Catalysis Science & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6cy00228e

Importance of consistent and well-dispersed catalyst datasets for machine learning in oxidative coupling of methane

Hinata Sudo, [a] Yoshiki Hasukawa, [a] Rensuke Koiwai, [a] Fernando Garcia-Escobar, [a] Shun Nishimura, [b] Lauren Takahashi *[a] and Keisuke Takahashi *[ac]

Received 23rd February 2026,
Accepted 7th April 2026

DOI: 10.1039/d6cy00228e

rsc.li/catalysis

The role of highly uniform, diverse experimental data in catalyst informatics is examined using an oxidative coupling of methane dataset measured by a single researcher under consistent devices and conditions. Broad compositional coverage and minimized experimental variability enable machine learning to capture composition–performance relationships using simple one-hot encoding. Inverse analysis of the compositional space identifies promising catalysts for experimental validation. These results demonstrate that carefully curated, well-distributed datasets, even if relatively small, enable machine learning to effectively capture composition–performance relationships.

Introduction

Catalyst informatics aims to accelerate catalyst design by identifying trends and patterns embedded in catalyst datasets.[1,2] Among the available approaches, machine learning has emerged as a powerful tool for extracting structure–performance relationships and guiding catalyst discovery, provided that sufficiently large and informative datasets are available.[3–7] However, the application of machine learning to heterogeneous catalysis faces several fundamental challenges. The first is the limited availability and quality of catalyst data. This issue has been addressed through the compilation of literature datasets, the use of high-throughput experimentation to generate large-scale catalyst data, or the development of methods tailored to small datasets.[8–11] Each strategy, however, has inherent limitations. Literature-derived datasets are often poorly dispersed, as researchers preferentially report successful results, while unsuccessful or so-called negative data are rarely disclosed. This lack of data diversity is a bottleneck not only for methane activation, but also across the broader landscape of carbon utilization, where high-quality data are essential for optimizing carbon efficiency in various thermochemical processes.[12] Furthermore, as the field matures, machine-learning workflows are increasingly being integrated with techno-economic and life cycle assessments to ensure that catalyst discovery aligns with broader sustainability goals.[13] Under these circumstances, high-throughput experimentation can mitigate both data scarcity and poor dispersion by systematically exploring broad catalyst spaces.[14,15] Nevertheless, high-throughput approaches also have intrinsic drawbacks, as catalysts are typically evaluated under uniform experimental conditions, which may obscure catalyst-specific optimal operating windows and mask intrinsic structure–performance relationships. A second critical challenge lies in the definition of catalyst descriptors for supervised machine learning. The construction of effective catalyst descriptors remains a long-standing problem, as there is no universally accepted representation of heterogeneous catalysts.[16,17] One-hot encoding, a commonly used representation, is particularly problematic for poorly dispersed or unevenly distributed datasets, as it treats chemically related compositions as independent categories and prevents the model from learning meaningful trends for sparsely sampled catalysts. An alternative strategy is to represent catalysts using physically meaningful quantities derived from periodic table properties, which has been shown to improve both predictive performance and interpretability.[18] Consequently, both the quality of catalyst datasets and the choice of catalyst descriptors are central to the successful application of machine learning in catalyst informatics.

In this work, the impact of uniformly generated and well-dispersed catalyst datasets produced by a single researcher and same experimental devices and environment is systematically investigated to elucidate how such data quality influences machine-learning performance. The oxidative

[a] Department of Chemistry, Hokkaido University, North 10, West 8, Sapporo 060-0810, Japan. E-mail: lauren.takahashi@sci.hokudai.ac.jp, keisuke.takahashi@sci.hokudai.ac.jp
[b] Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan
[c] List Sustainable Digital Transformation Catalyst Collaboration Research Platform, Institute for Chemical Reaction Design and Discovery, Hokkaido University, Sapporo 001-0021, Japan

coupling of methane (OCM) reaction is selected as a prototypical system. OCM aims at the direct conversion of methane into C_2 hydrocarbons, primarily C_2H_4 and C_2H_6.[19–21] The dataset used in this study is derived from previous work in which all catalysts are evaluated by a single researcher under strictly controlled experimental conditions, including identical experimental setups and standardized operating procedures.[22] By minimizing experimental variability while maintaining broad compositional diversity, this dataset provides an ideal platform for catalyst informatics. Using this dataset, comprehensive data analysis and supervised machine-learning approaches are applied to uncover structure–performance relationships and to guide catalyst design for the OCM reaction.

