Supporting Information - Machine learning for catalyst design: data matters

Pedro S. F. Mendes[1,2*†], Florence H. Vermeire[1,3 ‡], Thibaut Van Haute[1], Joris W. Thybaut[1]

[1]Laboratory for Chemical Technology & Department Of Materials, Textiles And Chemical Engineering, Ghent University, Technologiepark 125, 9052 Ghent, Belgium
[2]Department of Chemical Engineering & Centro de Química Estrutural (IMS), Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]Department of Chemical Engineering, KU Leuven, Celestijnenlaan 200F, 3001 Leuven, Belgium

December 19, 2023

*pedro.f.mendes@tecnico.ulisboa.pt
†These two authors contributed equally
‡These two authors contributed equally

S.1 Correlation between OCM features

Artificially introducing correlation in the *in silico* dataset clearly affects the feature importance by permutation. To better understand how such correlation can affect the feature importance of the experimental OCM dataset, the Pearson correlation coefficient between numerical features is calculate and reported in Figure S1. This matrix shows a high correlation between the molar composition of the different metals in the OCM catalyst ($M1\_mol$, $M2\_mol$ and $M3\_mol$). Additional, from our chemical intuition one would expect the sum of the inlet flow rates (Ar_flow, $O2\_flow$ and $CH4\_flow$) to be correlated to the contact time (CT). For this reason, also this sum (sum_flow) is used to calculate the correlation coefficients and a high degree of negative correlation with contact time is observed.

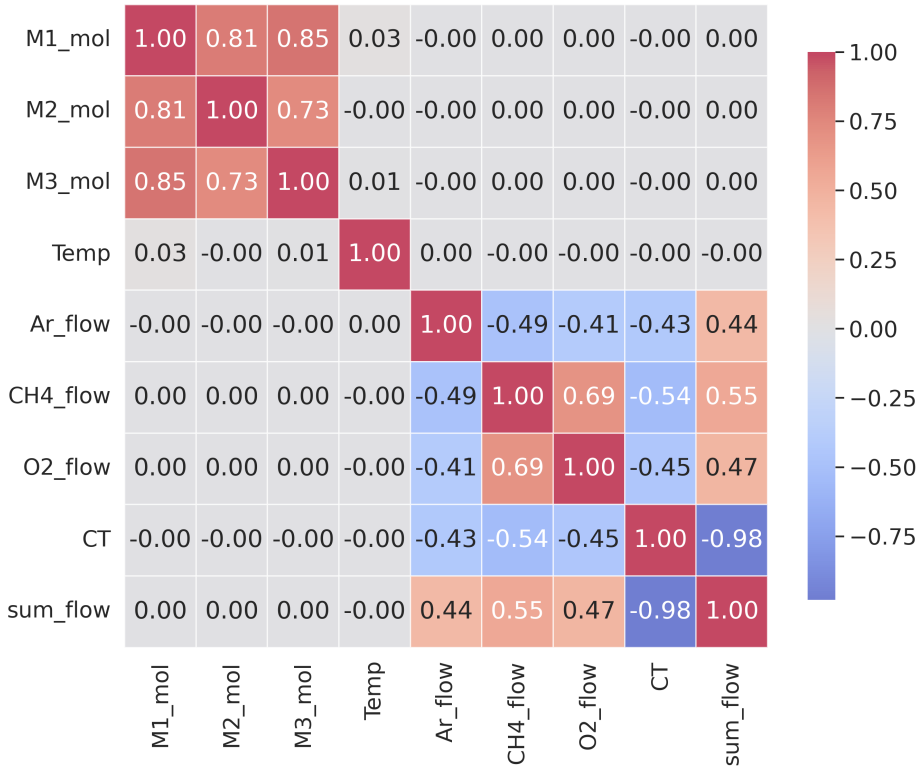

Figure S1: Pearson correlation coefficients between the features in the OCM experimental dataset. The sum of the flow rates (Ar, $O2$ and $CH4$) is added to show the correlation with contact time (CT)

S.2 Yield distribution for changing design space

The feature importance depends strongly on the design space of those features, as demonstrated in Figure 4. However, all features are scaled using the standard scalar when used as input to the model. As such, the actual model input does not change when changing the design space of the features. The dependency of the feature importance on the design space of those features results from the effect of this design space on the output, *i.e.* the yield. Changing the maximum value of the samples features changes the distribution of the yield. The resulting distribution can be seen in Figure S2 for changing the maximum metal concentration (n_m, left) and acid concentration (n_a, right). Higher metal concentration increases the yield, while smaller acid concentration have the same effect.

