Catalyst gene component manual

Shibata Kenshin

1 Introduction

1.1 Catalyst gene

Catalytic gene profiling is a useful method to explain similarities in catalytic performance that cannot be explained by chemical elements. The response of catalytic activity to synthesis and experimental conditions is unique to each catalyst, and therefore the conditions (high temperature and pressure, high temperature and low pressure, gas composition, etc.) that give the best catalytic activity can represent the catalyst. If sequences that represent the conditions leading to the highest activity are created, the similarities of catalysts can be discussed by investigating the similarities of created sequence, just as the similarities of organisms can be discussed by the similarities of genes.

1.2 Catalyst gene introduction

The design of catalytic genes is a method of representing the characteristics of each catalyst as a string based on numerical data. First, experimental conditions and experimental results are plotted on the x-axis, while the corresponding numerical data are plotted on the y-axis to create a line graph for each catalyst. Adjacent data points are connected by straight lines, and the area enclosed by these lines and the x-axis is calculated. This process is repeated for all data points of each catalyst to obtain a series of area values. The obtained area values are divided into 15 equal intervals (bins) ranging from the minimum to the maximum value. Each bin is assigned an alphabetical label in ascending order from A, B, C... to O. This allows each area value to be mapped to a specific letter. Finally, for each catalyst, the sequence of letters corresponding to the calculated area values is arranged in order, forming the catalytic gene.

1.3 Edit distance

Edit Distance is a metric that represents the minimum number of operations required to transform one string into another. This metric is highly useful for measuring the similarity between strings and is widely applied in various fields, especially in natural language processing and DNA sequence comparison. The most commonly used definition is Levenshtein Distance, which is calculated using three basic operations: insertion, deletion, and substitution.

2 Requried data preprocessing

2.1 Column Name Standardization

As shown in the red box in Figure 1, your data must contain a column that identifies the catalyst, and that column name must be "Catalyst" (with a capital C). Additionally, the elements in the Catalyst column must be unique.

2.2 Data Filtering

When multiple experimental conditions exist for the same catalyst name, retain only the data showing the best results for the property of interest and delete the rest.

2.3 Missing Value Imputation

If the experimental conditions or results columns contain missing values such as NaN, the columns will not be reflected in the design of catalytic genes. Please either complete them using an appropriate method (mean completion, mode completion, interpolation, etc.) or delete the data containing missing values.

2.4 Element Information

The dataset must include a column containing compositional information of the catalyst. The compositional information can be included either as a column indicating the elements or as a one-hot encoding representation.

