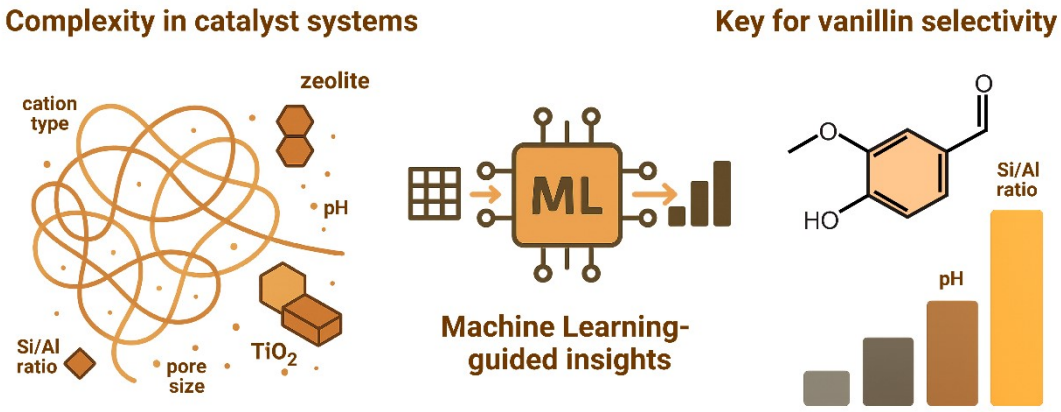

Supplementary Information (SI)

Data-driven design of TiO_2–zeolite photocatalysts for sustainable vanillin production

Daisuke Ino,*[a] Satoru Ito,[a†] Yoshihiro Kon,*[b] Dachao Hong,[c‡] Akira Yada[b] and Kazuhiko Sato[c]

a Technology Division, Panasonic Holdings Corporation. 3-1-1 Yagumo-Nakamachi, Moriguchi City, Osaka 570-8501, Japan; e-mail: ino.daisuke@jp.panasonic.com

b Catalytic Chemistry Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8565, Japan; e-mail: y-kon@aist.go.jp

c Interdisciplinary Research Center for Catalytic Chemistry, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8565, Japan

Contents

S1. Materials: suppliers and product information

P25 (anatase/rutile mixed phase) and P90 (anatase/rutile mixed phase) were obtained from Evonik Industries. FP-6 (anatase/rutile mixed phase) was obtained from Resonac Corporation. TITANIX JA-1 (anatase) was obtained from TAYCA Co., Ltd. Wako food grade (anatase) was commercially obtained from Fujifilm Wako Pure Chemical Corporation. 100808 EMSURE (anatase) was commercially obtained from Merck. 634662 (anatase-TiO_2), 637254 (anatase) and 637262 (rutile) were commercially obtained from Sigma-Aldrich. JRC-TIO-7 (anatase TiO_2) was kindly supplied by the Catalysis Society of Japan. ST-01 (anatase) was obtained from Ishihara Sangyo Kaisha, Ltd. TIO19PB (brookite) was commercially obtained from Kojundo Chemical Laboratory Co., Ltd.

CBV400, CBV720, CBV760, CBV780, CBV2314, CBV5524G, CBV8014 and CBV28014 were commercially obtained from Zeolyst International. HS-642, HS-320, HS-720, A-3, A-4 and F-9 were commercially obtained from Tosoh Corporation. JRC-Z-HY5.5, JRC-Z-Y5.5, JRC-Z-HY5.3, JRC-Z-Y5.3 and JRC-HM-20(5) were kindly supplied by the Catalysis Society of Japan.

4-vinyl-3-methoxyphenol (4-VG; QC-2540) was commercially obtained from Combi-Blocks Inc. Mixture of cis- and trans-isoeugenol (059-01952), Acetonitrile and ethanol were obtained from Fujifilm Wako Pure Chemical Corporation.