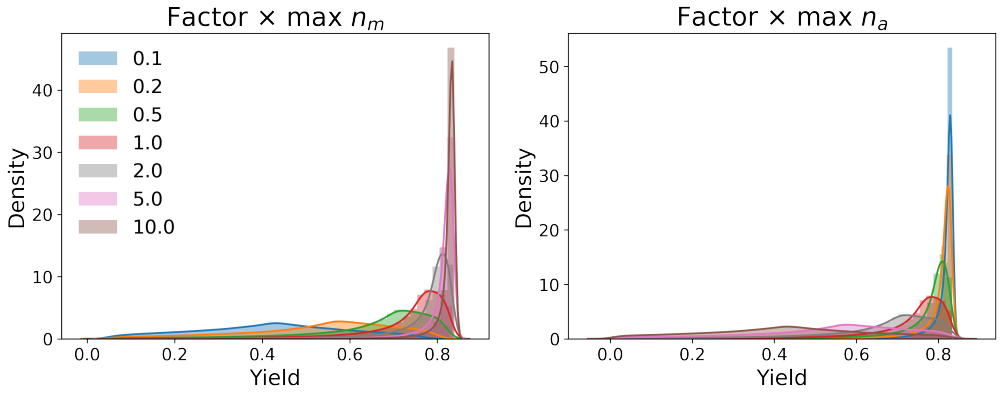

Figure S2: Distribution of the yield when changing the design space of the metal (left) and acid (right) concentration (n_m and n_a).

S.3 Effect of error on model score

There is a clear effect of experimental uncertainty on the model score when noise is applied to the yield. However it is not clear whether this effect results from (1) training the model on noisy data, or (2) evaluating the model score against noisy data. To distinguish between both, the model score is calculated using the predicted values from the model trained on data with noise applied to the yield against the experimental yield with noise (Figure S3, left) and without noise (Figure S3, right). From these results it can be concluded the drop in the model score is due to evaluating the model against noisy data rather than due to poor model performance.

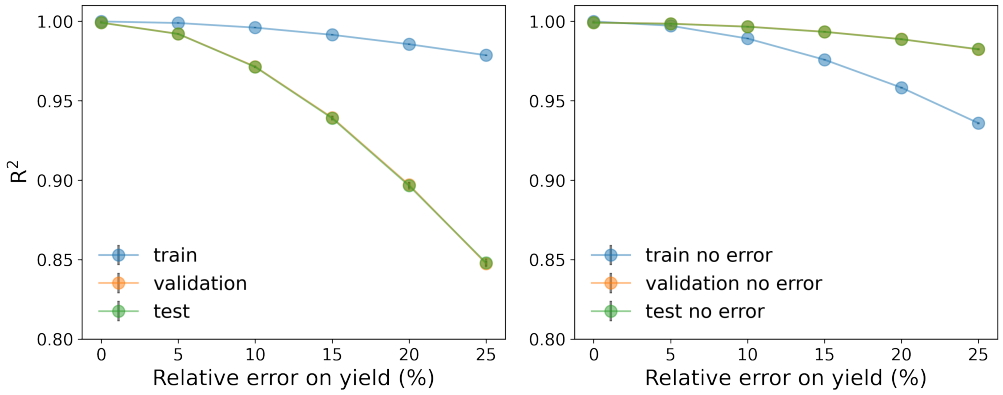

Figure S3: Model score by evaluating the model against data with noise applied to the yield (left) and to data with the true yield (right) in the training, validation and test set. The model is in both cases trained on data with noise applied to the yield.

