THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Letter

Catalysis Gene Expression Profiling: Sequencing and Designing Catalysts

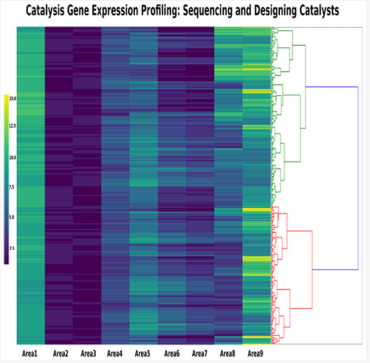
Keisuke Takahashi,* Jun Fujima, Itsuki Miyazato, Sunao Nakanowatari, Aya Fujiwara, Thanh Nhat Nguyen, Toshiaki Taniike, and Lauren Takahashi*

Cite This: *J. Phys. Chem. Lett.* 2021, 12, 7335−7341

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Identification of catalysts is a difficult matter as catalytic activities involve a vast number of complex features that each catalyst possesses. Here, catalysis gene expression profiling is proposed from unique features discovered in catalyst data collected by high-throughput experiments as an alternative way of representing the catalysts. Combining constructed catalyst gene sequences with hierarchical clustering results in catalyst gene expression profiling where natural language processing is used to identify similar catalysts based on edit distance. In addition, catalysts with similar properties are designed by modifying catalyst genes where the designed catalysts are experimentally confirmed to have catalytic activities that are associated with their catalyst gene sequences. Thus, the proposed method of catalyst gene expressions allows for a novel way of describing catalysts that allows for similarities in catalysts and catalytic activity to be easily recognized while enabling the ability to design new catalysts based on manipulating chemical elements of catalysts with similar catalyst gene sequences.



The rise of catalyst informatics has enabled the ability to design and understand catalysis from the underlying trends and patterns present in catalysts data.[1,2] Although data science techniques such as machine learning provide hints for descriptors representing catalytic activities, assigning the proper identification of catalysts remains a mystery.[3−8] In particular, identification of catalysts is a challenging matter as the manner in which catalysts respond to molecules and experimental conditions is strongly coupled with complex structural and compositional features in catalysts.[9−11] It is demonstrated that some catalysts exhibit performance patterns that are similar to each other during the reaction while combinations of chemical elements are unable to describe such similarities.[12] If these patterns can be organized in a manner similar to how gene expression is treated within genetics, it is possible to consider that these catalysts that exhibit similar series of performance patterns would also possess similar so-called "genetic profiles". Therefore, in principle, one can consider it possible to design catalysts via gene profiling.[8,13] Here, catalytic gene expressions are proposed where catalyst design via catalytic genetic profiling is performed where the designed catalysts are evaluated using high-throughput experiments.

The oxidative coupling of methane (OCM) reaction is chosen as a prototype reaction because of the availability of catalyst big data created via high-throughput experiments.[14,15] The OCM reaction involves the direct conversion of CH_4 to C_2H_4 and C_2H_6 where CO and CO_2 are known to be byproducts.[16,17] In the OCM reaction, CH_4 conversion, C_2 selectivity, and CO_x selectivity have been found to strongly couple with how catalysts respond against change in temperature and gas flows.[18] Given this, catalysts can be represented based on how catalysts behave against experimental conditions and resulting selectivities and conversion instead of representing catalysts as simple combinations of chemical elements. One can consider that an alternative way to express catalysts can be defined and the design of catalysts with this method can be achievable in principle if the unique features and patterns of each catalyst are properly extracted. Here, catalyst genetic expressions are designed and proposed by combining data science with OCM catalyst big data.

OCM catalyst big data is investigated where the data is previously collected using high-throughput experiments and data preprocessing is performed in order to check the outliers.[15] The data consists of 291 quaternary catalysts expressed as M1−M2−M3/Supports. M1, M2, and M3 are randomly selected from 28 elements with repetitive selection allowed while supports are randomly selected from 9 oxides.

Further detailed information regarding the data is provided in the Supporting Information. In particular, data contains 291 OCM catalysts with the following experimental conditions for producing the maximum C_2 yield: temperature (°C), CH_4 flow

Received: June 30, 2021
Accepted: July 15, 2021
Published: July 30, 2021

Table 1. Details of Select OCM Catalysts NaNiY−MgO, NiZnLa−Al_2O_3, and NiCuZr−Al_2O_3[a]

catalyst	T	CH_4	O_2	Ar	CH_4c	C_2s	C_2H_6s	C_2H_4s	COs	CO_2s
NaNiY−MgO	750	5.1	0.9	14	24.6	4.0	4.1	−0.1	106.9	3.6
NiZnLa−Al_2O_3	750	2.6	0.4	7	25.5	3.78	2.3	1.4	110.2	0.56
NiCuZr−Al_2O_3	750	4.8	1.2	14	37.8	4.2	2.9	1.3	81.5	12.0
catalyst	1	2	3	4	5	6	7	8	9	
NaNiY−MgO	40	3	7.4	19.3	14.3	4.1	2.0	53.4	55.2	
NiZnLa−Al_2O_3	38.8	1.5	3.7	16.3	14.7	3.1	1.9	55.7	55.3	
NiCuZr−Al_2O_3	39.9	3.0	7.6	25.9	21.0	3.5	2.0	41.4	46.7	

[a]T, CH_4, O_2, and Ar stand for temperature, CH_4 flow, O_2 flow, and Ar flow. The first table lists the following 10 descriptor variables for each catalyst: temperature (°C), CH_4 flow (mol/mL), O_2 flow (mol/mL), Ar flow (mol/mL), CH_4 conv, C_2s, C_2H_4s, C_2H_6s, COs, and CO_2s. The second table represents the nine areas of those catalysts calculated by the area under the curve.