Data and methods

Published OCM data are used in this work.[22] The dataset is generated by a single researcher using identical experimental devices and a consistent experimental environment. It consists of 2928 data points, including non-catalytic conditions, support-only catalysts, and binary and ternary oxide catalysts. In detail, the dataset includes 25 elements as active catalyst components (Li, Na, Mg, K, Ca, Ti, Mn, Zn, Rb, Sr, Y, Zr, Mo, Sn, Cs, Ba, La, Ce, Nd, Sm, Eu, Hf, W, Pb and Bi) and 8 oxides as supports (BaO, CaO, MgO, SiO_2, TiO_2, ZnO, Y_2O_3 and La_2O_3). Reaction temperature, the CH_4/O_2 gas flow ratio and furnace length are also included as experimental variables. The C_2 yield, defined as the sum of C_2H_4 and C_2H_6 yields, is used as the objective variable. Unless otherwise stated, the CH_4/O_2 ratio is fixed at 3 : 1, and the furnace length is fixed at 270 mm. Furthermore, for comparison, preprocessed literature data are also ultilized.[22] The physical quantities used for the catalyst descriptors are from the XenonPy library.[23]

Data preprocessing is performed prior to data analysis and machine learning. Data points exhibiting negative O_2 or CH_4 conversion values are removed. In addition, data points with selectivities exceeding 100% for H_2, CO, CO_2, C_2H_4, or C_2H_6 are excluded due to physical inconsistency and experimental noise. To eliminate the influence of varying experimental conditions and to focus on catalyst and support effects, data collected at 700 °C with a CH_4/O_2 ratio of 3 : 1 and a furnace length of 270 mm are extracted for machine-learning analysis. Furthermore, only binary and ternary catalyst compositions are retained. As a result, the number of data points is reduced to 2124. The catalyst composition and support are represented using one-hot encoding.

Supervised machine learning is performed using random forest regression (RFR) implemented in the scikit-learn.[24] The random state is fixed, and the number of trees is set to 100. Model performance is evaluated by cross-validation, in which the dataset is randomly split into 80% training and 20% test sets. The reported performance corresponds to the average R^2 score of the test data obtained from 10 independent train–test splits.

Catalyst synthesis

The catalysts listed in Table 1 are prepared by the impregnation method. Aqueous solutions containing metal precursors are impregnated onto a support, followed by drying and calcination to disperse the metal species on the support surface. Lanthanum oxide (La_2O_3, ≥99.5%, Junsei Chemical Co., Ltd.) is used as the catalyst support. Sodium nitrate ($NaNO_3$, ≥99.0%), magnesium nitrate hexahydrate ($Mg(NO_3)_2{\cdot}6H_2O$, ≥99.9%), calcium nitrate tetrahydrate ($Ca(NO_3)_2{\cdot}4H_2O$, ≥98.5%), barium nitrate ($Ba(NO_3)_2$, ≥99.0%), and lanthanum nitrate hexahydrate ($La(NO_3)_3{\cdot}6H_2O$) (FUJIFILM Wako Pure Chemical Corporation) are used as metal precursors.

For catalyst preparation, 2 g of La_2O_3 is added to 100 mL of deionized water. Metal nitrate salts are dissolved in 50 mL of deionized water so that the total molar fraction of the added metals is 3%. The metal precursor solution is added to the La_2O_3 suspension in the order M1, M2, and M3 at 5 min intervals under stirring, followed by stirring for 60 min and aging overnight at room temperature. The mixture is then heated under stirring to remove water. After drying at 80 °C for 8 h, the obtained solid is ground into a fine powder and calcined at 800 °C for 3 h, with heating and cooling rates of 800 °C h^{-1}. A reference catalyst consisting of La_2O_3 only is prepared following the same procedure.

Catalyst testing

Oxidative coupling of methane is carried out in a continuous-flow fixed-bed reactor. An alumina tubular reactor (length: 370 mm, inner diameter: 6.0 mm) is used, and 75 mg of catalyst is placed between two layers of quartz wool (1 mg each). The reactor temperature is monitored using a K-type thermocouple located near the catalyst bed. Before the reaction, the reactor is purged with N_2 at 200 °C for 30 min and then cooled to room temperature.

Catalytic performance is evaluated using a $CH_4/O_2/N_2$ gas mixture with flow rates of 8.0/4.0/16.0 mL min^{-1} at 600, 650, 675, 700, 725, 750, 800, and 850 °C. At each temperature, the reaction is conducted for 10 min before gas sampling. Reaction products are analyzed using a Shimadzu GC-2014 gas chromatograph equipped with a SHINCARBON ST 50/80 column. Conversions of CH_4 and O_2, as well as yields of CO, CO_2, C_2H_4, C_2H_6, and C_2 products and C_2 selectivity, are calculated using N_2 as an internal standard according to eqn (1)–(4). In eqn (1) and (2), in and out represent the inlet feed and outlet effluent streams, respectively, which are used to calculate the conversion of reagents (R) such as CH_4 and O_2. In eqn (2), n is equal to 1 for CO and CO_2 and 2 for C_2H_4 and C_2H_6.