S.4 Data for figures

Table S1: Model score used in Figure 1, x is the size of the training + validation set

x	**Training**	**Validation**	**Test ensemble**
100	0.99 ± 0.0	0.93 ± 0.12	0.91 ± 0.02
300	0.99 ± 0.0	0.95 ± 0.05	0.95 ± 0.01
500	0.99 ± 0.0	0.96 ± 0.02	0.96 ± 0.0
1000	1.0 ± 0.0	0.98 ± 0.01	0.97 ± 0.0
5000	1.0 ± 0.0	0.99 ± 0.0	0.99 ± 0.0
10000	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
50000	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
100000	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
500000	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

Table S2: Permutation importance used in Figure 1, x is the size of the training + validation set

x	n_m	n_a	γ	ϵ
100	0.09 ± 0.05	0.01 ± 0.0	1.73 ± 0.03	0.01 ± 0.0
300	0.13 ± 0.03	0.01 ± 0.0	1.75 ± 0.02	0.01 ± 0.0
500	0.16 ± 0.02	0.01 ± 0.0	1.74 ± 0.03	0.01 ± 0.0
1000	0.18 ± 0.01	0.02 ± 0.0	1.76 ± 0.02	0.02 ± 0.0
5000	0.21 ± 0.01	0.03 ± 0.0	1.74 ± 0.01	0.03 ± 0.0
10000	0.23 ± 0.01	0.04 ± 0.0	1.74 ± 0.01	0.03 ± 0.0
50000	0.23 ± 0.0	0.04 ± 0.0	1.75 ± 0.01	0.04 ± 0.0
100000	0.23 ± 0.0	0.04 ± 0.0	1.75 ± 0.0	0.04 ± 0.0
500000	0.23 ± 0.0	0.04 ± 0.0	1.75 ± 0.0	0.04 ± 0.0

Table S3: Model score used in Figure 2, x is the size of the training + validation set

x	Training	Validation	Test ensemble
100	0.91 ± 0.02	0.39 ± 0.24	0.3 ± 0.08
300	0.93 ± 0.01	0.48 ± 0.15	0.48 ± 0.03
500	0.94 ± 0.0	0.56 ± 0.11	0.54 ± 0.03
1000	0.95 ± 0.0	0.62 ± 0.07	0.63 ± 0.02
5000	0.97 ± 0.0	0.81 ± 0.02	0.8 ± 0.02
10000	0.98 ± 0.0	0.87 ± 0.01	0.85 ± 0.02

Table S4: Permutation importance used in Figure 2 (part 1), x is the size of the training + validation set

x	M1	M2	M3	Support_ID	M2_mol	M3_mol
100	0.05 ± 0.05	0.01 ± 0.01	0.0 ± 0.01	0.01 ± 0.01	0.07 ± 0.05	0.0 ± 0.0
300	0.04 ± 0.01	0.06 ± 0.06	0.01 ± 0.01	0.03 ± 0.01	0.16 ± 0.1	0.0 ± 0.0
500	0.06 ± 0.02	0.06 ± 0.07	0.02 ± 0.01	0.04 ± 0.02	0.22 ± 0.11	0.01 ± 0.0
1000	0.09 ± 0.02	0.07 ± 0.05	0.04 ± 0.02	0.07 ± 0.02	0.21 ± 0.08	0.02 ± 0.01
5000	0.2 ± 0.01	0.13 ± 0.05	0.1 ± 0.02	0.18 ± 0.02	0.38 ± 0.07	0.04 ± 0.01
10000	0.25 ± 0.01	0.15 ± 0.05	0.12 ± 0.02	0.25 ± 0.02	0.44 ± 0.09	0.06 ± 0.02