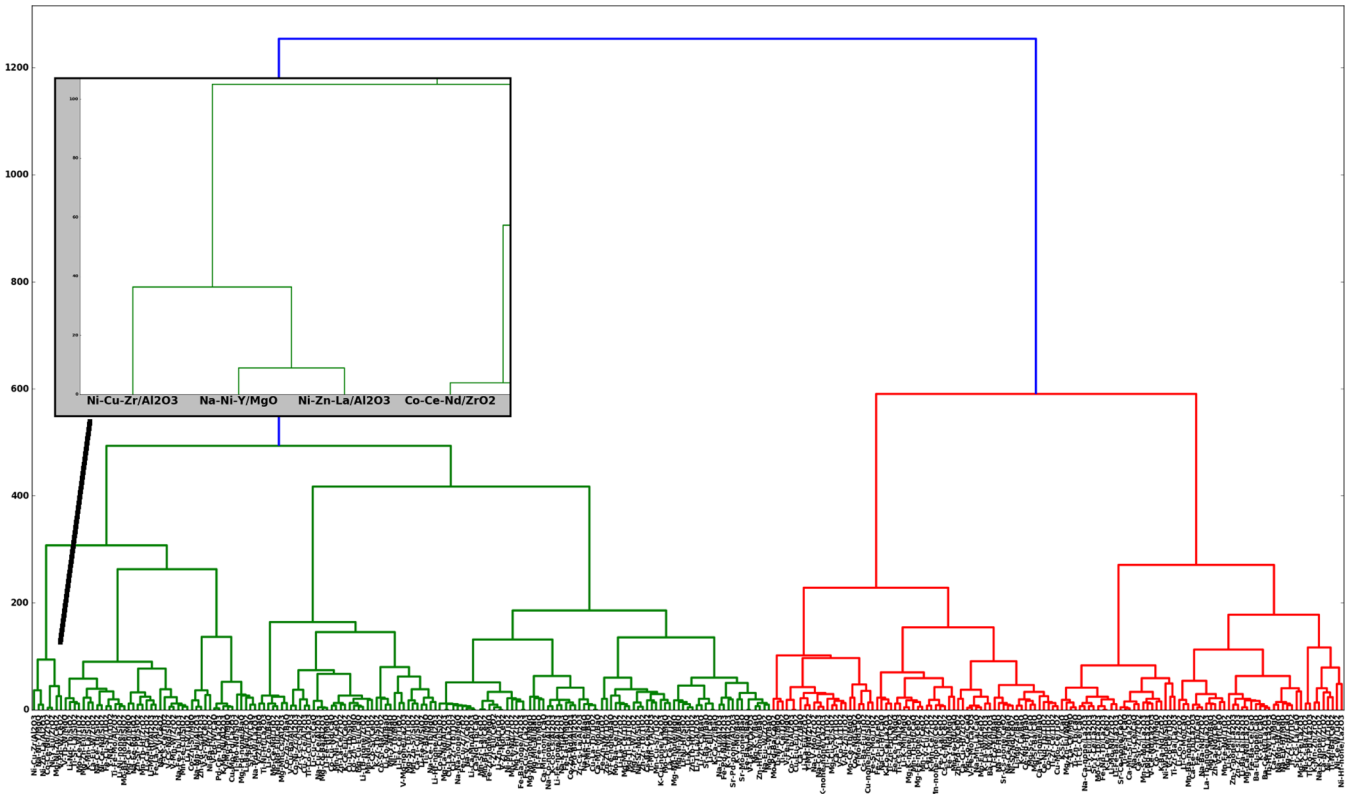


Figure 1. Dendrogram produced by hierarchical clustering for 291 OCM catalysts. The following 10 descriptor variables are used for hierarchical clustering: temperature (°C), CH_4 flow (mol/mL), O_2 flow (mol/mL), Ar flow (mol/mL), CH_4 conv, C_2s, C_2H_4s, C_2H_6s, COs, and CO_2s.

(mol/mL), O_2 flow(mol/mL), and Ar flow (mol/mL). Each catalyst's data also contains CH_4 conversion (CH_4 conv) and selectivity of C_2H_4 (C_2H_4s), C_2H_6 (C_2H_6s), CO (COs), CO_2 (C_2s), and sum of C_2H_4s and C_2H_6s (C_2s).

Validation experiments are performed on the catalysts designed by catalyst gene sequencing and natural language processing. The preparation and evaluation of the catalysts are performed according to the procedures identical to those employed for the 291 catalysts in the original data set. The catalysts are prepared by impregnating a specified oxide support with a solution containing salts of three specified elements, followed by calcination at 1000 (°C). The amount of each element is fixed at 0.37 mmol per gram of a support. The OCM performance of the prepared catalysts is evaluated under the same 135 sets of reaction conditions using the high-throughput screening instrument.[14,15] Out of the 135 sets, the data point corresponding to the maximum C_2 yield is extracted for the respective catalysts. Note that further details are found in previous work.[15]

Representations of the unique features found within 291 OCM catalysts are explored via data mining. In particular, each catalyst is defined by four experimental conditions for resulting in maximum C_2y (temperature (°C), CH_4 flow (mol/mL), O_2 flow (mol/mL), and Ar flow (mol/mL)) and 6 corresponding outputs (CH_4 conv, C_2s, C_2H_4s, C_2H_6s, COs, and CO_2s) (Table 1). Here, hierarchical clustering is performed using the above 10 variables for describing OCM catalysts in order to unveil the similarity in catalysts. A dendrogram produced using hierarchical clustering is illustrated in Figure 1 where OCM catalysts are classified based on similarities in catalytic performance with corresponding experimental conditions. For instance, Figure 1 indicates that catalysts NaNiY−MgO and NiZnLa−Al_2O_3 fall into the same group, suggesting that these catalysts would behave similarly during the OCM reaction. These two catalysts are found to have exhibited

https://doi.org/10.1021/acs.jpclett.1c02111
J. Phys. Chem. Lett. 2021, 12, 7335−7341

their best C2 yield at 750 C and the CH4/O2 ratio of 6 with the maximum selectivity toward CO. In a similar fashion, the dendogram also demonstrates that NiCuZr−Al_2O_3 is regarded to have performance traits similar to NaNiY−MgO and NiZnLa−Al_2O_3. Note that further details of the activities of these catalysts are collected in the Supporting Information. Ten descriptors for these catalysts are visualized in Figure 2,