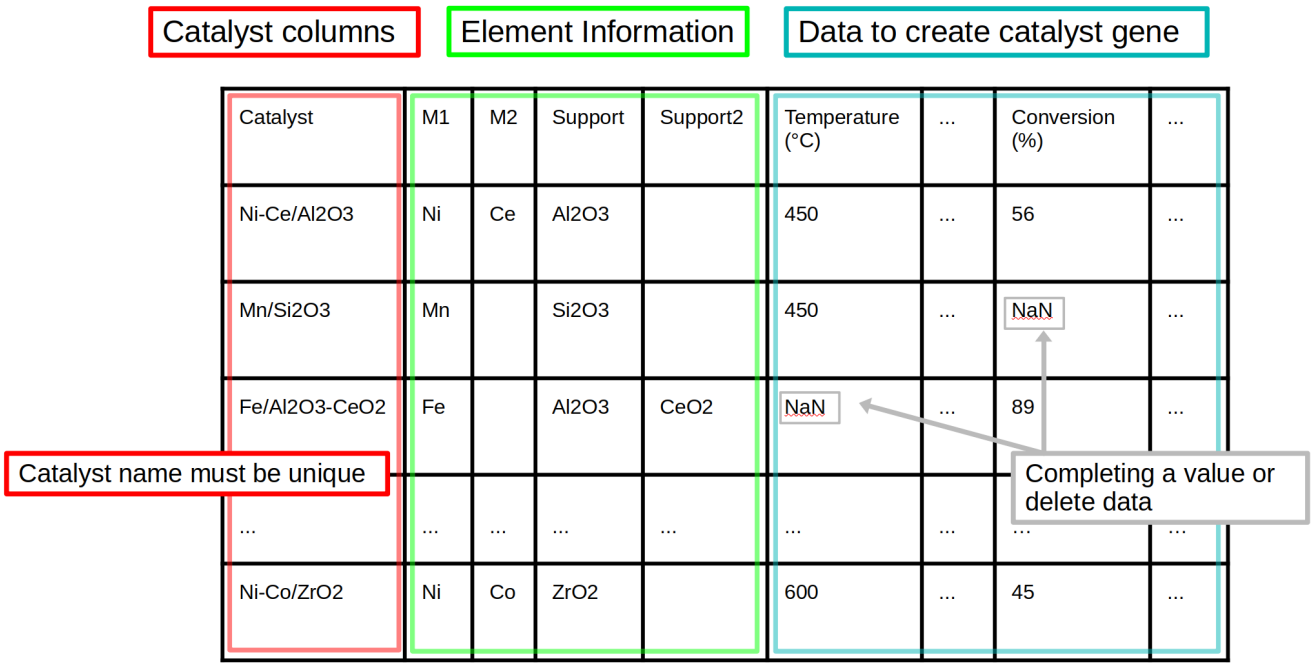


Figure 1: Requried data preprocessing

3 Component manual

3.1 Configure

- **Feature columns** : Select the columns you will use to create the catalytic gene. It is recommended to include the information of experimental conditions and measurement results.
- **Catalyst** : Select Base Catalyst that is the reference catalyst used as the 'origin' for evaluating the similarity of catalyst genes.
- **Onehot encoding element info** : Turn on this check box if the element information is held in one-hot encoding format.
- **Element columns** :
  (if "Onehot encoding element info" is OFF) Select column names that contain element information.
  (if "Onehot encoding element info" is ON) Specify the first and last columns of the onehot encoding of the elemental information
- **Apply Scaling** : Turn on this check box if the element information is held in one-hot encoding format.
- **Scaling Method** : If Apply data scaling is on, You can choose scaling method you want to use. Max and Min value is required when you choose MinMaxScaler.
- **Visualization** : Select visualization method. Each visualization is explained in 3.2(Visualization).
- **Clustering method** : When Hierarchical Clustering or Heatmap was selected in Visualization, Specify clustering method.
- **Color Palette** : when Heatmap is selected in Visualization, you need to select color palette for the heatmap.

3.2 Visualization

3.2.1 Clustering

In hierarchical clustering, the clustering is performed based on the method selected by the user. The clustering result example is shown in Figure 2

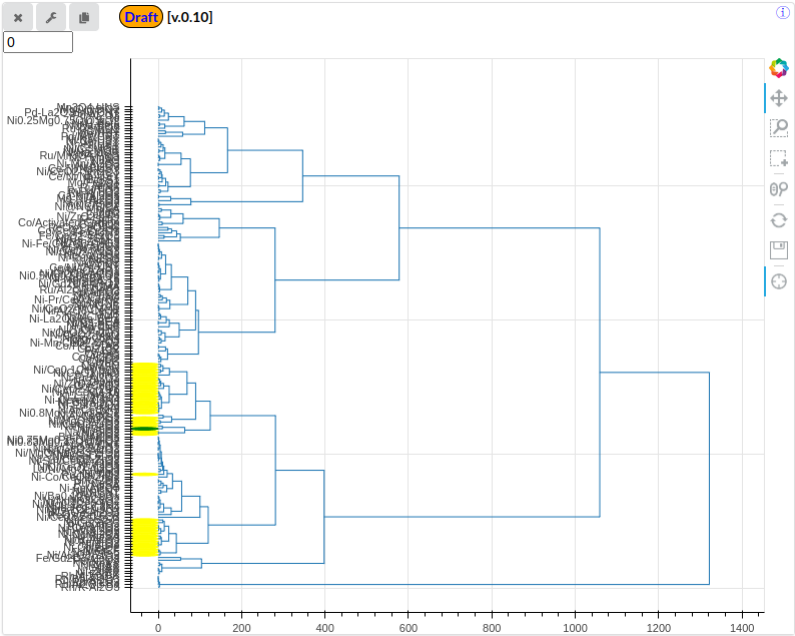


Figure 2: Clustering result.The root catalyst selected by the user is shown in green. The number displayed in the upper left represents the edit distance threshold, and catalysts with an edit distance equal to or below this value are marked in yellow. The edit distance threshold can be adjusted by hovering over the box in the upper left corner.

