This article can be cited before page numbers have been issued, to do this please use: T. Shimazaki and M. Tachikawa, *Phys. Chem. Chem. Phys.*, 2026, DOI: 10.1039/D5CP03946K.

Volume 19
Number 1
7 January 2017
Pages 1-896
PCCP
Physical Chemistry Chemical Physics
rsc.li/pccp
ISSN 1463-9076
ROYAL SOCIETY OF CHEMISTRY
PAPER
H.-P. Loock *et al.*
Determination of the thermal, oxidative and photochemical degradation rates of scintillator liquid by fluorescence EEM spectroscopy

View Article Online
DOI: 10.1039/D5CP03946K

# Theoretical study on the analyzability of modified convex regression for radical reaction

Tomomi Shimazaki and Masanori Tachikawa

Quantum Chemistry Division, Yokohama City University, Seto 22-2, Kanazawa-Ku,
Yokohama 236-0027, Kanagawa, Japan.
tshima@yokohama-cu.ac.jp

View Article Online
DOI: 10.1039/D5CP03946K

## Abstract

Analyzing data and extracting meaningful insights is essential across various research fields. To address acrylate and methacrylate radical reaction data, we propose a modified convex clustering (regression) method, in which representative points are directly selected from the training data to describe the dataset. Although machine learning (ML) models are often regarded as black boxes, making their predictions difficult to interpret, the (modified) convex clustering approach allows for straightforward analysis of model behavior. This study emphasizes the importance of selecting representative points to enhance the interpretability and transparency of ML models. We demonstrate that radical reaction energy barriers can be effectively described and predicted based on the contributions of similar reactions. The simplicity and transparency of the modified convex clustering (regression) method enable in-depth analysis of physicochemical data.

View Article Online
DOI: 10.1039/D5CP03946K

## 1. Introduction

Machine learning (ML) approaches, combined with large-scale molecular databases, have gained considerable prominence across various chemical fields, including functional molecules, drug discovery, and materials science.[1-21] However, despite its advantages, ML presents certain drawbacks. For instance, it is often difficult to comprehend how ML models operate, largely due to the complexity and lack of transparency in their prediction processes. Predictions based on correlations rather than causal relationships can further hinder interpretability. Although ML models excel at capturing multidimensional and nonlinear correlations, such complexity often exceeds human intuitive understanding, making it challenging to grasp underlying mechanisms. This lack of interpretability has limited the broader application of ML, particularly in extracting meaningful insights from physicochemical data. Although explainable artificial intelligence techniques—such as permutation importance and SHAP (SHapley Additive exPlanations)—are sometimes employed, the inherent complexity of ML models continues to obscure a comprehensive understanding of their behavior.[22-27] To address this issue, we focus on algorithms that facilitate interpretability. Particularly, we highlight the convex clustering algorithm, which offers several advantageous features, such as sophisticated data clustering, soft assignment, and direct representative point selection from training data.[28]

In a convex clustering algorithm, the distributions of all clusters (or classes) are defined a priori by a single shared parameter, resulting in uniform distribution sizes. In this study, we slightly modify the algorithm by relaxing this constraint: the distribution size of each cluster is governed by its own parameter, which is automatically determined during the model training process. The modified clustering process is

View Article Online
DOI: 10.1039/D5CP03946K

categorized as a soft assignment method, where each data point can partially belong to multiple classes, with probabilistic ratios representing its degree of membership. Further, we employ the modified clustering method to perform regression on radical reaction data for acrylate (ACR) and methacrylate (MA). Radical reactions involving ACR and MA are widely used in synthesizing various acrylic polymers, such as plastics, adhesives, paints, medical materials, and fibers.[29-31] To theoretically elucidate radical reaction mechanisms, transition state (TS) analyses based on density functional theory (DFT) are indispensable. DFT-based TS calculations have provided insights into reaction processes,[32, 33] and ML has been recently employed to predict complex features such as energy barriers and regioselectivity in radical reactions.[34, 35] We develop ML models to predict the energy barriers of ACR and MA radical reactions using a dataset derived from DFT calculations, and demonstrate that the number of representative points governs both the resolution of data reproduction and the prediction accuracy. This study highlights the simplicity of the convex clustering (regression) approach, particularly in selecting representative points from training data, as a key factor in enhancing the interpretability and transparency of ML models for analyzing physicochemical data.