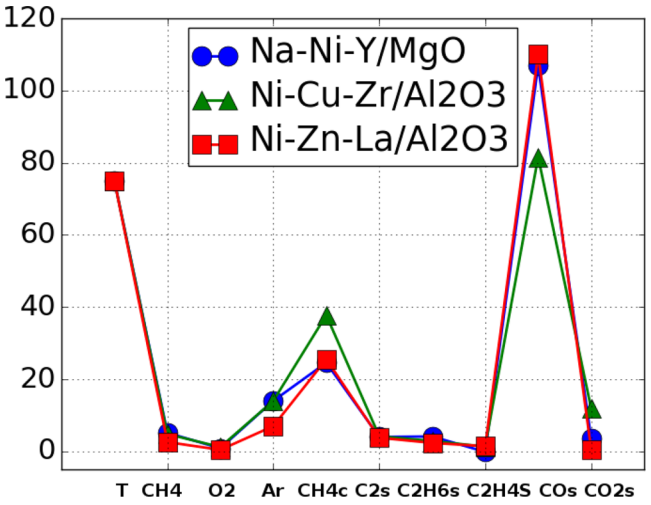


Figure 2. Visualization of the following 10 descriptor variables for NaNiY−MgO, NiZnLa−Al_2O_3, NiCuZr−Al_2O_3, and NiHf−La_2O_3: temperature (°C), CH_4 flow (mol/mL), O_2 flow (mol/mL), Ar flow (mol/mL), CH_4 conv, C_2s, C_2H_4s, C_2H_6s, COs, and CO_2s. Note that temperature is divided by 10.

where NaNiY−MgO and NiZnLa−Al_2O_3 are seen to have similar patterns, while NiCuZr−Al_2O_3 has a slightly different pattern compared to NaNiY−MgO and NiZnLa−Al_2O_3. Note that temperature is divided by 10. Thus, one could consider that NaNiY−MgO and NiZnLa−Al_2O_3 have a similar catalysis gene profile while a slightly different gene profile could be present in NiCuZr−Al_2O_3.

Hidden patterns that contribute to the dendrogram illustrated in Figure 1 are explored. More specifically, patterns that act as signatures for the catalyst profiles are extracted. This is carried out by measuring nine areas under the curves seen in Figure 3 where NaNiY−MgO is used as an example. In Figure 3, 10 descriptors are plotted in the following order: temperature, CH_4 flow, O_2 flow, Ar flow, CH_4 conv, C_2s, C_2H_6s, C_2H_4s, COs, and CO_2s. Note that it is important to keep the order of descriptors throughout the design of catalysts genes sequencings. The area under the curve between two variable points are sectioned, resulting in the creation of nine areas, as illustrated in Figure 3. Here, the areas under the curve for each of the nine areas in NaNiY−MgO are calculated as the following: 1, 40.1; 2, 3.0; 3, 7.4; 4, 19.3; 5, 14.3; 6, 4.1; 7, 2.0; 8, 53.4; and 9, 55.2. In the same fashion, the area under the curve for the proposed nine areas are calculated for the rest of the 291 OCM catalysts' data where the nine areas for each catalysts are illustrated in Figure 4. One can see that each area in the 291 catalyst produces the unique pattern as shown in Figure 4; thus, calculating the area demonstrated in Figure 3 unveils the unique features of catalysts. In order to create gene sequencings of the catalysts based on patterns shown in Figure 4, alphabetical letters are assigned in each pattern. In particular, a total of 15 alphabetical letters, A−O, are created along the *y* axis in increments of five ranging from 0 to 75 as shown in

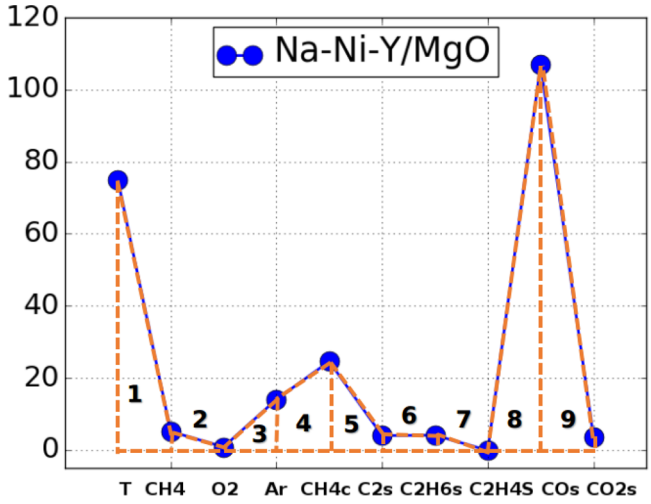


Figure 3. Nine areas where area under the curve is applied for calculating the areas. NaNiY−MgO in Figure 2 is used as an example.

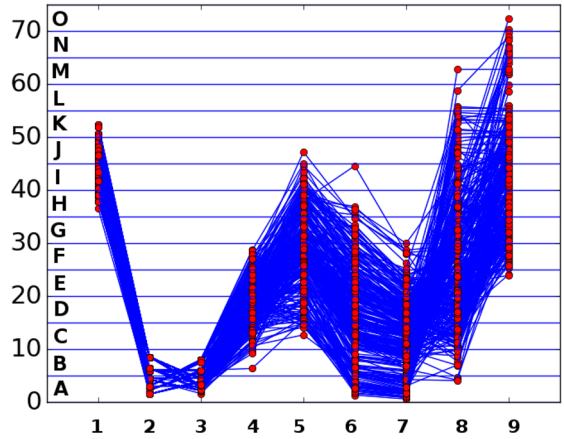


Figure 4. Plot of the area under the curve for each of the nine areas of 291 catalysts. Alphabet letters A through O are assigned in increments of five.