3.2.2 Area plot

In the area plot, a series of area values calculated to create the catalyst gene for each catalyst are plotted on a single graph. The selected root catalyst is displayed as an orange line, and by analyzing this plot, the characteristics and trends of the catalysts can be visually examined. Each data point represents the area formed by the values of two adjacent items and the x-axis, numerically expressing the characteristics of the catalyst. The area values correspond to the raw data before being converted into alphabets for designing the catalyst gene. The alphabet corresponding to each data point is displayed on the left end, and the data belonging to regions separated by horizontal lines are converted into that alphabet. Additionally, by clicking the button in the top-left corner, you can display the items used to calculate each area. This allows for an easy check of which data were used in the area calculations. The Area plot result example is shown in figure 3, 4

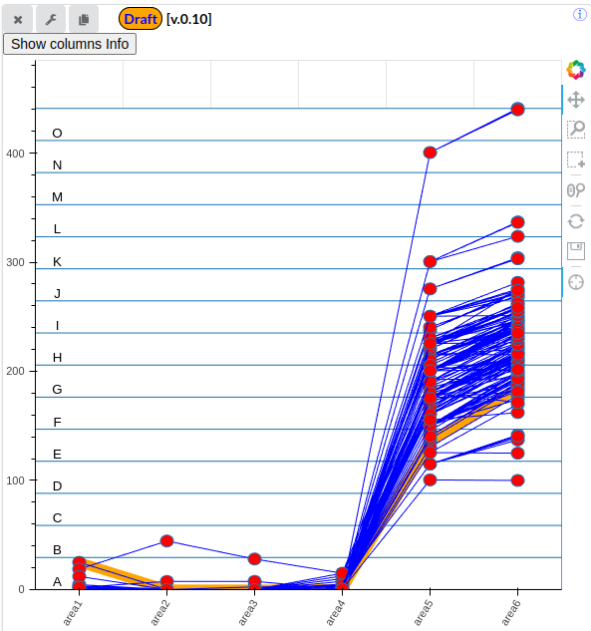


Figure 3: Area name are shown in x axis

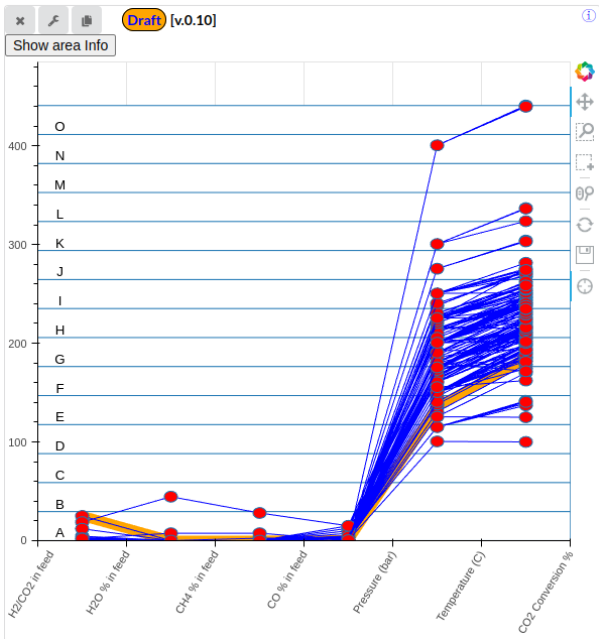


Figure 4: Column name are shown in x axis

3.2.3 Heatmap

In the heatmap, the area values calculated through the introduction of catalyst genes are represented by colors. While the area plot illustrates overall data trends, the heatmap provides detailed information for each catalyst. This visualization enables the analysis of the relationship between color patterns and the distribution of catalysts with similar genes, as well as the integration of this analysis with clustering results. The arrangement of catalysts is based on clustering results, ensuring that similar catalysts are positioned close to each other. As in clustering, the root catalyst is marked in green, while catalysts with similar genes are marked in yellow. You can move, zoom and select data by using tools om the right side. When you select the data, it connect to the Table. In the same way as the Area Plot, clicking the button in the top-left corner displays the items used to calculate each area. Additionally, like Clustering, the threshold for edit distance can also be adjusted using the box in the top-left corner. The Heatmap result example is shown in figure 5