Table 1 Materials used for each catalyst

Catalyst	Support	M1	M2	M3
$NaBaMg/La_2O_3$	La_2O_3	$NaNO_3$	$Mg(NO_3)_2{\cdot}6H_2O$	$Ba(NO_3)_2$
$NaCaBa/La_2O_3$	La_2O_3	$NaNO_3$	$Ca(NO_3)_2{\cdot}4H_2O$	$Ba(NO_3)_2$
$NaBaLa/La_2O_3$	La_2O_3	$NaNO_3$	$Ba(NO_3)_2$	$La(NO_3)_3{\cdot}6H_2O$

$$R_{\mathrm{Conv}} = \frac{(R_{\mathrm{In}}/N_{2\mathrm{In}}) - (R_{\mathrm{Out}}/N_{2\mathrm{Out}})}{(R_{\mathrm{In}}/N_{2\mathrm{In}})} \times 100 \quad (1)$$

$$P_{\mathrm{Yield}}\ (\%) = \frac{n \times (P_{\mathrm{Out}}/N_{2\mathrm{Out}})}{(\mathrm{CH}_{4\mathrm{In}}/N_{2\mathrm{In}})} \times 100 \quad (2)$$

$$C_{2\mathrm{yield}} = \mathrm{C_2H_{4Yield}} + \mathrm{C_2H_{6Yield}} \quad (3)$$

$$C_{2\mathrm{selectivity}} = \frac{C_{2\mathrm{yield}}}{\mathrm{CH_{4Conv}}} \times 100 \quad (4)$$

Results and discussion

OCM data are statistically analyzed. In particular, the distributions of catalyst elements, supports, and C_2 yield are visualized in Fig. 1. Fig. 1 demonstrates that the catalyst composition and support materials are broadly and uniformly distributed across the dataset, while the C_2 yield spans from 0% to high-yield regimes. Such wide coverage in both chemical composition and performance spaces is essential for machine-learning modeling, as it minimizes sampling bias, reduces extrapolation risk, and enables robust learning of composition–performance relationships. The well-dispersed nature of the dataset ensures that the model is trained on diverse catalytic environments, thereby improving generalizability and predictive reliability.

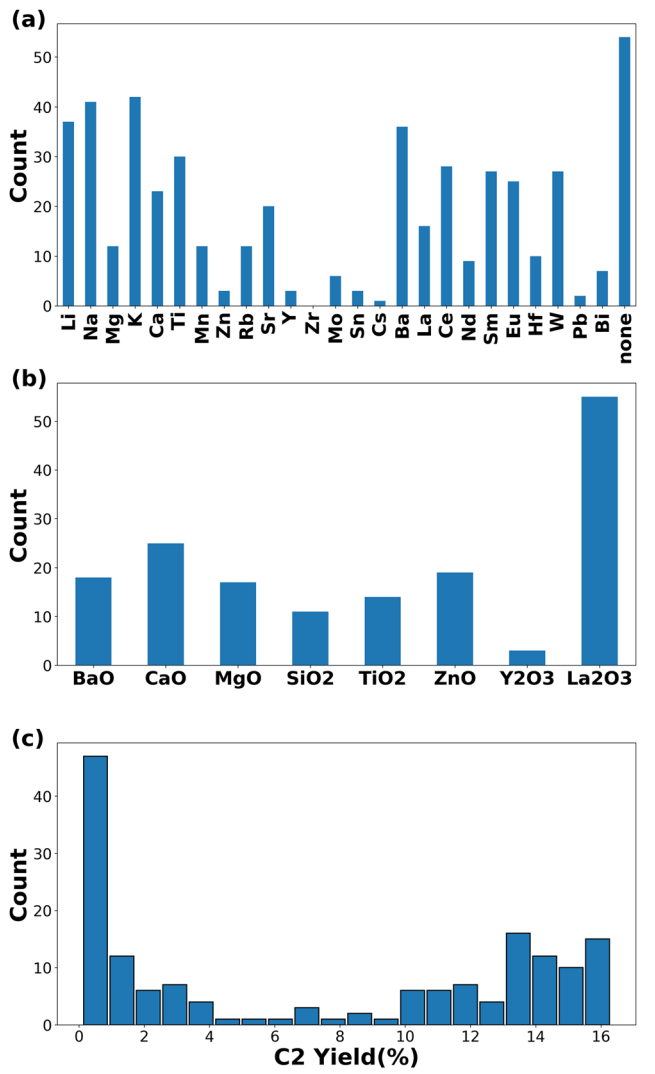

Fig. 1 Frequent data distribution of (a) catalyst elements, (b) supports and (c) C_2 yield.