Figure 4. The calculated area under the curve as illustrated in Figure 3 is transformed into a corresponding alphabetical letter. A series of alphabetical letters for each catalyst is then proposed as its catalyst gene sequence. Note that created catalyst gene sequences are collected in the Supporting Information. As a result, NaNiY−MgO, NiZnLa−Al_2O_3, and NiCuZr−Al_2O_3 are then represented as IABDCAAKL, HAADCAALL, and HABFEAAIJ, respctively. The same method of assigning a catalyst gene sequence is then applied toward the remaining catalysts within the data; thus, catalyst gene sequencings are created.

Catalyst gene expression profiling is carried out based on the developed catalyst gene sequences. In particular, catalyst sequences are visualized in the heatmap displayed in Figure 4 by combining the corresponding hierarchical clustering map seen in Figure 1. Note that the alphabetical letters are converted into numerical form for visualization where A−O becomes 1−15, respectively.

From the heatmap in Figure 5, one can see that there are several noticeable features present when comparing the green and red groups generated by hierarchical clustering. Here, key patterns in Figure 5 are investigated. To start, area 3, which represents O_2 and Ar flows, indicates that members of the red

https://doi.org/10.1021/acs.jpclett.1c02111
J. Phys. Chem. Lett. 2021, 12, 7335−7341

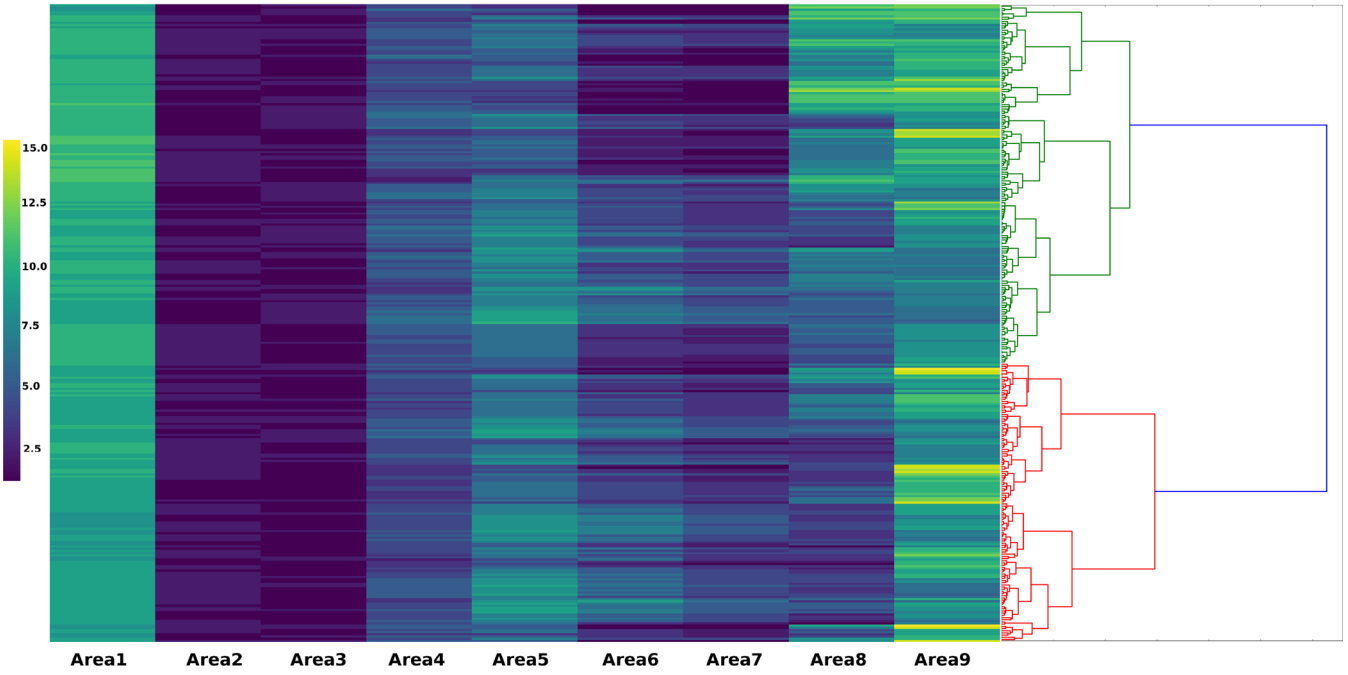


Figure 5. Catalysis gene expression profiling. Dendrogram produced by hierarchical clustering of 291 OCM catalysts as illustrated in Figure 1 with the corresponding calculated 9 areas of each of the 291 catalysts using area under the curve visualized as a heatmap.

group have a smaller area than those who belong to the green group. In particular, it is shown that catalysts found in the red group tend to have lower Ar flow than the catalysts found in the green group. Another noticeable feature is located at areas 5 and 6, which represent CH_4conv−C_2s and C_2s−C_2H_6s, respectively. One can see that areas 5 and 6 of the catalysts that fall within the red group are visually brighter than those that fall within the green group. Closer inspection reveals that catalysts that belong to the red group tend to possess high C_2s and C_2H_6s when compared against those of the green group. A clear difference is also observed for area 8, which represents C_2H_4s−COs, where catalysts that fall within the green group have a bright color while those of the red group have a dark color. The major difference can be seen as a reflection of COs of the green and red groups, where catalysts that fall within the green group have a much higher COs than those that fall within the red group. Given these results, one can say that catalysts that fall within the red group possess high C_2s and low COs with low Ar flow while catalysts that fall within the green group exhibit the opposite. These similarities can be seen within the catalyst gene expressions as well. For instance, NaNiY−MgO and NiZnLa−Al_2O_3 in the green group have similar catalyst gene sequences of IABDCAAKL and HAADCAALL, respectively. Thus, the proposed catalyst gene sequences are found to successfully represent the unique features of the catalysts while showing good agreement with hierarchical clustering and can be used to help compare catalysts and determine how similar they may be to each other.

