Supplementary Information (SI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2026

Supporting Information

Linh Thi Hoai Nguyen,[†] Edoardo Fabbrini,[‡] Andriy Olenko,[¶] Aleksandar Staykov,[§] and Pierluigi Cesana[*,‖]

†*Center for Energy Systems Design (CESD), International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, Japan*

‡*SACRA, Graduate School of Science, Kyoto University, Japan*

¶*School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, Australia*

§*International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, Japan*

‖*Institute of Mathematics for Industry (IMI), Kyushu University, Japan*

E-mail: nguyen.thi.hoai.linh.578@m.kyushu-u.ac.jp;fabbrini.edoardo.2w@kyoto-u.ac.jp;A.Olenko@latrobe.edu.au;cesana@math.kyushu-u.ac.jp

1 Dataset

The dataset used in this study comprises 2,098 molecules generated in[10] through a systematic exploration of the chemical space consisting of molecules with up to three mutations applied to the seven backbone structures considered in this study (see Figure 1 in the article). The percentages of mutations at each backbone site are summarized in Table 1.

This dataset forms the basis for training the machine learning models integrated into the automated framework.

Table 1: Statistics for percentages of mutations per site in the dataset.

backbone structure	$site_0$	$site_1$	$site_2$	$site_3$	$site_4$	$site_5$	$site_6$	$site_7$	$site_8$	$site_9$	$site_{10}$
5–1	32.69	45.67	42.31	32.21	27.88	44.71	38.94	0.00	0.00	0.00	0.00
5–2	53.05	57.72	0.00	50.20	44.92	61.18	0.00	0.00	0.00	0.00	0.00
6–1	36.36	27.27	35.45	19.55	27.27	20.00	25.91	43.18	35.45	0.00	0.00
6–2	35.82	52.30	0.00	38.70	28.16	28.54	23.95	55.75	0.00	0.00	0.00
6–3	47.17	44.81	0.00	53.30	0.00	51.42	0.00	68.87	0.00	0.00	0.00
7–1	40.71	32.74	24.34	27.88	23.45	19.91	0.00	26.55	0.00	30.09	41.59
7–2	35.78	23.85	0.00	21.10	0.00	30.28	29.36	23.39	22.02	46.33	34.86

2 Role of Mutation Sites as Predictive Features

For each molecular class (5-1, 5-2, 6-1, 6-2, 6-3, 7-1, and 7-2), we fit regression models with quadratic terms using the substituent constants at mutation sites (indexed from 0 to 10) as explanatory variables. For each molecular class, the best-performing model was selected using the Akaike Information Criterion (AIC), evaluated via leave-one-out cross-validation. Table 2 reports the number of times each explanatory variable appears in the best-performing (AIC-selected) regression models. We observe that the variable corresponding to mutation site 0 appears in the regression models of all seven molecular classes. Similarly, Table 2 highlights the importance of nonlinear effects associated with the substituent constant at site 0 (i.e., the $site_0^2$ term), which is included in the optimal regression model for every molecular class.

Table 2: Explanatory variables with a count of at least 2.

Variable	$site_0$	$site_1$	$site_3$	$site_4$	$site_5$	$site_7$	$site_9$	$site_0^2$	$site_1^2$	$site_3^2$
Count	7	5	4	2	3	3	2	7	3	4
Variable	$site_4^2$	$site_5^2$	$site_7^2$	$site_{10}^2$	$site_0 : site_7$		$site_7 : site_8$			
Count	2	3	2	2	2		2			

3 Substituent Constant

The Hammett equation, originally introduced by Hammett in 1937[11], establishes a linear free-energy relationship that correlates reaction rates and equilibrium constants of substituted benzoic acid derivatives through two empirical parameters: the substituent constant (σ) and the reaction constant (ρ). The σ parameter quantifies the electronic influence of a substituent relative to hydrogen, where positive values correspond to electron-withdrawing and negative values to electron-donating behavior.

The substituent constant reflects the combined contribution of two fundamental electronic effects: (a) inductive effects, transmitted through σ bonds as a result of electronegativity differences, and (b) resonance effects, transmitted through π-conjugation. The relative magnitude of these effects depends on the substituent position. Meta substituents predominantly exhibit inductive influences because they are not conjugated with the carboxyl π system, whereas para substituents contribute through both inductive and resonance interactions. As a result, para substituents generally exert stronger and more variable electronic perturbations on the aromatic ring compared to meta substituents.