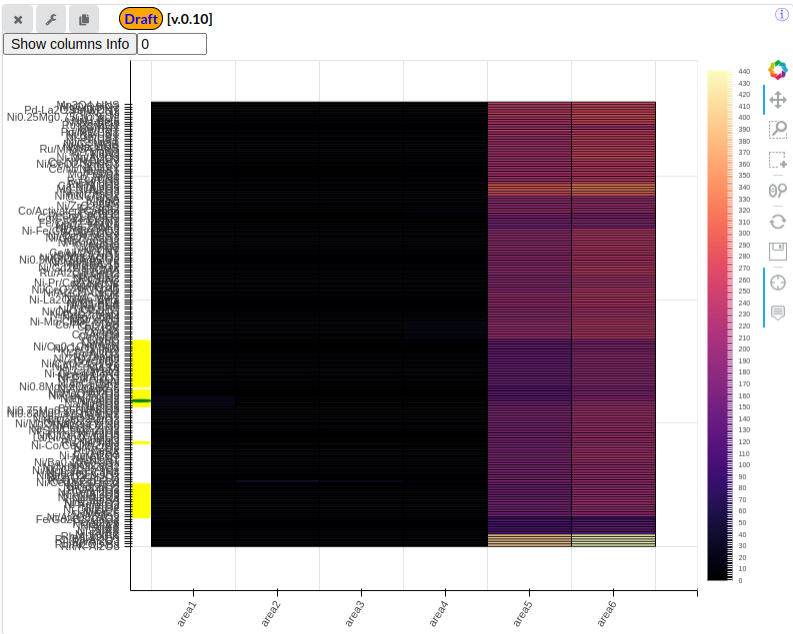
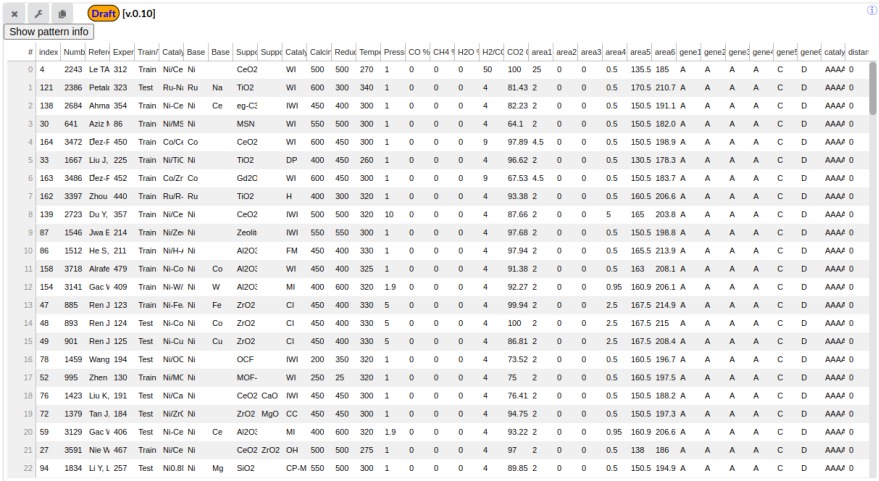


Figure 5: Heatmap visualization. In this result the edit distance threshold was set as 0, and area name are displayed

3.2.4 Table

The Table displays the original data with the addition of catalyst genes and the edit distance from the root catalyst. The data is initially sorted by edit distance from the root catalyst. However, clicking on a column name allows sorting by that column. This Table interacts with the Heatmap. Data selected in the Heatmap is highlighted in yellow in the Table, while selecting data in the Table causes it to be displayed in the Heatmap. By clicking the Show pattern info button in the top-left corner, you can check the combinations of elements that commonly appear in catalysts with an edit distance below the threshold, along with their occurrence frequency. The threshold value can also be adjusted using the box in the top-left corner. By combining this table with clustering and heatmaps, you can identify key patterns of catalysts with similar properties and analyze their relationships. The Table example is shown in figure 6, the pattern table example is shown in figure 7