Further investigation of each individual catalyst gene sequencing is conducted using natural language processing in order to evaluate similarities present in the catalyst genes. In particular, edit distance in NLTK is implemented, where edit distance evaluates the minimum number of letters to transform one sequence to another.[19] Here, Na−Ni−Y−MgO is chosen as a base catalyst for comparison, where its gene sequencing is defined as IABDCAAKL. The following five catalysts are compared to Na−Ni−Y−MgO using edit distance: Ni−Zn−La−Al_2O_3, Ni−Cu−Zr−Al_2O_3, Ni−Ce−W−TiO_2, Na−Ca−Mn−BaO, and Na−Tb−Hf−MgO. Further details regarding the edit distance and catalyst gene sequences are collected in Table 2. Ni−Zn−La−Al_2O_3 is found to have an edit distance of 3, allowing one to consider that Ni−Zn−La−Al_2O_3 has a gene sequence similar to that of Na−Ni−Y−MgO. Ni−Cu−Zr−Al_2O_3, meanwhile, is found to have an edit distance of 5. The edit distances show good agreement with the hierarchical clustering shown in Figure 5 where Na−Ni−Y−MgO is seen to be similar to Ni−Zn−La−Al_2O_3 with a slight difference from Ni−Cu−Zr−Al_2O_3. In the same fashion, the edit distances of catalysts that belong to the red group in Figure 5 are evaluated for the following catalysts: Ni−Ce−W−TiO_2, Na−Ca−Mn−BaO, and Na−Tb−Hf−MgO. Ni−Ce−W−TiO_2, Na−Ca−Mn−BaO, and Na−Tb−Hf−MgO are found to have edit distances of 6, 6, and 8, respectively. From this, one can come to understand the following: Ni−Ce−W−TiO_2, Na−Ca−Mn−BaO, and Na−Tb−Hf−MgO behave similarly to each other and, as a group, are distinct from Na−Ni−Y−

Table 2. Catalyst Gene Sequences of Randomly Selected Catalysts Where Three Catalysts Are Taken from the Green Group Illustrated in Figure 1 and Three Others Are Taken from the Red Group Illustrated in Figure 1[a]

catalyst	genes	ED	group
Na−Ni−Y−MgO	IABDCAAKL	0	green
Ni−Zn−La−Al_2O_3	HAADCAALL	3	green
Ni−Cu−Zr−Al_2O_3	HABFEAAIJ	5	green
Ni−Ce−W−TiO_2	IAADGDDHJ	6	red
Na−Ca−Mn−BaO	IBBDGEDCG	6	red
Na−Tb−Hf−MgO	JBADFECCH	8	red

[a]ED represents edit distance where Na−Ni−Y−MgO is chosen as the base catalyst gene sequence. "Group" represents the red and green groups visualized in Figure 1.

https://doi.org/10.1021/acs.jpclett.1c02111
J. Phys. Chem. Lett. 2021, 12, 7335−7341

MgO, Ni−Zn−La−Al_2O_3, and Ni−Cu−Zr−Al_2O_3. This grouping is reflected in the grouping provided by the hierarchical clustering demonstrated in Figure 5, showing agreement where green-group catalysts, the group that catalyst Na−Tb−Hf−MgO belongs to, have a small edit distance while the catalysts that belong to the red group have a large edit distance in comparison. Thus, natural language processing becomes the way to evaluate the similarities in catalysts.

In the same fashion, catalyst gene sequences having the highest C_2y, C_2s, and CH_4conv are investigated in order to determine the presence of similar genes. Sr−Mo−none−BaO, Zn−Zn−Nd−TiO_2, and Mn−Tb−Hf−ZrO_2 have the highest C_2y, C_2s, and CH_4conv, respectively. Using edit distance, the minimum edit distance for these catalysts are explored. The discovered similar catalysts and their corresponding gene sequences are collected in Table 3. Table 3 shows that catalyst

Table 3. Catalysts Having High C_2y (%) in First Table, High C_2s (%) in the Second Table, and High CH_4conv (%) in the Third Table Where Sr−Mo−none−BaO, Zn−Zn−Nd−TiO_2, and Mn−Tb−Hf−ZrO_2 Are Chosen as a Base Catalyst for Each Case[a]

catalyst	genes	C_2y	ED
Sr−Mo−none−BaO	IABFIFEDF	21.2	0
Li−Mg−Zr−BaO	IABFIGEFH	18.6	3
K−V−Mo−BaO	IABEIGFFF	18.5	3
Na−Eu−W−ZrO_2	IABFIFEEF	18.2	3
Mg−Sr−Ba−CaO	IABFIFEDF	17.5	3
Mg−K−Y−BaO	IABFIFEDF	17.0	3
Zn−Hf−BaO	IABFIFEEF	16.9	3
catalyst	genes	C_2s	ED
Zn−Zn−Nd−TiO_2	JBACHHFIF	59.9	0
Fe−Pd−Nd−Al_2O_3	JBACHHFFG	57.3	3
Mg−K−Fe−La_2O_3	IBACHIFCF	56.3	3
Na−Ti−Cu−Al_2O_3	JBBCHHFFG	55.9	3
catalyst	genes	C_4conv	ED
Mn−Tb−Hf−ZrO_2	IAAFEAAGJ	46.7	0
Zn−Sr−Pd−SiO_2	IABFEAAGJ	43.5	3
V−Pd−W−CeO_2	IAAEEAAHJ	42.9	3
Ni−Cu−Zr−Al_2O_3	HABFEAAIJ	37.8	3
Mo−Pd−Ba−CeO_2	IAADDAAGN	33.3	3

[a]Genes and ED represent catalyst gene sequencing and edit distance, respectively.

