area for Bayesian optimization investigation, 264 unpublished data points from 39 catalysts prepared using the same preparation protocol and evaluated with the same reactor and profiles were applied in our laboratory. Then the populated Dataset 3 with 3335 data points was implemented into the Bayesian optimization (denoted as

Bayesian-1 in Scheme 1). There are 406 lines with C_2 yields higher than 16.00%; 38.2% (155 lines) are still the binary element supported catalysts. Consequently, 20 catalysts selected tentatively based on the high EI score without duplicate catalysts from Cat. No. 1–95 in Table S3 in the ESI,† were tested for validation 3. Repeatedly, we examined validation 4 by Bayesian-2 and validation 5 by Bayesian-3 according to the same catalyst selection procedures: duplicate components with earlier validation stages were skipped for the next validation. The corresponding prediction lists are included in the ESI† as the **List4**, **List5**, and **List6.csv** files. In variation 5 based on Bayesian-3, we did not test the 20 catalysts selected from the higher EI value, but instead we tested all 22 catalysts in the range of the predicted C_2 yields higher than 16.00%. These three-times Bayesian-based validation results presented by the best C_2 yield values are shown in Fig. 3. It is readily apparent that the trends of the best C_2 yield values are moving gradually to the higher value by the validation steps. Indeed, the numbers of occurrences with the best C_2 yield value higher than 15.00% were 3, 8, and 14 catalysts, respectively, in the first (20 catalysts), second (20 catalysts), and third (22 catalysts) validation by Bayesian optimization. Therefore, Bayesian optimization investigation based on the EI index is helpful to guide the next experiment to improve the OCM performance and C_2 yield. However, the results indicated that the maximum C_2 yields were not changed at around 16.0–16.5%. In addition, La_2O_3-based catalysts are frequently found: 36 types among 62 catalysts. It can be considered that La_2O_3-based categories possess potentially high performance for OCM. Moreover, spinning the roulette wheel for selection of appropriate M1–M2–M3 components for the La_2O_3 support from the selected fields by one-hot encoding becomes a mother target for Bayesian optimization. However, it has the C_2 yield limit as its nature at around 16% under the present reaction conditions. In fact, bare La_2O_3 exhibited the highest C_2 yield of 14.0% in the present state among the bare support catalysts studied, including bare anatase-TiO_2, SiO_2, MgO, CaO, BaO, ZnO, and Y_2O_3 from the references in Table S3 in the ESI.† When further Bayesian optimization is implemented by Dataset 6 (as Bayesian-4 in Scheme 1), several components aside from Cat. No. 1–158 in Table S3 in the ESI,† and the Th element are still suggested, as well as the following 18 catalysts with C_2 yields higher than 16.00%: 17 components of CaBaLa, CaBaSm, SrBaEu, NaCaBa, MgBaEu, SrBaNd, MgKBa, CaCsBa, KSrBa, MgBaNd, BaNdNone, SrSmNone, LiCaBa, CaSnBa, CaSrEu, MgCaEu, and SrCaNd for the La_2O_3 support, and one component of KBaCe for the CaO support. The corresponding prediction lists are included in the ESI† as the **List7.csv** file. These do not appear to be attractive for additional validation because of their strong convergence toward La_2O_3-derivatives in the predictions.