Table S5: Permutation importance used in Figure 2 (part 2), x is the size of the training + validation set

x	Temp	Ar_flow	CH4_flow	O2_flow	M1_mol	CT
100	0.43 ± 0.11	0.02 ± 0.02	0.02 ± 0.04	0.05 ± 0.07	0.02 ± 0.04	0.0 ± 0.0
300	0.66 ± 0.06	0.05 ± 0.02	0.01 ± 0.02	0.05 ± 0.03	0.0 ± 0.01	0.0 ± 0.01
500	0.74 ± 0.05	0.07 ± 0.01	0.03 ± 0.01	0.06 ± 0.02	0.01 ± 0.01	0.0 ± 0.0
1000	0.85 ± 0.05	0.1 ± 0.02	0.04 ± 0.01	0.08 ± 0.02	0.01 ± 0.01	0.0 ± 0.0
5000	1.04 ± 0.06	0.15 ± 0.02	0.05 ± 0.01	0.12 ± 0.02	0.02 ± 0.0	0.01 ± 0.0
10000	1.11 ± 0.06	0.19 ± 0.01	0.07 ± 0.01	0.14 ± 0.02	0.02 ± 0.0	0.01 ± 0.0

Table S6: Model score used in Figure 3, x is the number of features removed from the model input

x	Training	Validation	Test
0	0.98 ± 0.0	0.87 ± 0.01	0.87 ± 0.01
1	0.98 ± 0.0	0.87 ± 0.01	0.87 ± 0.01
2	0.98 ± 0.0	0.87 ± 0.01	0.87 ± 0.01
3	0.98 ± 0.0	0.87 ± 0.01	0.87 ± 0.01
4	0.98 ± 0.0	0.86 ± 0.01	0.86 ± 0.01
5	0.91 ± 0.0	0.7 ± 0.02	0.7 ± 0.03
6	0.8 ± 0.0	0.67 ± 0.03	0.68 ± 0.03
7	0.76 ± 0.0	0.65 ± 0.04	0.65 ± 0.02
8	0.58 ± 0.0	0.56 ± 0.02	0.57 ± 0.02
9	0.51 ± 0.0	0.49 ± 0.02	0.49 ± 0.02
10	0.42 ± 0.0	0.42 ± 0.02	0.42 ± 0.02

Table S7: Permutation importance used in Figure 3 (part 1), x is the number of features removed from the model input

x	M1	M2	M3	Support_ID	M2_mol	M3_mol
0	0.25 ± 0.02	0.22 ± 0.04	0.14 ± 0.02	0.26 ± 0.02	0.4 ± 0.02	0.05 ± 0.02
1	0.27 ± 0.02	0.23 ± 0.02	0.13 ± 0.02	0.26 ± 0.02	0.41 ± 0.03	0.05 ± 0.02
2	0.34 ± 0.02	0.23 ± 0.02	0.13 ± 0.03	0.26 ± 0.02	0.42 ± 0.03	0.05 ± 0.02
3	0.35 ± 0.02	0.23 ± 0.02	0.27 ± 0.02	0.26 ± 0.02	0.41 ± 0.03	nan
4	0.35 ± 0.03	0.23 ± 0.03	0.28 ± 0.02	0.27 ± 0.02	0.42 ± 0.03	nan
5	0.16 ± 0.02	0.2 ± 0.02	nan	0.17 ± 0.02	0.37 ± 0.03	nan
6	0.17 ± 0.01	0.2 ± 0.04	nan	0.16 ± 0.02	0.36 ± 0.03	nan
7	0.18 ± 0.02	nan	nan	0.16 ± 0.02	0.47 ± 0.03	nan
8	0.16 ± 0.02	nan	nan	0.14 ± 0.01	0.42 ± 0.03	nan
9	0.14 ± 0.02	nan	nan	nan	0.38 ± 0.02	nan
10	nan	nan	nan	nan	0.34 ± 0.02	nan

Table S8: Permutation importance used in Figure 3 (part 2), x is the number of features removed from the model input

x	Temp	Ar_flow	CH4_flow	O2_flow	M1_mol	CT
0	1.14 ± 0.04	0.2 ± 0.02	0.08 ± 0.01	0.15 ± 0.01	0.03 ± 0.0	0.01 ± 0.0
1	1.11 ± 0.07	0.23 ± 0.01	0.09 ± 0.01	0.14 ± 0.02	0.02 ± 0.0	nan
2	1.13 ± 0.05	0.23 ± 0.01	0.09 ± 0.01	0.14 ± 0.01	nan	nan
3	1.11 ± 0.06	0.23 ± 0.01	0.09 ± 0.01	0.14 ± 0.01	nan	nan
4	1.11 ± 0.05	0.26 ± 0.01	nan	0.16 ± 0.01	nan	nan
5	0.98 ± 0.06	0.2 ± 0.02	nan	0.09 ± 0.02	nan	nan
6	0.97 ± 0.06	0.26 ± 0.01	nan	nan	nan	nan
7	0.93 ± 0.06	0.25 ± 0.01	nan	nan	nan	nan
8	0.71 ± 0.03	nan	nan	nan	nan	nan
9	0.65 ± 0.04	nan	nan	nan	nan	nan
10	0.57 ± 0.03	nan	nan	nan	nan	nan