As shown in Fig. 2(a), a pairwise correlation map is constructed between CH_4 conversion, C_2 yield, and selectivity, and the presence of catalyst elements and supports at 700 °C. Additionally, the feature importance of these catalyst elements and supports toward C_2 yield is evaluated. Fig. 2(a) indicates that the La_2O_3 support exhibits a positive correlation with both CH_4 conversion and C_2 yield, suggesting that La_2O_3 is an active support under OCM reaction conditions. In a similar manner, certain alkaline earth elements, such as Sr and Ca, show weak but positive correlations with C_2 yield, implying a potential promoting effect on C_2 formation. Note that while Fig. 2(b) and (c) present the feature importance obtained from the random forest model, a high importance score does not necessarily imply a positive impact on the target variable. Random forest importance analysis of elements and supports is also performed for C_2 yield as shown in Fig. 2(b) and (c), respectively. Fig. 2(b) reveals that Ca has the most predominant importance for C_2 yield, followed by other alkaline earth metals such as Sr and Ba. Combining these results with the pairwise correlations in Fig. 2(a), it can be inferred that these specific alkaline earth elements exert a beneficial effect on C_2 formation. Furthermore, although W also exhibits relatively high importance, its negative correlation in Fig. 2(a) suggests a deleterious effect on C_2 yield. Similarly, Fig. 2(c) shows that La_2O_3 possesses the highest importance, which, consistent with Fig. 2(a), indicates its significant role in enhancing catalytic performance. While MgO also shows relatively high importance and a certain trend in Fig. 2(a), the underlying factors behind its contribution are discussed in detail in the following section in conjunction with the results shown in Fig. 3.

To evaluate the overall influence of catalyst elements and supports on C_2 yield, violin plots are constructed. The distributions of C_2 yield as a function of individual elements and supports are shown in Fig. 3. Fig. 3 indicates that catalysts containing Sr and Ca tend to exhibit higher C_2 yields, consistent with the positive correlations observed in the pairwise correlation map as shown in Fig. 2. In addition, elements such as La and Mg are associated with relatively high C_2 yields, whereas Cs, Bi, and Pb are generally linked to lower C_2 yields. Notably, the interpretation is nontrivial because catalytic performance strongly depends on elemental combinations. As shown in Fig. 3(a), elements such as K, Na, Mn, and W are associated with both high and low C_2 yields, depending on their pairing with other elements. This highlights the importance of combination effects rather than single-element contributions. Similarly, Fig. 3(b) shows that

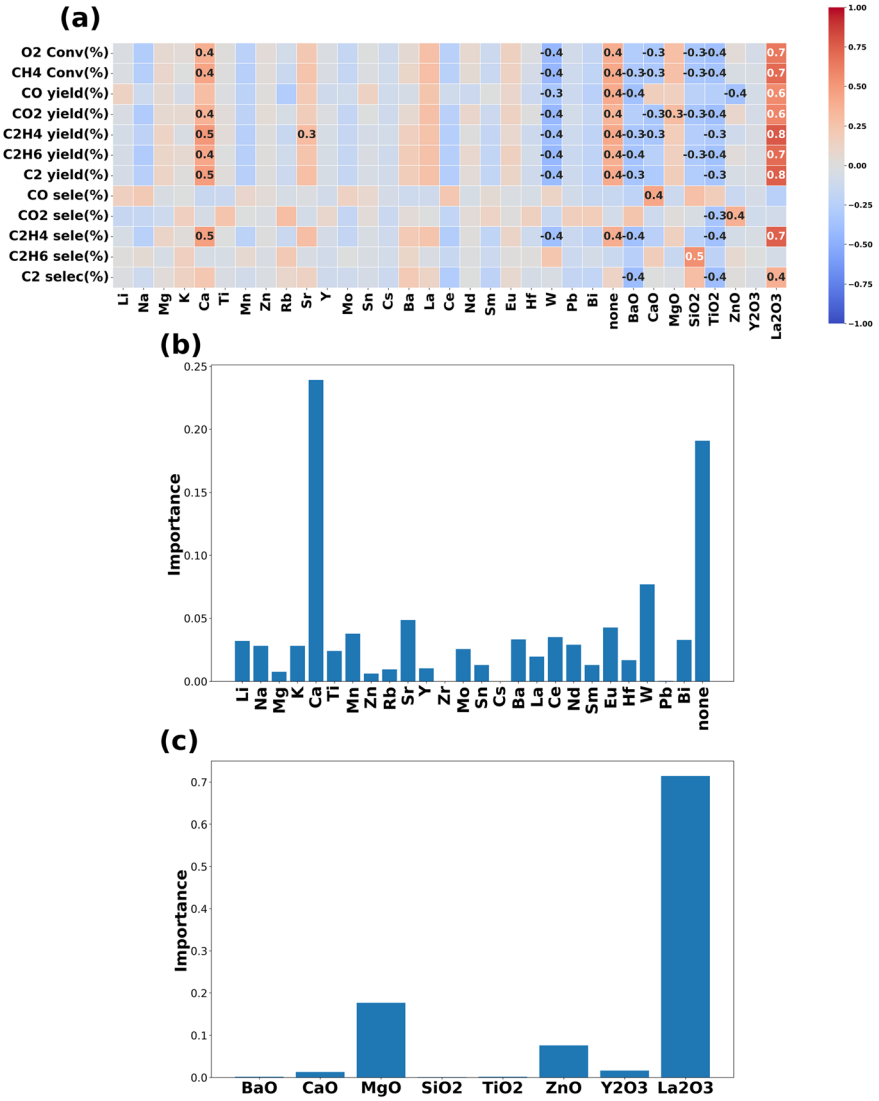

Fig. 2 (a) Pairwise correlation of conversion, yield and selectivity against elements and supports at 700 °C. (b) Random forest feature importance of elements at 700 °C. (c) Random forest feature importance of supports for C_2 yield at 700 °C.