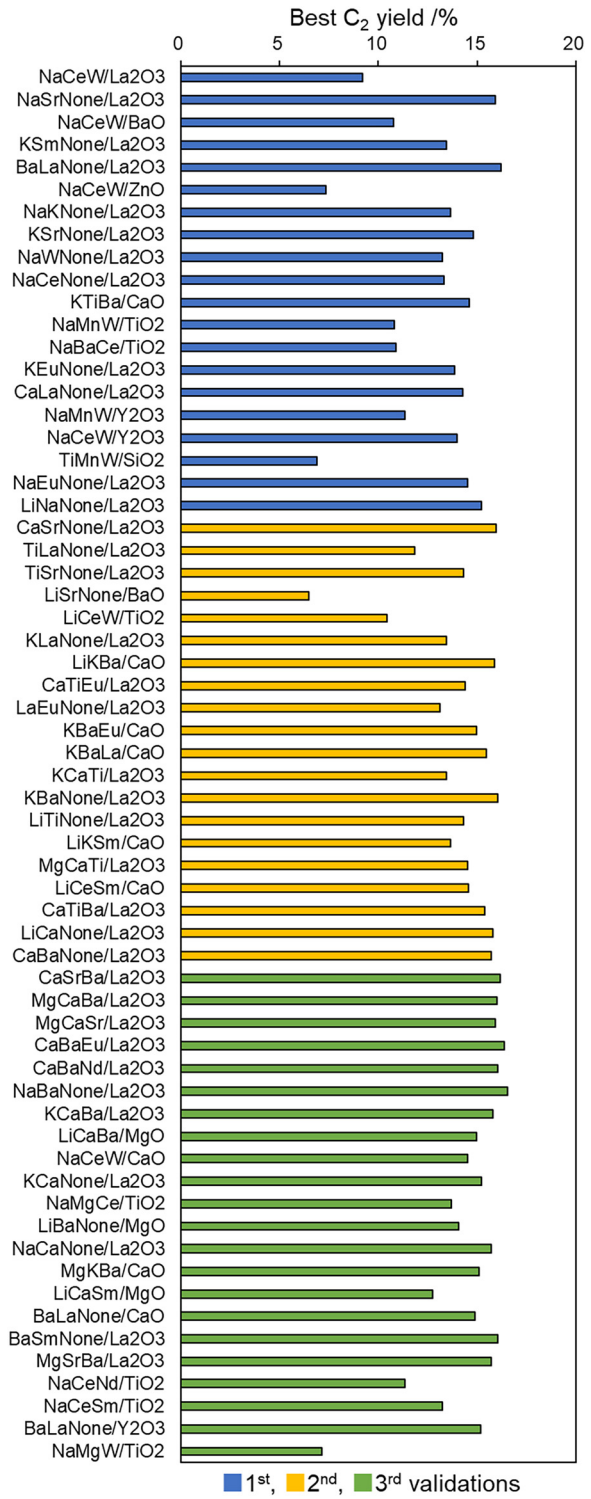


Fig. 3 Best C_2 yield plots of the first, second and third validations based on Bayesian optimization based on the EI.

4. Discussion

These results indicate that both ML-assisted catalyst investigations present some challenging issues. The first SVR has well assisted in finding OCM catalyst trends providing high C_2 yield made from Na–Mn (or Ce)–W/SiO_2-derivatives and M1–M2–none/La_2O_3 categories. The highest C_2 value in this report was obtained here by KMnW/SiO_2 (18.8%). However, the second SVR has demonstrated spatial shrinkage

in the field of C_2 prediction. Bayesian optimization based on EI provides an excellent guide for C_2 improvement. Particularly, it shows improvement of the superior La_2O_3-derived performance by M1–M2–M3 components. However, to a limited degree, it has opened a path to the extraordinary potential of catalyst design.

For the initial Dataset 1, which comprises 2842 data points, the distribution of C_2 yield values, which were rounded to the nearest integer, is shown in Fig. 4(a). The maximum value of C_2 yield in this dataset is 32.0%. However, the distributions at such high C_2 yields are not mother fields. Briefly, the C_2 yield below 8.0% includes about 50% of the data points. About 95% have occupied the range of C_2 yield below 20.0%. The mean C_2 yield was 8.9%, with a standard deviation of 6.30%. Therefore, this trend of the original dataset is one reason for the upper limit of 22.61% of the C_2 yield predicted by SVR-1. Fig. 4(b) shows the distribution of C_2 yield values, which were rounded to the nearest integer, in the additional data points based on experiment validation 1 based on SVR-1 in Dataset 2. It includes the C_2 yield in the range of 3.4% to 18.8%; the mean C_2 yield was 11.2%. It is noteworthy that approximately 85% of the points in the additional data are located at C_2 yields below 14%. One can infer that these additional data are associated with an increase in the effect of not the high but the medium C_2 yield area in the next dataset of Dataset 2, leading to difficulties in the improvement of the second prediction by SVR.