The electronic character captured by the Hammett σ parameter is closely associated with the energies of the frontier molecular orbitals. Electron-donating substituents (negative σ) typically increase the HOMO energy by donating electron density into the π system, whereas electron-withdrawing substituents (positive σ) generally lower both HOMO and LUMO energies, with the LUMO often being more strongly stabilized[33,40]. In π-conjugated systems, para substituents capable of resonance coupling directly modulate the delocalized π orbitals, producing significant shifts in the frontier orbital energies.

In the present study, the para substituent constants are employed as quantitative electronic descriptors to construct feature vectors for machine learning models aimed at predicting the HOMO–LUMO energy gap of diarylethene derivatives.

4 Machine Learning Model Construction and Interpretation

4.1 Model selection

We report a list of models we trained for comparison in Table 3. While R^2 is included for completeness, it is acknowledged that this metric may not be fully reliable for assessing nonlinear models, and it is evident that RMSE gives much accurate results.

The analysis demonstrated that both RMSE and R^2 values remained consistently stable across multiple iterations and between the training and test datasets. Furthermore, models trained on the full dataset were much better than those trained on data from individual molecules. This improvement can be attributed to the significantly smaller size of single-type backbone structure datasets (approximately sevenfold smaller) and the broader variability in descriptor space that incorporates multiple backbone structures.

Table 3: Performance of machine learning models on the entire dataset

Model	Train RMSE	Train R^2	Test RMSE	Test R^2
Multiple linear regression	0.6139	0.3313	0.6235	0.3283
Multiple linear regression model with interaction and quadratic terms	0.3412	0.7934	0.3444	0.7952
Bayesian generalized linear model	0.3271	0.8111	0.3300	0.8081
Artificial Neural Network (ANN) with one hidden layer (5 nodes)	0.2699	0.8823	0.2886	0.8658
ANN with two hidden layers (50 nodes each)	0.3439	0.8230	0.3503	0.8162
K-nearest neighbours (KNN)	0.1934	0.9344	0.2835	0.8585
Support vector machine (SVM)	0.2477	0.8926	0.2737	0.8684
Ridge regression	0.3271	0.8111	0.3300	0.8081
Random forest	0.1333	0.9690	0.2371	0.9010
XGBoost	0.1823	0.9415	0.2368	0.9014

4.2 Fitting and interpretation of machine learning model

To obtain the optimal hyperparameters for the XGB algorithm, the Bayesian optimisation for hyperparameter tuning was employed. The search ranges for each parameter are listed in the “Range" column in Table 4. For each case, the hyperparameters were tuned separately. The corresponding optimal values are presented in the last two columns of Table 4.

Table 4: XGBoost hyperparameter search ranges optimised via Bayesian tuning.

Hyperparameter	**Range**	**With group labels**	**Without group labels**
learning rate	(0.01, 0.5)	0.328	0.237
gamma	(0, 10)	0.104	0.069
max depth	(3, 15)	13	15
min child weight	(3, 10)	4	7
subsample	(0.5, 1)	0.930	0.829
colsample by tree	(0.1, 1)	0.997	0.969

Alongside RMSE, MAE, and R^2, the mean absolute percentage error (MAPE) was employed to evaluate the performance of the optimised XGB models on both training and test sets for each case. The results are summarised in Table 5.

Table 5: Performance of machine learning models on training and test sets.

		Train				Test			
Missing values	Label	RMSE	MAE	MPE(%)	R_2	RMSE	MAE	MPE(%)	R_2
NaN	Yes	0.196	0.146	2.833	0.933	0.231	0.172	3.326	0.903
NaN	No	0.176	0.130	2.510	0.946	0.238	0.173	3.335	0.897
0	Yes	0.242	0.183	3.517	0.897	0.252	0.187	3.609	0.884
0	No	0.404	0.312	5.889	0.714	0.464	0.374	7.066	0.608

Table 5 shows the comparison of two different Strategies, I and II (described in Section 5.1 of the main article).