La_2O_3 supports tend to result in higher C_2 yields, consistent with the analysis in Fig. 2. The catalytic performance may be attributed to the formation of subsurface peroxide species acting as active oxygen centers, which remain stable even at high temperatures.[25] According to microkinetic analysis, these stable surface oxygen sites are suggested to promote methane dissociation and subsequent methyl radical formation, thereby facilitating gas-phase C_2 coupling reactions.[26] In a similar manner, Fig. 3(b) also indicates that MgO supports lead to enhanced C_2 yields, which supports the trends observed in Fig. 2(a) and (c). This performance may be attributed to the presence of surface defects, particularly steps, which serve as active centers for the oxidative activation of methane. Kinetic studies suggest that these step sites facilitate a surface-mediated coupling process, contributing to high initial C_2 selectivity.[27] However, the support effect is also strongly coupled with the choice of active elements, as support performance varies substantially depending on elemental pairing.

Machine learning modeling is performed to predict highly active OCM catalysts. One-hot encoded representations of catalyst elements and supports are used as descriptor variables, while the objective variable is set to the C_2 yield. Because experimental conditions strongly affect catalytic performance and can obscure composition performance relationships in machine learning analysis, the reaction temperature and CH_4/O_2 ratio are fixed at 700 °C and 3, respectively. The comparison between predicted and experimentally measured C_2 yields is shown in Fig. 4(a). The model achieves a cross-validated coefficient of determination (R^2) of 0.74, indicating good predictive performance under these constrained conditions. For comparison, 58 physical quantities in XenonPy are used, which result in an R^2 of 0.74 as shown in Fig. 4(b); however, the MAE is a slightly better score in the one hot encoding case. Furthermore, literature

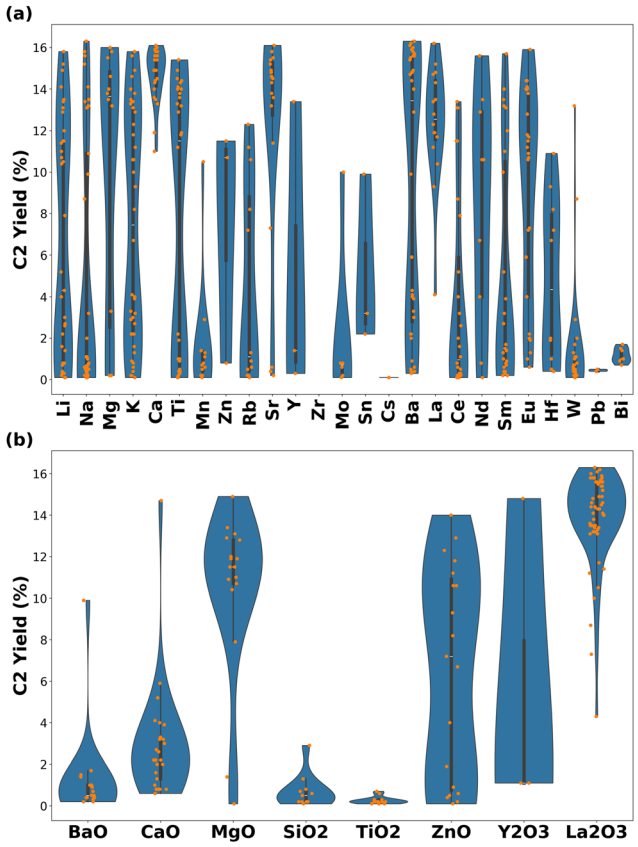

Fig. 3 Violin plot of C_2 yield against each (a) element and (b) support.

data are also evaluated with one hot encoding where the temperature and CH_4/O_2 ratio are fixed at 700 °C and 3, respectively, and collected in Fig. 4(c). Fig. 4(c) shows that inconsistent data result in poor machine learning performance, and thus consistent data are quite important. For comparison, 58 physical quantities in XenonPy are used, which result in an R^2 of 0.74 as shown in Fig. 4(b); however, the MAE is a slightly better score in the one-hot encoding case. Furthermore, literature data are also evaluated with one-hot encoding where the temperature and CH_4/O_2 ratio are fixed at 700 °C and 3, respectively, and collected in Fig. 4(c). Fig. 4(c) shows that inconsistent data result in poor machine learning performance, and thus consistent data are quite important. It should be noted that the well-diverse dataset, generated by a single researcher under identical devices and conditions, carries rich information about the catalysts, as the broad coverage of compositional space ensures that even simple one-hot encoding enables the model to extract relationships between catalyst composition and performance directly from the data. This work demonstrates that the quality and consistency of the dataset are critical. Collecting such data is labor intensive but reduces experimental noise and variability. Carefully curated, well-dispersed datasets can be more effective for machine learning than larger but heterogeneous datasets compiled from multiple sources.