This is one characteristic of ML prediction that distinguishes it from catalyst investigations conducted based on human intuition. A human can revisit considerations along the trail and change the views of angles for the next plan to find a high-performance catalyst. In contrast, data-driven catalyst investigation requires "upper" changes close to the target performance in the dataset because it includes consideration of the trend of the base dataset for the prediction. To overcome such common issues in ML prediction, taking actions to create serendipity "proactively" is a key technology for data-driven approaches targeting exceptional performance catalysts. Furthermore, one-hot encoding, by which the numbers 0 and 1 are filled in the selected elemental indexes, was applied for this study to describe the catalyst components. It helps to reduce the complexity of the catalyst description method. In other words, the regression field cannot extend its views to other factors such as wt%, element ratios, and the preparation conditions. This lack of extension is another factor that has led to shrinkage of the SVR prediction domain at such an early stage of catalyst investigation in this study. There remains the dilemma for catalyst description between cost and accuracy, as described above. Therefore, time is necessary to discuss how to represent appropriate catalyst information for ML studies, especially for prediction. In addition, the effects of molten salts[47,48] and methyl radical ($CH_3^{\bullet}$) generation capability[49,50] in OCM at higher reaction temperatures have been discussed by *in situ* analytical techniques. How to apply such information on active states and surface/gas-phase changes based on experimental evidence would be the next subject.

Fig. 5 presents the distribution of C_2 yield values, which were rounded to the nearest integer, at additional data points based on experiment validations 3, 4, and 5, into the dataset for the subsequent Bayesian optimization processes. It is readily apparent that the additional experimentally obtained data points improved the OCM performance of the C_2 yield value gradually; the mean C_2 yield was changed from 12.6% to 13.1% and 14.3% in the data. Therefore, the Bayesian optimization certainly conducted the upper changes of the experimentally obtained data. However, as described above, La_2O_3 is recognized as an active support in OCM during the Gaussian process regression. Therefore, La_2O_3 is converged as an optimal support within the train data. The upper limit of C_2 yield was believed to be around 16% under the present reaction conditions. Compared to the bare La_2O_3 reactivity (*viz.* 14.0%), some selected M1–M2–M3 components have positive potential for C_2 yield in OCM. Therefore, it is apparent that the Bayesian optimization greatly reduces the

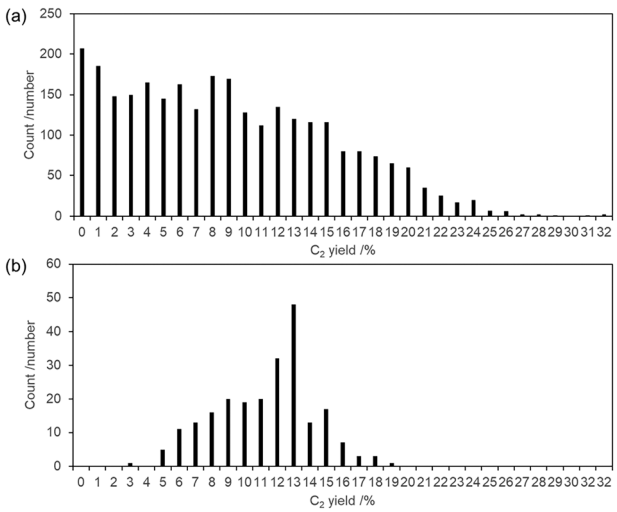


Fig. 4 Distribution of round-off C_2 yield values for (a) the initial dataset 1 consisting of 2842 data points from the literature and HTS, and (b) the additional data points based on experiment validation 1.