Sr−Mo−none−BaO possesses genes similar to six catalysts, Zn−Zn−Nd−TiO_2 to three catalysts, and Mn−Tb−Hf−ZrO_2 to four catalysts, where all catalysts are found to have edit distances of 3. Further investigation reveals that the catalysts are found to have C_2y, C_2s, and CH_4conv similar to Sr−Mo−none−BaO, Zn−Zn−Nd−TiO_2, and Mn−Tb−Hf−ZrO_2, respectively. Here, it raises the question of whether catalysts having similar catalyst gene sequences truly have similar catalytic activities, including experimental conditions and other selectivities. Parallel coordinates shown in Figure 6 are explored in order to evaluate how the catalytic performances of the catalysts listed in Table 3 are similar. In particular, the following 10 variables are visualized: temperature, CH_4 flow, O_2 flow, Ar flow, CH_4 conv, C_2s, C_2H_6s, C_2H_4s, COs, and CO_2s. Note that color is assigned for similar catalysts having high C_2y, C_2s, and CH_4conv as shown in Table 3. Figure 6 demonstrates that similar catalyst gene sequences in Table 3 have very similar catalytic performance in terms of

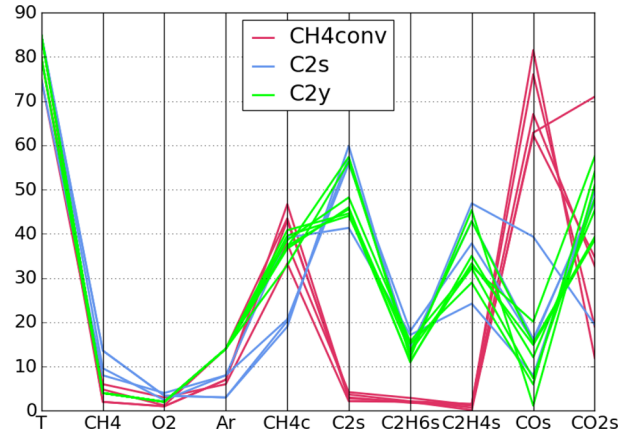


Figure 6. Parallel coordinates of the following 10 descriptor variables of catalysts collected in Table 3: temperature, CH_4 flow, O_2 flow, Ar flow, CH_4 conv, C_2s, C_2H_4s, C_2H_6s, COs, and CO_2s. Color is assigned based on catalysts having high C_2y, C_2s, and CH_4conv.

experimental conditions for achieving the highest C_2y, C_2s, and CH_4 conversion. Analyzing catalysts as listed in Table 3 according to their chemical elements makes it very difficult to determine any similarities that may be present as they are composed of different chemical elements. However, listing catalysts according to assigned catalyst gene sequences uncovers the similarities present between catalysts which would otherwise be unobservable if observed only by combinations of chemical elements. Thus, the proposed catalyst gene sequences unveil the hidden similarities present within catalysts including support while also leading toward searching and evaluating similar catalysts using edit distance within natural language processing.

Catalyst design is performed using catalyst gene sequencing and natural language processing. Given the previous observations, one can consider that catalysts with similar catalyst gene sequences would also possess similar catalytic performance. In this sense, if catalysts having similar gene sequences can be designed, it should therefore be possible to design catalysts that have different chemical elements but have duplicate performance. Therefore, the following steps are taken to design the catalysts. Catalysts having similar catalyst gene sequences are first explored using edit distance. Then, common chemical elements are identified in the catalysts having similar gene sequences where catalysts are then designed by rearranging the uncovered common chemical elements. Here, four types of catalyst gene sequences are investigated, as shown in Table 4, where the parallel coordinates of the four catalysts and the designed catalysts are also shown in Figure 7. Li−Fe−Ba−La_2O_3, Fe−Nd−Tb−La_2O_3, and K−La−Ce−CeO_2 are found to have the catalyst gene sequences of HAADHGECG with similar C_2y. These catalysts hint toward the suggestion that the combinations of Fe−La and Ce−La could be a key combination. Therefore, rearranging the Li−Fe−Ba−La_2O_3 with Ce is considered, resulting in Li−Fe−Ce−La_2O_3 and Fe−Ba−Ce−La_2O_3. Li−Fe−Ce−La_2O_3 and Fe−Ba−Ce−La_2O_3 are then evaluated in a high-throughput experiment where the C_2y of Li−Fe−Ce−La_2O_3 and Fe−Ba−Ce−La_2O_3 are found to be 12.8 % and 14.0%, respectively. These are found to be similar to those of Li−Fe−Ba−La_2O_3, Fe−Nd−Tb−La_2O_3, and K−La−Ce−CeO_2. Furthermore, catalyst gene sequences of Li−Fe−Ce−La_2O_3 and Fe−Ba−Ce−La_2O_3 are found to be HAADHGECG and HAADHGECG, respectively.

https://doi.org/10.1021/acs.jpclett.1c02111
J. Phys. Chem. Lett. 2021, 12, 7335−7341

Table 4. Four Types of Catalysts Having Similar Catalyst Gene Sequences[a]

catalyst 1	genes	C_2y	type	ED
Li−Fe−Ba−La_2O_3	HAADHGECG	15.2	data	0
Fe−Nd−Tb−La_2O_3	HAADHGECG	13.4	data	0
K−La−Ce−CeO_2	HAADHGECG	13.4	data	0
Li−Fe−Ce−La_2O_3	HBADHFDCI	12.8	designed	4
Fe−Ba−Ce−La_2O_3	HAAEHFEBH	14.0	designed	4
catalyst 2	genes	C_2y	type	ED
Li−K−Mn−MgO	IAADFDCEJ	8.2	data	0
Na−K−Hf−Al_2O_3	IAADFDCEJ	8.2	data	0
K−none−none−La_2O_3	IAADFDCFJ	7.9	data	1
Mg−K−none−CeO_2	IAADFDCFJ	8.5	data	1
K−Ce−none−MgO	IBADGFDCJ	11.5	designed	3
K−Hf−none−CeO_2	JBADFCCEL	7.63	designed	4
Na−Mn−Sr−Al_2O_3	IABFHEDCF	14.1	designed	6
catalyst 3	genes	C_2y	type	ED
Mg−Sr−Ba−CaO	IABFIFEDF	17.5	data	0
Mg−K−Y−BaO	IABFIFEDF	17.0	data	0
Mg−Ca−Sr−BaO	IABDJKHDD	17.4	designed	5
Mg−K−Ba−BaO	IABEHFDAF	12.8	designed	4
catalyst 4	genes	C_2y	type	ED
Na−Eu−W−ZrO_2	IABFIFEEF	18.2	data	0
Zn−Hf−none−BaO	IABFIFEEF	16.9	data	0
Eu−Hf−W−BaO	IABFIGECF	16.9	designed	2

[a]C_2y, genes, and ED represent catalyst C_2 yield, catalyst gene sequence, and edit distance, respectively. Type represents whether catalysts are from the 291 catalysts data ("data") or are newly designed catalysts evaluated in experiment ("designed"). Note that "none" represents the cases where no additional chemical elements are added.