An inverse analysis is performed to identify promising OCM catalysts. A total of 20 800 hypothetical binary and ternary catalyst combinations are generated from 25 elements (Na, Li, Mn, K, Sr, La, Ti, Ca, Ba, Mg, Rb, Y, Sm, Ce, Zn, Mo, Zr, Eu, Cs, Nd, Sn, W, Bi, Hf, and Pb) and 8 supports (BaO, CaO, La_2O_3, MgO, SiO_2, TiO_2, Y_2O_3, and ZnO). Each element–support combination is converted into a one-hot encoded representation and used as input for a trained random forest regression (RFR) model. The model is then applied to screen the full compositional space, and the top 3 predicted high C_2 yield catalysts are summarized in Table 2.

Based on machine learning predictions, three catalyst compositions, Na–Ba–Mg/La_2O_3, Na–Ca–Ba/La_2O_3, and Na–Ba–La/La_2O_3, as shown in Table 2, are selected for experimental validation. The catalytic performances of these catalysts, together with La_2O_3 as a reference, are shown in Fig. 5. All experiments are independently repeated twice to confirm reproducibility. As shown in Fig. 5, all three machine-learning-guided catalysts exhibit substantially higher C_2 yields than the La_2O_3 reference, demonstrating the effectiveness of the inverse design strategy. The maximum C_2 yields achieved by each catalyst are 19.2% at 750 °C for Na–

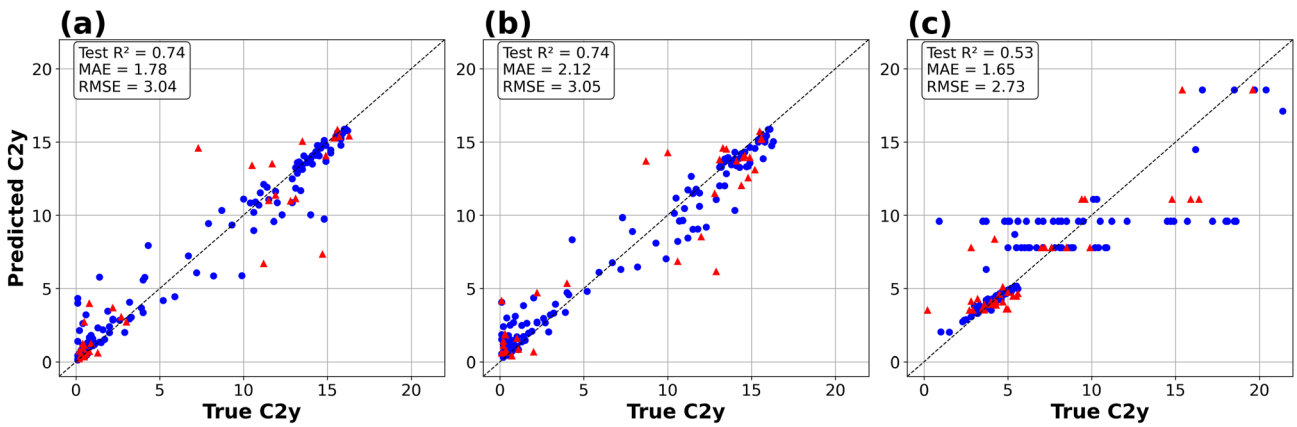

Fig. 4 True and predicted C_2 yield using (a) one hot encoding descriptors, (b) physical quantity descriptors, and (c) one hot encoding with literature data.

Table 2 Top 3 predicted and unreported catalyst compositions and predicted with high C_2 yield

Support	M1	M2	M3	Predicted C_2y
La_2O_3	Na	Ba	Mg	15.914
La_2O_3	Na	Ca	Ba	15.896
La_2O_3	Na	Ba	La	15.869

Ba–Mg/La_2O_3, 19.2% at 750 °C for Na–Ca–Ba/La_2O_3, and 18.6% at 725 °C for Na–Ba–La/La_2O_3, respectively.

Conclusion

In summary, the role of data quality in catalyst informatics is systematically examined using a uniformly generated OCM