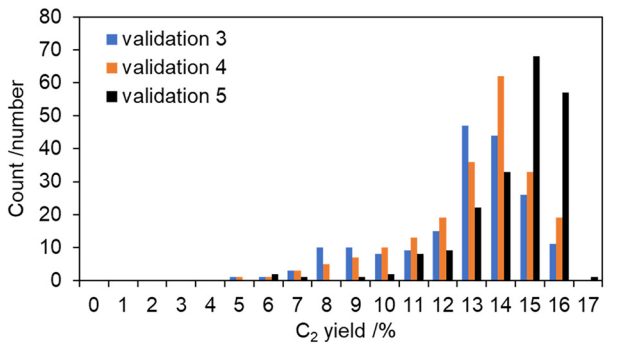


Fig. 5 Distribution of round-off C_2 yield values at additional data points based on experiment validations 3, 4 and 5, into the dataset for the subsequent Bayesian optimization processes.

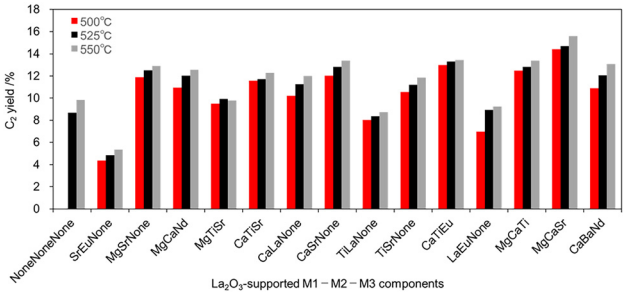


Fig. 6 Characteristic lower-temperature OCM features of La_2O_3-based M1–M2–M3 catalysts determined from the experimentally obtained data (Table S3, in the ESI†).

number of experiment trials needed to find optimal catalysts. This feature is difficult for humans in such a large survey area of element combinations as M1–M2–M3 and the support. However, this still falls short of the desired goal of data-driven catalyst investigation: the discovery of unexpected catalysts. It is noted as an important difficulty that the prediction field continues to emphasize examination of the existence of high-potential candidates such as La_2O_3-derivatives during Bayesian optimization processes. Bayesian optimization processes can benefit from incorporating a broader range of experimentally obtained data into the catalyst component. To overcome the continual turning of the roulette wheel to select the M1–M2–M3 components of the interesting support (*e.g.*, La_2O_3) from the selected fields by one-hot encoding, it is necessary to broaden the scope of support utilization. This approach can enhance the likelihood of encountering fortuitous discoveries and of achieving desirable outcomes.

From the viewpoints of the lower-temperature OCM feature, which is one attractive subject, especially for La_2O_3-based catalysts,[39,51,52] very attractive catalysts are presented in Table S3 in the ESI.† The 12 (+2) components among 48 (+5) examinations in the La_2O_3-derived catalysts are found to be positive components assisting the lower-temperature OCM based on the La_2O_3 nature in the experimentally obtained data. Actually, two appearances in five catalysts are from unpublished datasets. As shown in Fig. 6, SrEuNone (4.4%), MgSrNone (11.9%), MgCaNd (10.9%), MgTiSr (9.5%), CaTiSr (11.6%), CaLaNone (10.2%), CaSrNone (12.0%), TiLaNone (8.0%), TiSrNone (10.6%), CaTiEu (13.0%), LaEuNone (7.0%), MgCaTi (12.5%), MgCaSr (14.4%), and CaBaNd (10.9%) gave C_2 yield at 500 °C under the present conditions. The numbers in parentheses are the C_2 yields at 500 °C. Because the bare La_2O_3 was inactive at 500 °C, these 12 components were found serendipitously to be the positive compositions for activating the La_2O_3-based lower-onset temperature OCM. Another attempt is made to investigate the effects of total water production on OCM performance. It has been discussed that adding water vapor to the OCM atmosphere has both positive and negative effects on its performance.[53–55] In this study, the total water collected in a trap tube during the pre-treatment and reaction in the experiment sequence shown in Fig. S1 is also recorded in the ESI† (Table S3). Therefore, if some correlations were found between the amount of water produced and OCM reactivity, then it would be helpful for additional discussion of water effects. However, the contributions remain unclear, as shown in Fig. S3 in the ESI.† It can be inferred that the sum of water production includes a variety history in the reaction, which makes it difficult to show trends with the C_2 yield. Further considerations based on experimentally obtained data can be discussed freely as an open source *via* the CADS platform.[52]‡ The authors infer that additional opportunities exist for knowledge extraction into the next views of the OCM in a data-driven manner.