The remainder of this article is organized as follows. In Section 2, we describe the modified convex clustering algorithm. Section 3 presents the fundamental behavior of the algorithm based on a simple dataset as well as its application to ACR and MA radical reaction data. Finally, Section 4 summarizes this study.

View Article Online
DOI: 10.1039/D5CP03946K

## 2. Methods and Computational Conditions

### 2-1. Convex clustering method

Convex clustering is an unsupervised ML algorithm designed to assign data points to clusters. It is theoretically grounded in the Gaussian mixture model (GMM) and is classified as a soft assignment method, wherein each data point is probabilistically associated with multiple clusters.[36, 37] Despite this probabilistic framework, a definitive cluster assignment can be made by selecting the cluster with the highest associated probability. Meanwhile, hard assignment algorithms, such as K-means clustering, assign each data point to only one cluster.[36, 37] We present a modified K-means clustering algorithm (K-near) in Section S1 of the Supporting Information as a representative hard assignment approach. Although the modified convex clustering and K-near methods represent soft and hard assignment techniques, respectively, both select representative points from the training data to describe the underlying data structure or class distribution.

First, we describe the GMM to highlight the distinctive characteristics of the convex clustering method. We consider a dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ comprising $N$ data points, where each data point is represented by a $d$-dimensional vector. We assume that each data point is generated from a single class through probabilistically independent sampling (trials), although the specific class from which each point originates is unknown. Under these assumptions, the probability of observing the dataset, denoted as $p_{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, can be expressed as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{k=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{x}_k) \tag{1-1}$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}_k) = \sum_{i}^{C} \pi_i p_{\theta_i}(\mathbf{x}_k | \Pi_i) \tag{1-2}$$

View Article Online
DOI: 10.1039/D5CP03946K

$$\sum_{i}^{C} \pi_i = 1, \tag{1-3}$$

where $\boldsymbol{\theta}$ represents parameters of the probabilistic model, $p_{\boldsymbol{\theta}}(\mathbf{x}_k)$ denotes the probability of observing a data point of $\mathbf{x}_k$, $p_{\theta_i}(\mathbf{x}_k | \Pi_i)$ denotes the conditional probability of observing the data point $\mathbf{x}_k$ given that the class $\Pi_i$ was selected under the parameter $\theta_i$, $\pi_i$ denotes the probability (or proportion) associated with class $\Pi_i$, and $C$ denotes the total number of classes. Eq. (1-3) indicates that the sum of all class proportions equals 1.0. Based on the assumption of probabilistic independence, the probability $p_{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ can be expressed as the product of the probabilities for each data point. In the GMM framework, we assume the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{\Sigma}_i$ for class $\Pi_i$ with respect to $\theta_i$. Accordingly, $p_{\theta_i}(\mathbf{x}_k | \Pi_i)$ can be computed as follows:

$$\begin{aligned} p_{\theta_i}(\mathbf{x}_k | \Pi_i) &= p_{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i}(\mathbf{x}_k | \Pi_i) \\ &= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left( -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_i) \right). \end{aligned} \tag{2}$$

Given the number of classes $C$, the parameters $\pi_i$, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ can be determined using the expectation–maximization (EM) algorithm by maximizing the log-likelihood of $\log p_{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.[36, 37] Notably, the number of classes $C$ becomes a predefined parameter (hyperparameter) in the GMM. Meanwhile, the convex clustering method allows for automatic determination of $C$, although the covariance matrix $\boldsymbol{\Sigma}_i$ must be specified in advance.