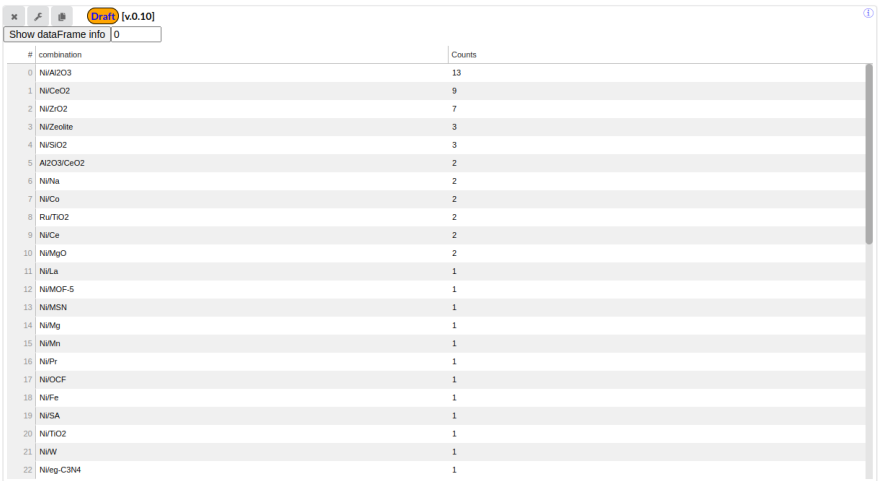


Draft [v.0.10]