Conclusions

Support vector regression (SVR) and Bayesian optimization based on the Expected Improvement (EI) index were implemented for ternary element component-supported catalyst investigation for the oxidative coupling of methane (OCM) reaction. The dataset was compiled from the published literature and patents, including a systematic high-throughput screening (HTS) experiment, after careful verification of its accuracy by humans, one-by-one, and was updated with experimentally obtained data in validations during the progress of this study. The first trial of SVR afforded some potential OCM catalysts with C_2 yield higher than 15% under the conditions used for this study, but the second trial was unable to show improved validation results towards the higher C_2 yield field because the additional data points based on the first validation for updating use in the second validation did not include exceptional results that cannot contribute to improvement of the SVR method. The Bayesian optimizations were repeated three times using updated experimentally obtained data from validations. It is noteworthy that the Bayesian optimizations showed gradual improvement in the appearance of higher C_2 yields in data points collected from the validation experiment. However, another difficulty was observed. The prediction field continues to emphasize examination of the existence of high-potential candidates such as La_2O_3-derivatives during Bayesian optimization processes, leading to the lack of diversity of predicted materials. Although the La_2O_3-derivatives certainly show good potential for OCM, there is apparently limited performance at around a C_2 yield of 16% under the present reaction conditions. The Bayesian optimization-assisted heterogeneous catalyst investigation can access the high potential area towards the goal, but enhancement of the likelihood of encountering fortuitous discoveries is necessary to avoid spatial shrinkage in the prediction fields. Developing new active discovery algorithms to create serendipity proactively through ML is one action for

‡ Data Availability: All data generated during this study are available free of charge in the web platform Catalyst Acquisition by Data Science (CADS) for shared usage, https://cads.eng.hokudai.ac.jp.

the next challenge. From the perspectives of the different characteristics of data-driven and human-intuition-driven catalyst investigation, synergistic cooperation for the greatest benefits achieved together is another path to expand the future prospects for ML engineering use.

Author contributions

SN and KT conducted most experiments, analyzed the data, and composed the manuscript. XL prepared the catalysts whereas JO deeply participated in discussions of catalyst performance evaluation and data analysis manner. KT performed data pre-treatment of original datasets and the ML analysis. All authors contributed to the conceptual design of this study, discussion of the results, and review of the manuscript, and have approved the final version of the manuscript.

Conflicts of interest

The authors declare that they have no competing financial interest.

Acknowledgements

This work is supported by the Japan Science and Technology Agency (JST) CREST (grant no. JPMJCR17P2).