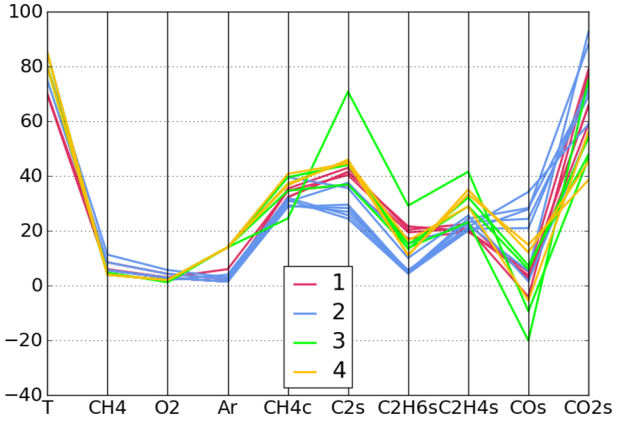


Figure 7. Parallel coordinates of the following 10 descriptor variables of catalysts collected in Table 4: temperature, CH_4 flow, O_2 flow, Ar flow, CH_4 conv, C_2s, C_2H_4s, C_2H_6s, COs, and CO_2s. Color is assigned based on the 4 assigned groups listed in Table 4.

In both cases, the edit distance is calculated to be 4, which can be considered to be similar, as Table 3 and Figure 6 demonstrate that edit distances of 3 result in similar catalytic performances. It must also be noted that Li−Fe−Ba−La_2O_3, Fe−Nd−Tb−La_2O_3, and K−La−Ce−CeO_2 reach maximum performance at a temperature of 700 °C, whereas designed catalysts Li−Fe−Ce−La_2O_3 and Fe−Ba−Ce−La_2O_3 also have a temperature of 700 °C. Thus, the designed catalysts are found to have similar catalytic performances. Note that the details of designed catalysts are listed in the Supporting Information.

In the same manner, catalysts are designed based on Li−K−Mn−MgO, Na−K−Hf−Al_2O_3, K−none−none−La_2O_3, and Mg−K−none−CeO_2 where these catalysts have low C_2y. Note that "none" indicates that chemical element is not assigned. These catalysts have similar genes of either IAADFDCFJ or IAADFDCFJ with an edit difference of 1. Catalysts are designed by rearranging the chemical elements within the 4 catalysts having similar catalyst gene sequences. Here, K−Hf−none−CeO_2 is designed by subsituting Mg in Mg−K−none−CeO_2 with Hf, resulting in the gene sequence of JBADFCCEL. The edit distance of JBADFCCEL is 4 with aC_2y of 7.63 %; thus, similar catalytic performance of low C_2y is demonstrated. On the other hand, K−Ce−none−MgO and Na−Mn−Sr−Al_2O_3 are also designed by rearranging the elements in Li−K−Mn−MgO, Na−K−Hf−Al_2O_3, K−none−none−La_2O_3, and Mg−K−none−CeO_2. C_2y and the edit distances of K−Ce−none−MgO and Na−Mn−Sr−Al_2O_3 are slightly different from those based on Li−K−Mn−MgO, Na−K−Hf−Al_2O_3,K−none−none−La_2O_3, and Mg−K−none−CeO_2. However, the overall pattern of parallel coordinates shown in Figure 7 is still similar.

High active catalysts Mg−Sr−Ba−CaO and Mg−K−Y−BaO have the same gene sequences of IABFIFEDF. Rearranging the chemical elements provides designed catalysts Mg−Ca−Sr−BaO and Mg−K−Ba−BaO which are then evaluated in high-throughput experiment and result in aC_2y of 17.4% and 12.8%, respectively, with gene sequences of IABDJKHDD and IABEHFDAF, respectively. In the case of Mg−Ca−Sr−BaO, it has similar C_2y with a large edit distance of 5. Further analysis reveals that Mg−Ca−Sr−BaO has a different C_2H_6s compared to Mg−Sr−Ba−CaO and Mg−K−Y−BaO, which can be also seen at C_2s−C_2H_6s shown in Figure 7. On the other hand, Mg−K−Ba−BaO has an edit distance of 4 which has a similar gene sequence to Mg−Sr−Ba−CaO and Mg−K−Y−BaO; however, C_2y in Mg−K−Ba−BaO is slightly lower than C_2y for Mg−Sr−Ba−CaO and Mg−K−Y−BaO.

Finally, Na−Eu−W−ZrO_2 and Zn−Hf−none−BaO are investigated; they have the same catalyst gene sequence of IABFIFEEF. Eu−Hf−W−BaO is created by rearranging the chemical elements of catalysts Na−Eu−W−ZrO_2 and Zn−Hf−none−BaO. Evaluation via high-throughput experiment reveals that Eu−Hf−W−BaO has a C_2y and gene sequence of 16.9% and IABFIGECF, respectively. The edit distance is also calcualted to be 2; thus, both C_2y and gene sequence are found to be similar to Na−Eu−W−ZrO_2 and Zn−Hf−none−BaO, as shown in Table 4.