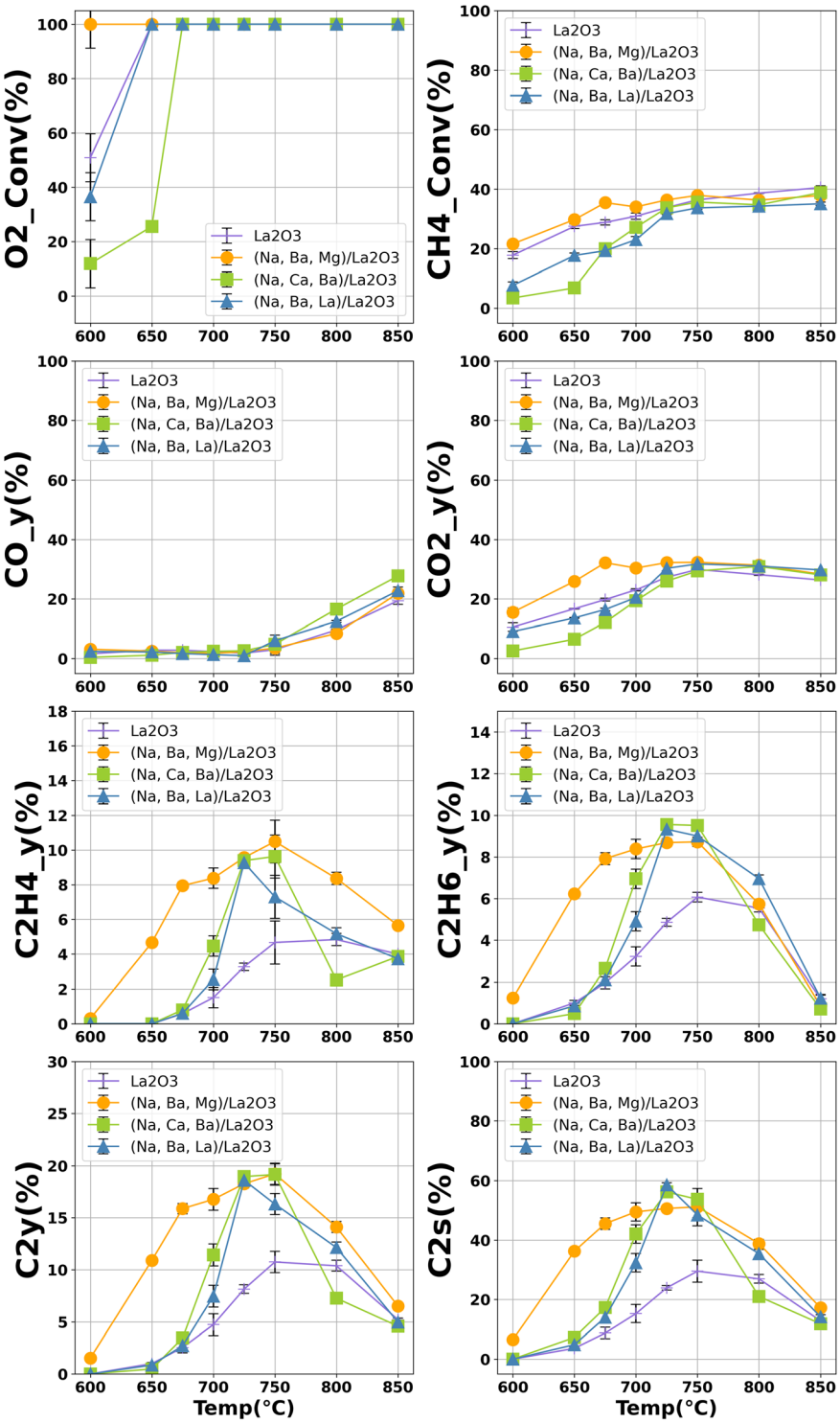

Fig. 5 Conversion and yield of Na–Ba–Mg–La_2O_3, Na–Ca–Ba–La_2O_3, Na–Ba–La–La_2O_3 and La_2O_3.

catalyst dataset. Statistical analysis confirms that catalyst elements, supports, and C_2 yield are well dispersed, providing a favorable foundation for machine learning. When experimental conditions are fixed, supervised learning using simple one-hot encoding successfully captures composition–performance relationships, achieving reliable predictive accuracy. Inverse analysis further enables efficient screening of a large compositional space, leading to the experimental validation of previously unreported high-performance OCM catalysts. Although generating these datasets by a single researcher is labor-intensive, the resulting consistency and broad compositional coverage substantially enhance the effectiveness of data-driven catalyst design. This work demonstrates that carefully curated, well-dispersed experimental datasets can enable meaningful machine learning even without large-scale high-throughput experiments, highlighting data quality as a key factor in advancing catalyst informatics.

Conflicts of interest

There are no conflicts to declare.

Data availability

All data are included in the manuscript.

Acknowledgements

This work was funded by the Japan Science and Technology Agency (JST) ERATO Grant Number (JPMJER1903), PRESTO Grant Number (JPMJPR24T5), the JST Mirai Program Grant Number (JP-MJMI25G1), and the JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Numbers (JP23H01762) and (24K01241).

References

1 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, The rise of catalyst informatics: towards catalyst genomics, *ChemCatChem*, 2019, **11**, 1146–1152.

2 K. Takahashi, J. Ohyama, S. Nishimura, J. Fujima, L. Takahashi, T. Uno and T. Taniike, Catalysts informatics: paradigm shift towards data-driven catalyst design, *Chem. Commun.*, 2023, **59**, 2222–2238.

3 J. R. Kitchin, Machine learning in catalysis, *Nat. Catal.*, 2018, **1**, 230–232.

4 W. Yang, T. T. Fidelis and W.-H. Sun, Machine learning in catalysis, from proposal to practicing, *ACS Omega*, 2019, **5**, 83–88.