Notes and references

1 J. Bures and I. Larrosa, *Nature*, 2023, **613**, 689–695.
2 B. MacQueen, R. Jayarathna and J. Lauterbach, *Curr. Opin. Chem. Eng.*, 2022, **36**, 100781.
3 B. R. Goldmith, J. Esterhuizen, J. Liu, C. J. Bartel and C. Sutton, *AIChE J.*, 2018, **64**, 2311–2323.
4 C. B. Santiago, J. Guo and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 2398–2412.
5 T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K. Shimizu, *ACS Catal.*, 2020, **10**, 2260–2297.
6 K. Takahashi, J. Ohyama, S. Nishimura, J. Fujima, L. Takahashi, T. Uno and T. Taniike, *Chem. Commun.*, 2023, **59**, 2222–2238.
7 U. Zavyalova, M. Holena, R. Schlogl and M. Baerns, *ChemCatChem*, 2011, **3**, 1935–1947.
8 E. V. Kondratenko, T. Peppel, D. Seeburg, V. A. Kondratenko, N. Kalevaru, A. Martin and S. Wohlrab, *Catal. Sci. Technol.*, 2017, **7**, 366–381.
9 R. Schmack, A. Friedrich, E. V. Kondratenko, J. Polte, A. Werwatz and R. Kraehnert, *Nat. Commun.*, 2019, **10**, 441.
10 K. Takahashi, I. Miyazato, S. Nishimura and J. Ohyama, *ChemCatChem*, 2018, **10**, 3223–3228.
11 S. Mine, M. Takao, T. Yamaguchi, T. Toyao, Z. Maeno, S. M. A. H. Siddiki, S. Takakusagi, K. Shimizu and I. Takigawa, *ChemCatChem*, 2021, **13**, 3636–3655.
12 X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist and J. Schrier, *Nature*, 2019, **573**, 251–255.
13 S. Nishimura, J. Ohyama, T. Kinoshita, S. D. Le and K. Takahashi, *ChemCatChem*, 2020, **12**, 5888–5892.
14 F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen and F. Glorius, *Angew. Chem.*, 2022, **61**, e202204647.
15 A. Lazaridou, L. R. Smith, S. Pattisson, N. F. Dummer, J. J. Smit, P. Johnston and G. J. Hutchings, *Nat. Rev. Chem.*, 2023, **7**, 287–295.
16 J. J. Hanak, *J. Mater. Sci.*, 1970, **5**, 964–971.
17 L. Olivier, S. Haag, H. Pennemann, C. Hofmann, C. Mirodatos and A. C. van Veen, *Catal. Today*, 2008, **137**, 80–89.
18 Z. Aydin, A. Zanina, V. A. Kondratenko, J. Rabeah, J. Li, J. Chen, Y. Li, G. Jiang, H. Lund, S. Bartling, D. Linke and E. V. Kondratenko, *ACS Catal.*, 2022, **12**, 1298–1309.
19 R. J. Hendershot, P. T. Fanson, C. M. Snively and J. A. Lauterbach, *Angew. Chem., Int. Ed.*, 2003, **42**, 1152–1155.
20 T. N. Nguyen, T. T. P. Nhat, K. Takimoto, A. Thakur, S. Nishimura, J. Ohyama, I. Miyazato, L. Takahashi, J. Fujima, K. Takahashi and T. Taniike, *ACS Catal.*, 2020, **10**, 921–932.
21 E. V. Kondratenko, M. Schluter, M. Baerns, D. Linke and M. Holena, *Catal. Sci. Technol.*, 2015, **5**, 1668–1677.
22 C. Hoffmann, H. Schmidt and F. Schüth, *J. Catal.*, 2001, **198**, 348–354.
23 C. Ortega, D. Otyuskaya, E. Ras, L. D. Virla, G. S. Patience and H. Dathe, *Can. J. Chem. Eng.*, 2021, **99**, 1288–1306.
24 Y. Shinke, T. Miyazawa, M. Hiza, I. Nakamura and T. Fujitani, *React. Chem. Eng.*, 2021, **6**, 1381–1385.
25 G. Rothenberg, *Catal. Today*, 2008, **137**, 2–10.
26 T. Taniike and K. Takahashi, *Nat. Catal.*, 2023, **6**, 108–111.
27 A. Takagaki, Y. Tsuji, T. Yamasaki, S. Kim, T. Shishido, T. Ishihara and K. Yoshizawa, *Chem. Commun.*, 2023, **59**, 286–289.
28 Y. Tsuji and K. Yoshizawa, *J. Phys. Chem. C*, 2018, **122**, 15359–15381.
29 K. Takahashi, L. Takahashi, S. D. Le, T. Kinoshita, S. Nishimura and J. Ohyama, *J. Am. Chem. Soc.*, 2022, **144**, 15735–15744.
30 D. Roy, S. C. Mandal and B. Pathak, *ACS Appl. Mater. Interfaces*, 2021, **13**, 56151–56163.
31 K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, J. Ohyama, T. N. Nguyen, S. Nishimura and T. Taniike, *ChemCatChem*, 2019, **11**, 1146–1152.
32 T. Ishida, T. Murayama, A. Taketoshi and M. Haruta, *Chem. Rev.*, 2020, **120**, 464–525.
33 G. E. Keller and M. M. Bhasin, *J. Catal.*, 1982, **73**, 9–19.
34 W. Hinsen and M. Baerns, *Chem.-Ztg.*, 1983, **107**, 223–226.
35 A. H. Tullo, *Chem. Eng. News*, 2014, **92**, 20–21.
36 Y. Gao, L. Neal, D. Ding, W. Wu, C. Baroi, A. M. Gaffney and F. Li, *ACS Catal.*, 2019, **9**, 8592–8621.
37 B. L. Farrell, V. O. Igenegbai and S. Linic, *ACS Catal.*, 2016, **6**, 4340–4346.
38 T. N. Nguyen, S. Nakanowatari, T. P. N. Tran, A. Thakur, L. Takahashi, K. Takahashi and T. Taniike, *ACS Catal.*, 2021, **11**, 1797–1809.