Show pattern info

#	index	Numb	Refer	Exper	Train	Catal	Base	Base	Supp	Supp	Catal	Calcir	Reduc	Temp	Press	CO %	CH4 %	H2O %	H2/C	CO2	area1	area2	area3	area4	area5	area6	gene1	gene2	gene3	gene4	gene5	gene6	catalyl	distar
0	4	2243	Le TA	312	Train	Ni/Ce	Ni		CeO2		WI	500	500	270	1	0	0	0	50	100	25	0	0	0.5	135.5	185	A	A	A	A	C	D	AAAA	0
1	121	2386	Petak	323	Test	Ru-Ni	Ru	Na	TiO2		WI	600	300	340	1	0	0	0	4	81.43	2	0	0	0.5	170.5	210.7	A	A	A	A	C	D	AAAA	0
2	138	2684	Ahma	354	Train	Ni-Ce	Ni	Ce	eg-C3		IWI	450	400	300	1	0	0	0	4	82.23	2	0	0	0.5	150.5	191.1	A	A	A	A	C	D	AAAA	0
3	30	641	Aziz M	86	Train	Ni/MS	Ni		MSN		WI	550	500	300	1	0	0	0	4	64.1	2	0	0	0.5	150.5	182.0	A	A	A	A	C	D	AAAA	0
4	164	3472	Đez-F	450	Train	Co/Ce	Co		CeO2		WI	600	450	300	1	0	0	0	9	97.89	4.5	0	0	0.5	150.5	198.9	A	A	A	A	C	D	AAAA	0
5	33	1667	Liu J,	225	Train	Ni/TiC	Ni		TiO2		DP	400	450	260	1	0	0	0	4	96.62	2	0	0	0.5	130.5	178.3	A	A	A	A	C	D	AAAA	0
6	163	3486	Đez-F	452	Train	Co/Zr	Co		Gd2C		WI	600	450	300	1	0	0	0	9	67.53	4.5	0	0	0.5	150.5	183.7	A	A	A	A	C	D	AAAA	0
7	162	3397	Zhou	440	Train	Ru/R-	Ru		TiO2		H	400	300	320	1	0	0	0	4	93.38	2	0	0	0.5	160.5	206.6	A	A	A	A	C	D	AAAA	0
8	139	2723	Du Y,	357	Train	Ni/Ce	Ni		CeO2		IWI	500	500	320	10	0	0	0	4	87.66	2	0	0	5	165	203.8	A	A	A	A	C	D	AAAA	0
9	87	1546	Jwa E	214	Train	Ni/Ze	Ni		Zeolit		IWI	550	550	300	1	0	0	0	4	97.68	2	0	0	0.5	150.5	198.8	A	A	A	A	C	D	AAAA	0
10	86	1512	He S,	211	Train	Ni/H-J	Ni		Al2O3		FM	450	400	330	1	0	0	0	4	97.94	2	0	0	0.5	165.5	213.9	A	A	A	A	C	D	AAAA	0
11	158	3718	Alrafe	479	Train	Ni-Co	Ni	Co	Al2O3		WI	450	400	325	1	0	0	0	4	91.38	2	0	0	0.5	163	208.1	A	A	A	A	C	D	AAAA	0
12	154	3141	Gac V	409	Train	Ni-W/	Ni	W	Al2O3		MI	400	600	320	1.9	0	0	0	4	92.27	2	0	0	0.95	160.9	206.1	A	A	A	A	C	D	AAAA	0
13	47	885	Ren J	123	Train	Ni-Fe	Ni	Fe	ZrO2		CI	450	400	330	5	0	0	0	4	99.94	2	0	0	2.5	167.5	214.9	A	A	A	A	C	D	AAAA	0
14	48	893	Ren J	124	Test	Ni-Co	Ni	Co	ZrO2		CI	450	400	330	5	0	0	0	4	100	2	0	0	2.5	167.5	215	A	A	A	A	C	D	AAAA	0
15	49	901	Ren J	125	Test	Ni-Cu	Ni	Cu	ZrO2		CI	450	400	330	5	0	0	0	4	86.81	2	0	0	2.5	167.5	208.4	A	A	A	A	C	D	AAAA	0
16	78	1459	Wang	194	Test	Ni/OC	Ni		OCF		IWI	200	350	320	1	0	0	0	4	73.52	2	0	0	0.5	160.5	196.7	A	A	A	A	C	D	AAAA	0
17	52	995	Zhen	130	Train	Ni/MC	Ni		MOF-		WI	250	25	320	1	0	0	0	4	75	2	0	0	0.5	160.5	197.5	A	A	A	A	C	D	AAAA	0
18	76	1423	Liu K,	191	Test	Ni/Ca	Ni		CeO2	CaO	IWI	450	450	300	1	0	0	0	4	76.41	2	0	0	0.5	150.5	188.2	A	A	A	A	C	D	AAAA	0
19	72	1379	Tan J,	184	Test	Ni/ZrC	Ni		ZrO2	MgO	CC	450	450	300	1	0	0	0	4	94.75	2	0	0	0.5	150.5	197.3	A	A	A	A	C	D	AAAA	0
20	59	3129	Gac V	406	Test	Ni-Ce	Ni	Ce	Al2O3		MI	400	600	320	1.9	0	0	0	4	93.22	2	0	0	0.95	160.9	206.6	A	A	A	A	C	D	AAAA	0
21	27	3591	Nie W	467	Train	Ni/Ce	Ni		CeO2	ZrO2	OH	500	500	275	1	0	0	0	4	97	2	0	0	0.5	138	186	A	A	A	A	C	D	AAAA	0
22	94	1834	Li Y, L	257	Test	Ni0.8I	Ni	Mg	SiO2		CP-M	550	500	300	1	0	0	0	4	89.85	2	0	0	0.5	150.5	194.9	A	A	A	A	C	D	AAAA	0

Figure 6: Table with Catalyst gene and edit distance



Draft [v.0.10]

Show dataFrame info 0

#	combination	Counts
0	Ni/Al2O3	13
1	Ni/CeO2	9
2	Ni/ZrO2	7
3	Ni/Zeolite	3
4	Ni/SiO2	3
5	Al2O3/CeO2	2
6	Ni/Na	2
7	Ni/Co	2
8	Ru/TiO2	2
9	Ni/Ce	2
10	Ni/MgO	2
11	Ni/La	1
12	Ni/MOF-5	1
13	Ni/MSN	1
14	Ni/Mg	1
15	Ni/Mn	1
16	Ni/Pr	1
17	Ni/OCF	1
18	Ni/Fe	1
19	Ni/SA	1
20	Ni/TiO2	1
21	Ni/W	1
22	Ni/eg-C3N4	1

Figure 7: Table showing pattern information