Overall, catalysts are designed based on the rearrangement of chemical elements within catalysts having similar catalyst gene sequencing. As shown in Figure 7, the overall patterns of the designed catalysts are similar to the catalysts they are based off of. However, cases like group 3 have slightly different patterns, and a further, more precise way to design those catalysts should be considered.

In summary, catalyst gene sequences are proposed according to the catalytic performance patterns present within 291 catalysts. The proposed catalyst gene sequences prove to be an alternative way of representing catalysts where catalysts have been traditionally expressed according to their chemical element combinations. This method allows for catalyst gene sequences to be applied toward unreported catalysts that may be discovered in order to see how similar they may be to

https://doi.org/10.1021/acs.jpclett.1c02111
J. Phys. Chem. Lett. 2021, 12, 7335−7341

previously reported catalysts. If similar catalyst gene sequences are not observed, researchers can move on to establish the new catalyst as a unique catalyst. Using catalyst gene sequences would provide a method for searching for similarities within the catalysts and thereby provide a way to classify catalysts based on similarities in their catalyst gene sequences. It is also demonstrated that catalyst design according to catalyst gene sequence is also possible by rearranging chemical elements found within catalysts with similar gene sequences where new catalysts having similar catalytic performance can be intentionally designed. Hence, the proposed catalyst gene sequences not only change how catalysts are represented but also give an alternative way to design catalysts with particular sets of catalytic performances.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.1c02111.

Catalyst information with corresponding catalyst genes (Table S1) and designed catalysts' information with corresponding catalyst genes (Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Keisuke Takahashi − *Department of Chemistry, Hokkaido University, Sapporo 060-8510, Japan;* orcid.org/0000-0002-9328-1694; Email: keisuke.takahashi@sci.hokudai.ac.jp

Lauren Takahashi − *Department of Chemistry, Hokkaido University, Sapporo 060-8510, Japan;* orcid.org/0000-0001-9922-8889; Email: lauren.takahashi@sci.hokudai.ac.jp

Authors

Jun Fujima − *Department of Chemistry, Hokkaido University, Sapporo 060-8510, Japan*

Itsuki Miyazato − *Department of Chemistry, Hokkaido University, Sapporo 060-8510, Japan;* orcid.org/0000-0002-1533-9790

Sunao Nakanowatari − *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan*

Aya Fujiwara − *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan*

Thanh Nhat Nguyen − *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan*

Toshiaki Taniike − *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan;* orcid.org/0000-0002-4870-837X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.1c02111

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is funded by Japan Science and Technology Agency (JST) CREST Grant Number JPMJCR17P2.

■ REFERENCES

(1) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37.

(2) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data Through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403−7429.

(3) Zavyalova, U.; Holena, M.; Schlögl, R.; Baerns, M. Statistical Analysis of Past catalytic Data on Oxidative Methane Coupling for New Insights into the Composition of High-performance Catalysts. *ChemCatChem* **2011**, *3*, 1935−1947.

(4) Kitchin, J. R. Machine learning in catalysis. *Nature Catalysis* **2018**, *1*, 230−232.

(5) Li, Z.; Wang, S.; Xin, H. Toward Artificial Intelligence in Catalysis. *Nature Catalysis* **2018**, *1*, 641−642.

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

(7) Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine Learning for Computational Heterogeneous Catalysis. *ChemCatChem* **2019**, *11*, 3581−3601.

(8) Takahashi, K.; Takahashi, L.; Miyazato, I.; Fujima, J.; Tanaka, Y.; Uno, T.; Satoh, H.; Ohno, K.; Nishida, M.; Hirai, K.; Ohyama, J.; Nguyen, T. N.; Nishimura, S.; Taniike, T.; et al. The Rise of Catalyst Informatics: Towards Catalyst Genomics. *ChemCatChem* **2019**, *11*, 1146−1152.

(9) VOSKRESENSKAYA, E. N.; ROGULEVA, V. G.; ANSHITS, A. G. Oxidant activation over structural defects of oxide catalysts in oxidative methane coupling. *Catal. Rev.: Sci. Eng.* **1995**, *37*, 101−143.

(10) Ji, S.-f.; Xiao, T.-c.; Li, S.-b.; Xu, C.-z.; Hou, R.-l.; Coleman, K. S.; Green, M. L. The relationship between the structure and the performance of Na-W-Mn/SiO2 catalysts for the oxidative coupling of methane. *Appl. Catal., A* **2002**, *225*, 271−284.

(11) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.

(12) Takahashi, K.; Takahashi, L.; Nguyen, T. N.; Thakur, A.; Taniike, T. Multidimensional Classification of Catalysts in Oxidative Coupling of Methane through Machine Learning and High-Throughput Data. *J. Phys. Chem. Lett.* **2020**, *11*, 6819−6826.

(13) Nørskov, J. K.; Bligaard, T. The Catalyst Genome. *Angew. Chem., Int. Ed.* **2013**, *52*, 776−777.

(14) Nguyen, T. N.; Nhat, T. T. P.; Takimoto, K.; Thakur, A.; Nishimura, S.; Ohyama, J.; Miyazato, I.; Takahashi, L.; Fujima, J.; Takahashi, K.; Taniike, T.; et al. High-Throughput Experimentation and Catalyst Informatics for Oxidative Coupling of Methane. *ACS Catal.* **2020**, *10*, 921−932.

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

(16) Keller, G.; Bhasin, M. Synthesis of Ethylene via Oxidative Coupling of Methane: I. Determination of Active Catalysts. *J. Catal.* **1982**, *73*, 9−19.

(17) Galadima, A.; Muraza, O. Revisiting the Oxidative Coupling of Methane to Ethylene in the Golden Period of Shale Gas: A Review. *J. Ind. Eng. Chem.* **2016**, *37*, 1−13.

(18) Ohyama, J.; Nishimura, S.; Takahashi, K. Data Driven Determination of Reaction Conditions in Oxidative Coupling of Methane via Machine Learning. *ChemCatChem* **2019**, *11*, 4307−4313.

(19) Bird, S. *NLTK: The natural language toolkit*; 2006; pp 69−72.