39 S. Nishimura, S. D. Le, I. Miyazato, J. Fujima, T. Taniike, J. Ohyama and K. Takahashi, *Catal. Sci. Technol.*, 2022, **12**, 2766–2774.
40 S. Nishimura, J. Ohyama, X. Li, I. Miyazato, T. Taniike and K. Takahashi, *Ind. Eng. Chem. Res.*, 2022, **61**, 8462–8469.
41 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher and M. P. E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
42 J. Ohyama, Y. Tsuchimura, H. Yoshida, M. Machida, S. Nishimura and K. Takahashi, *J. Phys. Chem. C*, 2022, **126**, 19660–19666.
43 J. Ohyama, S. Nishimura and K. Takahashi, *ChemCatChem*, 2019, **11**, 4307–4313.
44 H. R. Godini, A. Gili, O. Görke, S. Arndt, U. Simon, A. Thomas, R. Schomäcker and G. Wozny, *Catal. Today*, 2014, **236**, 12–22.
45 A. Cruellas, T. Melchiori, F. Gallucci and M. van Sint-Annala, *Catal. Rev.: Sci. Eng.*, 2017, **59**, 234–294.
46 L. A. Vandewalle, R. Van de Vijver, K. M. Van Geem and G. B. Marin, *Chem. Eng. Sci.*, 2019, **198**, 268–289.
47 S. Sourav, Y. Wang, D. Kiani, J. Baltrusaitis, R. R. Fushimi and I. E. Wachs, *Angew. Chem.*, 2021, **60**, 21502.
48 K. Takanabe, A. M. Khan, Y. Tang, L. Nguyen, A. Ziani, B. W. Jacobs, A. M. Elbaz, S. M. Sarathy and F. F. Tao, *Angew. Chem., Int. Ed.*, 2017, **56**, 10403.
49 Q. Zhou, Z. Wang, Z. Li, J. Wang, M. Xu, S. Zou, J. Yang, Y. Pan, X. Gong, L. Xiao and J. Fan, *ACS Catal.*, 2021, **11**, 14651.
50 L. Luo, X. Tang, W. Wang, Y. Wang, S. Sun, F. Qi and W. Huang, *Sci. Rep.*, 2013, **3**, 1625.
51 J. Ohyama, T. Kinoshita, E. Funada, H. Yoshida, M. Machida, S. Nishimura, T. Uno, J. Fujima, I. Miyazato, L. Takahashi and K. Takahashi, *Catal. Sci. Technol.*, 2021, **11**, 524–530.
52 S. Nishimura, High-Throughput Screening and Literature Data Driven Machnine Learning Assistaing Discovery of La_2O_3-based Catalysts for Low-Temperature Oxidative Coupling of Methane, *Proceedings of the 31st Annual Saudi-Japan Symposium on Technology in Fuels & Petrochemicals – Innovative Catalyst Development*, KFUPM, Dhahran, Saudi Arabia, 2022, pp. 32–42.
53 H. Wang, C. Yang, C. Shao, S. Alturkistani, G. Magnotti, J. Gascon, K. Takanabe and S. M. Sarathy, *ChemCatChem*, 2022, **14**, e202200927.
54 V. I. Lomonosov and M. Y. Sinev, *Kinet. Catal.*, 2016, **57**, 647–676.
55 K. Takanabe and E. Iglesia, *Angew. Chem., Int. Ed.*, 2008, **47**, 7689–7693.