Table S9: Model score used in Figure 4, x is the factor multiplied by the maximum boundary of the metal concentration (n_m) or acid concentration (n_a) to modify the design space

x	Training	Validation	Test
$n_m \cdot$ 0.1	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_m \cdot$ 0.2	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_m \cdot$ 0.5	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_m \cdot$ 1.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_m \cdot$ 2.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_m \cdot$ 5.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_m \cdot$ 10.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_a \cdot$ 0.1	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_a \cdot$ 0.2	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_a \cdot$ 0.5	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_a \cdot$ 1.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_a \cdot$ 2.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_a \cdot$ 5.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$n_a \cdot$ 10.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

Table S10: Permutation importance used in Figure 4, x is the factor multiplied by the maximum boundary of the metal concentration (n_m) or acid concentration (n_a) to modify the design space

x	n_m	n_a	γ	ϵ
$n_m \cdot$ 0.1	1.13 ± 0.01	0.89 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_m \cdot$ 0.2	1.37 ± 0.01	0.66 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_m \cdot$ 0.5	1.61 ± 0.01	0.46 ± 0.02	0.0 ± 0.0	0.0 ± 0.0
$n_m \cdot$ 1.0	1.74 ± 0.02	0.36 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_m \cdot$ 2.0	1.83 ± 0.02	0.3 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_m \cdot$ 5.0	1.9 ± 0.03	0.26 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_m \cdot$ 10.0	1.93 ± 0.04	0.24 ± 0.02	0.0 ± 0.0	0.0 ± 0.0
$n_a \cdot$ 0.1	1.99 ± 0.02	0.01 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
$n_a \cdot$ 0.2	1.95 ± 0.02	0.08 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
$n_a \cdot$ 0.5	1.85 ± 0.01	0.23 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_a \cdot$ 1.0	1.75 ± 0.02	0.36 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_a \cdot$ 2.0	1.58 ± 0.01	0.5 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_a \cdot$ 5.0	1.29 ± 0.02	0.76 ± 0.01	0.0 ± 0.0	0.0 ± 0.0
$n_a \cdot$ 10.0	1.06 ± 0.01	0.98 ± 0.01	0.0 ± 0.0	0.0 ± 0.0

Table S11: Model score used in Figure 5, x is the maximum relative uncertainty (δ) applied to the metal concentration (n_m), the acid concentration (n_a), and the yield (Y)

x	**Training**	**Validation**	**Test**
$\delta_{n_m}\pm$ 0.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_m}\pm$ 5.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_m}\pm$ 10.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_m}\pm$ 15.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_m}\pm$ 20.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_m}\pm$ 25.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_a}\pm$ 0.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_a}\pm$ 5.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_a}\pm$ 10.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_a}\pm$ 15.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_a}\pm$ 20.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_{n_a}\pm$ 25.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_Y\pm$ 0.0%	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
$\delta_Y\pm$ 5.0%	1.0 ± 0.0	0.99 ± 0.0	0.99 ± 0.0
$\delta_Y\pm$ 10.0%	1.0 ± 0.0	0.97 ± 0.0	0.97 ± 0.0
$\delta_Y\pm$ 15.0%	0.99 ± 0.0	0.94 ± 0.0	0.94 ± 0.0
$\delta_Y\pm$ 20.0%	0.99 ± 0.0	0.9 ± 0.0	0.9 ± 0.0
$\delta_Y\pm$ 25.0%	0.98 ± 0.0	0.85 ± 0.0	0.85 ± 0.0