5 P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen and T. Bligaard, Machine learning for computational heterogeneous catalysis, *ChemCatChem*, 2019, **11**, 3581–3601.

6 T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K.-I. Shimizu, Machine learning for catalysis informatics: recent applications and prospects, *ACS Catal.*, 2019, **10**, 2260–2297.

7 J. A. Esterhuizen, B. R. Goldsmith and S. Linic, Interpretable machine learning for knowledge generation in heterogeneous catalysis, *Nat. Catal.*, 2022, **5**, 175–184.

8 J. S. Hummelshoj, F. Abild-Pedersen, F. Studt, T. Bligaard and J. K. Norskov, Cat-App: a web application for surface chemistry and heterogeneous catalysis, *Angew. Chem., Int. Ed.*, 2012, **51**, 272–274.

9 A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders and R. Fushimi, Extracting knowledge from data through catalysis informatics, *ACS Catal.*, 2018, **8**, 7403–7429.

10 K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich and T. Bligaard, Catalysis-Hub.org, an open electronic structure database for surface reactions, *Sci. Data*, 2019, **6**, 75.

11 R. Tran, J. Lan, M. Shuaibi, B. M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, A. Sriram, F. Therrien, J. Abed, O. Voznyy, E. H. Sargent, Z. Ulissi and C. L. Zitnick, The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts, *ACS Catal.*, 2023, **13**, 3066–3084.

12 H. Shahbeik, S. Hu, S. Motamedi, S. A. F. S. A. Tajuddin, M. Tabatabaei and M. Aghbashlo, A comprehensive review of CO2 integration in thermochemical biomass conversion for enhanced biofuel production, *Renewable Sustainable Energy Rev.*, 2026, **226**, 116442.

13 M. Golvirdizadeh, S. A. G. Nia, H. Shahbeik, A. Shafizadeh, H. Hosseinzadeh-Bandbafha, M. Kiehbadroudinezhad, S. A. Seyedalikhani, M. Ahmadi, A. Hajiahmad, S. A. F. S. A. Tajuddin, M. Tabatabaei and M. Aghbashlo, Machine learning–guided CO2 methanation: From catalyst design optimization to techno-economic and life cycle assessment analyses, *Energy*, 2025, 139708.

14 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, High-throughput experimentation and catalyst informatics for oxidative coupling of methane, *ACS Catal.*, 2019, **10**, 921–932.

15 T. Taniike and K. Takahashi, The value of negative results in data-driven catalysis research, *Nat. Catal.*, 2023, **6**, 108–111.

16 Z. Yang and W. Gao, Applications of machine learning in alloy catalysts: rational selection and future development of descriptors, *Adv. Sci.*, 2022, **9**, 2106043.

17 L.-H. Mou, T. Han, P. E. Smith, E. Sharman and J. Jiang, Machine learning descriptors for data-driven catalysis study, *Adv. Sci.*, 2023, **10**, 2301020.

18 S. Ishioka, A. Fujiwara, S. Nakanowatari, L. Takahashi, T. Taniike and K. Takahashi, Designing catalyst descriptors for machine learning in oxidative coupling of methane, *ACS Catal.*, 2022, **12**, 11541–11546.

19 G. Keller and M. Bhasin, Synthesis of ethylene via oxidative coupling of methane: I. Determination of active catalysts, *J. Catal.*, 1982, **73**, 9–19.

20 G. Hutchings, M. Scurrell and J. Woodhouse, Oxidative coupling of methane using oxide catalysts, *Chem. Soc. Rev.*, 1989, **18**, 251–283.

21 J. H. Lunsford, The catalytic oxidative coupling of methane, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 970–980.
22 S. Nishimura, X. Li, J. Ohyama and K. Takahashi, Leveraging machine learning engineering to uncover insights into heterogeneous catalyst design for oxidative coupling of methane, *Catal. Sci. Technol.*, 2023, **13**, 4646–4655.
23 S. Wu, G. Lambard, C. Liu, H. Yamada and R. Yoshida, iQSPR in XenonPy: a bayesian molecular design algorithm, *Mol. Inf.*, 2020, **39**, 1900107.
24 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, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
25 X. Zhou, Y. Pang, Z. Liu, E. I. Vovk, A. P. van Bavel, S. Li and Y. Yang, Active oxygen center in oxidative coupling of methane on La2O3 catalyst, *J. Energy Chem.*, 2021, **60**, 649–659.
26 Z. Xiong, J. Guo, Y. Deng, B. Liu, H. Lou, M. Zeng, Z. Wang, Z. Zhou, W. Yuan and F. Qi, Elucidating the mechanism for oxidative coupling of methane catalyzed by La2O3: experimental and microkinetic modeling studies, *ACS Catal.*, 2024, **14**, 1267–1280.
27 P. Schwach, W. Frandsen, M.-G. Willinger, R. Schlögl and A. Trunschke, Structure sensitivity of the oxidative activation of methane over MgO model catalysts: I. Kinetic study, *J. Catal.*, 2015, **329**, 560–573.