View Article Online
DOI: 10.1039/D5CP03946K

In the convex clustering algorithm, the representative point $\boldsymbol{\mu}_i$ is selected from a point $\mathbf{x}_i$ in the training dataset, and $\boldsymbol{\Sigma}_i$ is assumed to be a simple diagonal matrix, as follows:

$$\boldsymbol{\mu}_i = \mathbf{x}_i \tag{3-1}$$

$$\boldsymbol{\Sigma}_i = \sigma \mathbf{I}_d, \tag{3-2}$$

where $\mathbf{I}_d$ is the $d$ -dimensional identity matrix. From Eq. (3-2), all classes share the same distribution, which is controlled by the hyperparameter $\sigma$. Under these conditions, the relationship $p_{\theta_i}\left(\mathbf{x}_k \middle| \Pi_i\right) = p_{\mathbf{x}_i,\sigma}\left(\mathbf{x}_k \middle| \Pi_i\right) = f_{i,\sigma}\left(\mathbf{x_k}\right)$ can be derived using the following equation:

$$f_{i,\nu}\left(\mathbf{x}\right) \equiv \frac{1}{\left(2\pi\nu^2\right)^{d/2}} \exp\left(-\frac{1}{2\nu^2}\left|\mathbf{x}-\mathbf{x}_i\right|^2\right). \tag{4}$$

Therefore, the log-likelihood of the convex clustering method can be explicitly expressed as follows:

$$\log p_{\boldsymbol{\theta}}\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) = \sum_k^N \log\left(\sum_i^N \pi_i p_{\boldsymbol{\mu}_i,\sigma}\left(\mathbf{x}_k \middle| \omega_i\right)\right) = \sum_k^N \log\left(\sum_i^N \pi_i f_{i,\sigma}\left(\mathbf{x}_k\right)\right). \tag{5}$$

The EM algorithm provides a self-consistent procedure to determine $\pi_i$. The log-likelihood (Eq. (5)) increases monotonically and converges through an iterative loop in which the values of $\pi_i^{(next)}$ at each step are updated based on the current $\pi_i$ values as follows:

$$\pi_i^{(next)} = \frac{1}{N}\sum_{k=1}^N P_i\left(\mathbf{x}_k\right) \tag{6-1}$$

$$P_i\left(\mathbf{x}\right) \equiv P\left(\Pi_i \middle| \mathbf{x}\right) \equiv \frac{\pi_i p_{\mathbf{x}_i,\sigma}\left(\mathbf{x} \middle| \Pi_i\right)}{\sum_{j=1}^N \pi_j p_{\mathbf{x}_j,\sigma}\left(\mathbf{x} \middle| \Pi_j\right)} = \frac{\pi_i f_{i,\sigma}\left(\mathbf{x}\right)}{\sum_{j=1}^N \pi_j f_{j,\sigma}\left(\mathbf{x}\right)}, \tag{6-2}$$

View Article Online
DOI: 10.1039/D5CP03946K

where $P\left(\Pi_i\left|\mathbf{x}\right.\right)$ denotes the probability of class $\Pi_i$ given that the data point $\mathbf{x}$ was observed. In the self-consistent loop, several data points acquire very small $\pi_i$ values. Points with $\pi_i$ values below a certain threshold are excluded from the set of representative candidates; thus, they no longer contribute to the clustering process. To exclude such data points, their $\pi_i$ values are set to zero. As a result, the number of representative points gradually decreases during the self-consistent loop. The remaining points after the convergence are used as the final representatives for clustering. Notably, the number of remaining points, which corresponds to the number of clusters, depends on the hyperparameter $\sigma$. For example, when a small value is assigned to $\sigma$, a larger number of data points tend to survive the self-consistent procedure.