Table S12: Permutation importance used in Figure 5, x is the maximum relative uncertainty (δ) applied to the metal concentration (n_m), the acid concentration (n_a), and the yield (Y)

x	n_m	n_a	γ	ϵ
$\delta_{n_m}\pm$ 0.0	0.22 ± 0.01	0.04 ± 0.0	1.75 ± 0.01	0.04 ± 0.0
$\delta_{n_m}\pm$ 5.0	0.22 ± 0.01	0.04 ± 0.0	1.74 ± 0.01	0.04 ± 0.0
$\delta_{n_m}\pm$ 10.0	0.22 ± 0.01	0.04 ± 0.0	1.74 ± 0.02	0.04 ± 0.0
$\delta_{n_m}\pm$ 15.0	0.22 ± 0.01	0.04 ± 0.0	1.76 ± 0.01	0.04 ± 0.0
$\delta_{n_m}\pm$ 20.0	0.22 ± 0.01	0.04 ± 0.0	1.75 ± 0.01	0.04 ± 0.0
$\delta_{n_m}\pm$ 25.0	0.22 ± 0.0	0.04 ± 0.0	1.75 ± 0.01	0.04 ± 0.0
$\delta_{n_a}\pm$ 0.0	0.22 ± 0.01	0.04 ± 0.0	1.74 ± 0.01	0.04 ± 0.0
$\delta_{n_a}\pm$ 5.0	0.22 ± 0.01	0.04 ± 0.0	1.75 ± 0.01	0.04 ± 0.0
$\delta_{n_a}\pm$ 10.0	0.22 ± 0.01	0.04 ± 0.0	1.75 ± 0.01	0.04 ± 0.0
$\delta_{n_a}\pm$ 15.0	0.22 ± 0.01	0.04 ± 0.0	1.74 ± 0.01	0.04 ± 0.0
$\delta_{n_a}\pm$ 20.0	0.22 ± 0.01	0.04 ± 0.0	1.75 ± 0.01	0.04 ± 0.0
$\delta_{n_a}\pm$ 25.0	0.22 ± 0.0	0.04 ± 0.0	1.74 ± 0.01	0.04 ± 0.0
$\delta_Y\pm$ 0.0	0.22 ± 0.01	0.04 ± 0.0	1.74 ± 0.01	0.04 ± 0.0
$\delta_Y\pm$ 5.0	0.22 ± 0.01	0.04 ± 0.0	1.73 ± 0.02	0.04 ± 0.0
$\delta_Y\pm$ 10.0	0.22 ± 0.01	0.04 ± 0.0	1.7 ± 0.02	0.04 ± 0.0
$\delta_Y\pm$ 15.0	0.21 ± 0.01	0.04 ± 0.0	1.65 ± 0.02	0.04 ± 0.0
$\delta_Y\pm$ 20.0	0.21 ± 0.01	0.04 ± 0.0	1.57 ± 0.01	0.04 ± 0.0
$\delta_Y\pm$ 25.0	0.19 ± 0.01	0.04 ± 0.0	1.5 ± 0.01	0.03 ± 0.0

Table S13: Model score used in Figure 6, x is the degree of linear correlation between the metal and acid concentration

x	**Training**	**Validation**	**Test**
0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
0.2	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
0.4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
0.6	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
0.8	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
1.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

Table S14: Permutation importance used in Figure 6, x is the degree of linear correlation between the metal and acid concentration

x	n_m	n_a	γ	ϵ
0.0	0.23 ± 0.01	0.04 ± 0.0	1.75 ± 0.02	0.04 ± 0.0
0.2	0.22 ± 0.01	0.03 ± 0.0	1.8 ± 0.01	0.04 ± 0.0
0.4	0.2 ± 0.01	0.02 ± 0.0	1.83 ± 0.02	0.03 ± 0.0
0.6	0.15 ± 0.01	0.01 ± 0.0	1.87 ± 0.02	0.03 ± 0.0
0.8	0.09 ± 0.0	0.0 ± 0.0	1.92 ± 0.01	0.02 ± 0.0
1.0	0.01 ± 0.0	0.01 ± 0.0	1.97 ± 0.01	0.02 ± 0.0