S2. Machine learning method and hyperparameter optimization

Machine learning (ML) algorithms applied in this study included Random Forest (RF), Neural Network (NN), Ridge regression (Ri), LASSO regression (LAS), Elastic Net (EN), Ordinary Least Squares regression (OLS), Partial Least Squares regression (PLS) and Support Vector Machine (SVM). All calculations were performed in Python (ver. 3.8.10). The algorithms were implemented using scikit-learn (ver. 1.3.2) and TensorFlow (ver. 2.4.1). A fixed random seed (random_state = 42) was employed for reproducibility.

The dataset consisted of 20 zeolite samples, among which 3 were classified as rare cases (selectivity, Sel. > 30%) and 17 as normal cases. To address the imbalance in the distribution of response values, a stratified 3-fold cross-validation scheme was adopted. Each fold contained exactly one high-selectivity sample together with 5–6 normal samples, ensuring an approximately balanced distribution. This balanced splitting was adopted to avoid biased evaluation due to uneven distribution of critical data points. This stratification was applied consistently to both hyperparameter tuning and out-of-fold (OOF) performance evaluation.

For each ML algorithm, hyperparameters were optimized by grid search within the stratified 3-fold cross-validation framework. The optimal hyperparameters were defined as those maximizing the mean coefficient of determination (R^2) across folds. Final models were subsequently retrained on the entire dataset using the selected hyperparameters. Feature importance was extracted from the RF model, while standardized regression coefficients were derived from linear regression models (Ri, LAS, EN, OLS, PLS).

In the case of the Neural Network (NN) model, we employed a single hidden layer with 20 neurons, L2

regularization (alpha = 0.001), and a maximum of 5000 iterations (Table S1). However, as shown in Table S1, the NN exhibited limited predictive performance (R^2_cv = –0.1053, RMSE_cv = 8.621), indicating overfitting due to the small dataset size. Therefore, NN was included only as a comparative reference, while the main discussion focuses on the Random Forest (RF) model, which demonstrated the most reliable predictive accuracy under the present conditions.

Table S1 Hyperparameter settings and cross-validation performance of the machine learning models: (Ri, LAS, EN, OLS, PLS, SVM, RF, and NN) applied to the zeolite dataset.

Model	Hyperparameter	Search space / settings	Description	Best hyperparameters	R^2_cv	RMSE_cv
Ri	alpha	0.01, 0.1, 1, 10	Regularization strength; controls L2 penalty	10	0.5441	5.913
LAS	alpha	0.001, 0.01, 0.1, 1	Regularization strength; controls L1 penalty	0.1	0.5443	5.834
EN	alpha	0.001, 0.01, 0.1	Overall regularization strength combining L1 and L2 penalties	0.1	0.5494	5.816
	l1_ratio	0.2, 0.5, 0.8	Ratio between L1 (lasso) and L2 (ridge) penalties	0.8		
OLS	-	-	No hyperparameters	-	-0.4594	10.41
PLS	n_components	1, 2	Number of latent components to retain in Partial Least Squares regression	2	0.5102	6.177
RF	n_estimators	50, 100, 150, 200	Number of trees in the forest	200	0.7287	4.603
	max_depth	None, 3, 5	Maximum depth of each tree (None means fully expanded)	None		
SVM	C	0.1, 1, 10	Regularization parameter controlling trade-off between training error and margin	1	0.4732	6.246
	epsilon	0.01, 0.1, 0.5	Width of the epsilon-insensitive tube in regression; defines tolerance of error	0.1		
NN	hidden_layer_sizes	(10,), (20,)	Number of neurons in hidden layers	20	-0.1053	8.621
	alpha	0.0001, 0.001	L2 regularization parameter (weight decay)	0.001		
	max_iter	5000	Maximum number of iterations for training	-		

S3. Training and test data sets for machine learning analysis

To construct predictive models, we employed a dataset of zeolite descriptors and vanillin selectivity values obtained from the photocatalytic oxidation experiments. Six physicochemical descriptors were considered: Si/Al ratio, specific surface area, pore size, suspension pH, cation type, and framework type.

A total of 20 zeolite samples were available. Among them, 16 samples were used as the training dataset for model development (Table S2), while 4 samples were reserved as the independent test dataset for external validation (Table S3). This division enabled us to evaluate the predictive performance of machine learning (ML) models under conditions not seen during training.

Table S23 Training dataset for machine learning analysis: experimental vanillin selectivity and physicochemical descriptors of various zeolites. The dataset includes 16 zeolite samples used to construct predictive models for vanillin selectivity in the photocatalytic oxidation of 4-vinylguaiacol (4-VG). Six physicochemical descriptors —including Si/Al ratio, specific surface area, pore size, pH, cation type, and framework type—served as explanatory variables, while vanillin selectivity was the response variable. Conv. = Conversion. Sel. = Selectivity.

Entry	Zeolite	Si/Al	Surface area	Pore size	pH	Cation				Zeolite type					Conv.	Yield	Sel.
		(mol/mol)	(m^2/g)	(nm)		H^+	NH_4^+	Na^+	K^+	MOR	FER	FAU	MFI	LTA	(%)	(%)	(%)
1	HS-642	9	360	7	9.7	0	0	1	0	1	0	0	0	0	16	2.6	16
2	HS-320	2.8	700	7.4	8.4	0	0	1	0	0	0	1	0	0	15	2.5	17
3	HS-720	9	170	4.8	8.2	0	0	0	1	0	1	0	0	0	20	3	15
4	CBV400	2.6	730	9	5.4	1	0	0	0	0	0	1	0	0	99	6.1	6.1
5	CBV780	40	780	9	5.1	1	0	0	0	0	0	1	0	0	42	4.7	11
6	CBV2314	11.5	425	5.8	7.5	0	1	0	0	0	0	0	1	0	76	6.9	9.1
7	CBV5524G	25	425	5.8	6.4	0	1	0	0	0	0	0	1	0	41	5.7	14
8	CBV8014	40	425	5.8	6.2	0	1	0	0	0	0	0	1	0	32	4.8	15
9	CBV28014	140	425	5.8	6.4	0	1	0	0	0	0	0	1	0	23	4.1	18
11	A-3	1	230	4.2	10.5	0	0	0	1	0	0	0	0	1	99	36	36
12	F-9	1.3	525	9	10.5	0	0	1	0	0	0	1	0	0	91	32	35
13	JRC-Z-HM-20	18.3	360	7	4.5	1	0	0	0	1	0	0	0	0	29	4.6	16
14	JRC-Z-HY5.5	5.6	570	9	7.5	1	0	0	0	0	0	1	0	0	99	5.2	5.2
15	JRC-Z-Y5.3	5.3	728	9	8.3	0	0	1	0	0	0	1	0	0	21	4.0	19
16	JRC-Z-HY5.3	5.3	692	9	8.7	1	0	0	0	0	0	1	0	0	99	9.2	9.2

Table S3 Test dataset for machine learning analysis: experimental and predicted vanillin selectivity along with physicochemical descriptors of various zeolites. The dataset contains four zeolite samples excluded from the training set, listing their structural and physicochemical descriptors together with experimental selectivity (Sel.Ex) and predicted selectivity (Sel.ML) obtained from the RF (Random Forest) model. These data were used to validate the predictive accuracy and generalization performance of the constructed models. The predicted values closely matched the experimental results (maximum deviation within ±2 %), confirming the robustness of the RF model. Conv. = Conversion. Sel. = Selectivity.

Entry	Zeolite	Si/Al	Surface area	Pore size	pH	Cation				Zeolite type					Conv.	Yield	Sel. Ex.	Sel.ML
		(mol/mol)	(m^2/g)	(nm)		H^+	NH_4^+	Na^+	K^+	MOR	FER	FAU	MFI	LTA	(%)	(%)	(%)	(%)
1	CBV720	15	780	9	4.2	1	0	0	0	0	0	1	0	0	16	1.6	10	10
2	CBV760	30	720	9	4.1	1	0	0	0	0	0	1	0	0	15	1.7	11	11
3	JRC-Z-Y5.5	5.6	660	9	7.9	0	0	1	0	0	0	1	0	0	20	2.8	14	14
4	A-4	1	290	4.2	10.3	0	0	0	1	0	0	0	0	1	99	34	34	31

S4. Machine learning-based correlation analysis between zeolite properties and vanillin selectivity (Additional analysis by Ri and LAS models)

Ridge Regression (Ri) exhibited a larger root mean squared error (RMSE) than Random Forest (RF) as shown in Fig. 2A (main text) but retained nonzero coefficients for correlated variables. In this model, both Si/Al ratio and pH received nonzero coefficients, indicating that these descriptors are important for vanillin selectivity. Because Si/Al ratio and pH are strongly correlated within the present dataset, the positive Si/Al coefficient should be interpreted with caution and does not imply that higher Si/Al is beneficial; rather, it reflects multicollinearity between these descriptors. This is consistent with the RF analysis and experimental observations, which indicate that lower Si/Al ratios, combined with stronger basic buffering (higher suspension pH), are associated with improved selectivity. Notably, the appearance of the LTA framework is consistent with the RF analysis as shown in Fig. 2C (main text), whereas the FAU framework type did not clearly appear in either RF or Ri.

In the case of LTA, a lower Si/Al ratio generates abundant AlO_4^- sites that enhance ion-exchange capacity. The three-dimensional channel system of LTA zeolites further promotes rapid ion migration, leading to fast adsorption kinetics, suggesting that the internal AlO_4^- sites can act as basic buffer sites, whereas this does not appear to be the case for FAU. This result is considered evidence that the ML did not merely memorize the training data but actually learned to identify the relevant zeolite descriptor.

Pore size appeared as a positive descriptor, although with only a minor contribution, suggesting a possible relation to diffusion or local concentration effects rather than a direct control of selectivity. Na^+, and K^+ also appeared with positive coefficients, but these factors were not clearly highlighted by RF.

In contrast, H^+ and surface area showed negative coefficients. The negative weight of H^+ is consistent with the detrimental effect of Brønsted acidity observed for H-type zeolites. The negative coefficient of surface

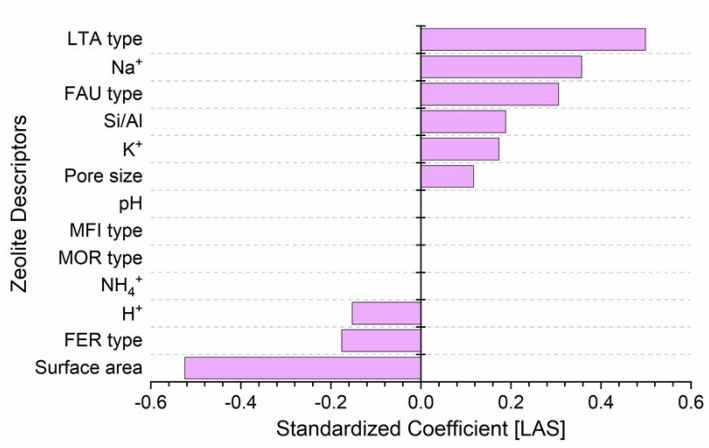

Fig. S1 Standardized regression coefficients of zeolite descriptors in the LASSO Regression (LAS) model.

area may indicate that a larger surface area, dominated by microporous regions, does not directly contribute to effective buffering or enhanced selectivity, because cation sites located deep inside pores are less accessible to the solution. These results collectively indicate that, despite lower predictive accuracy, linear models provide complementary information by clarifying the direction of descriptor contributions and support the central role of Si/Al ratio and pH in selectivity control.

LASSO Regression (LAS) yielded a relatively large RMSE shown in Fig. 2A (main text) similar to Ri, the coefficient analysis was nevertheless informative. Fig. S1 shows Si/Al ratio, Na^+, and LTA framework showed positive coefficients, while H^+ and surface area, exhibited negative coefficients. Notably, the weight of pH was reduced to nearly zero. This behavior is likely due to the strong correlation between pH and Si/Al ratio: the LASSO regularization selected Si/Al as a representative descriptor of basicity, and the exact sign of its coefficient should not be overinterpreted. Together with the experimental observations, the RF analysis, and the Ri regression, these findings converge on a consistent interpretation: the buffering effect arising from framework composition (low Si/Al) and exchangeable cations plays a central role in stabilizing the suspension pH and thereby governing vanillin selectivity.