### 2-2. Modified convex clustering method

In the convex clustering method, the hyperparameter $\sigma$ uniformly defines the distributions of all clusters, as seen in Eq. (3-2). In this study, we modify this distributional assumption as follows:

$$\mathbf{\Sigma}_i = \sigma_i \mathbf{I}_d . \tag{7}$$

Eq. (7) allows each class to have its own distribution corresponding to its specific $\sigma_i$. Thus, the conditional probability $p_{\theta_i}\left(\mathbf{x}_k\left|\Pi_i\right.\right)$ can be expanded as follows:

$$f_{i,\sigma_i}\left(\mathbf{x}\right) = \frac{1}{\left(2\pi\sigma_i^{\ 2}\right)^{d/2}} \exp\left(-\frac{1}{2\sigma_i^{\ 2}}\left|\mathbf{x}-\mathbf{x}_i\right|^2\right), \tag{8}$$

where $p_{\theta_i}\left(\mathbf{x}_k\left|\Pi_i\right.\right) = p_{\mathbf{x}_i,\sigma_i}\left(\mathbf{x}_k\left|\Pi_i\right.\right) = f_{i,\sigma_i}\left(\mathbf{x_k}\right)$. Therefore, the conditional probability

View Article Online
DOI: 10.1039/D5CP03946K

$P_i(\mathbf{x})$ is slightly modified as follows:

$$P_i(\mathbf{x}) \equiv P\left(\Pi_i \middle| \mathbf{x}\right) = \frac{\pi_i f_{i,\sigma_i}(\mathbf{x})}{\sum_{j=1}^{N} \pi_j f_{j,\sigma_i}(\mathbf{x})}. \tag{9}$$

To determine the $\sigma_i$ value, we examine the following conditions:

$$\sigma_i = \frac{1}{2}\min\{L_{i1}, L_{i2}, \cdots, L_{il}, \cdots\}, \tag{10}$$

where $L_{il}$ denotes the distance between representative points $\mathbf{x}_i$ and $\mathbf{x}_l$. Accordingly, $\sigma_i$ is set to half the distance to the nearest class. In addition, $\sigma_i$ is constrained to lie between a minimum threshold $\sigma^{\min}$ and a maximum threshold $\sigma^{\max}$. The number (or granularity) of classes can be controlled by adjusting these threshold values.

Besides the modification given by Eq. (7), we performed a purification process to further reduce the number of clusters (i.e., representative data points; **Figure 1a**). In the convex clustering method, representative candidates are typically removed based on a threshold applied to $\pi_i$. However, some redundant points may remain in the cluster representation even after the self-consistent procedure. To eliminate such redundant points, we introduce a purification process based on the condition $j \neq \underset{i}{\operatorname{argmax}}\left\{P_i\left(\mathbf{x}_j\right)\right\}$. When this condition is satisfied, even data point $\mathbf{x}_j$ does not yield the maximum probability for its associated class $\Pi_j$. Such classes contribute only marginally to the clustering and can be removed without significantly affecting the overall model behavior. We apply this purification condition to simplify the clustering model by removing redundant representative points.

View Article Online
DOI: 10.1039/D5CP03946K

### 2-3. Regression based on convex clustering

The (modified) convex clustering method discussed in the previous section can be easily expanded to perform regression. In regression tasks, we typically consider two types of data, explanatory variables (feature vectors) $X$ and target values $Y$, as follows:

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \tag{11-1}$$

$$Y = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}. \tag{11-2}$$

As the first step in constructing a regression model based on the (modified) convex clustering method, a clustering model $M_{clustering}(\mathbf{x})$ is built using only explanatory (feature) data $X$. For a given point $\mathbf{x}$, a set of probabilities over representative classes can be obtained using Eq. (9) as follows:

$$M_{clustering}(\mathbf{x}) \to \{P_1(\mathbf{x}), P_2(\mathbf{x}), \cdots, P_N(\mathbf{x})\}. \tag{12}$$

Here, the relation $\sum_{i=1}^{N} P_i(\mathbf{x}) = 1$ holds. Based on the probabilities, we can predict a target value $\mathbf{y}_{predict}$ for data point $\mathbf{x}$ as follows:

$$\mathbf{y}_{predict} = \sum_{i}^{N} \mathbf{y}_i P_i(\mathbf{x}). \tag{13}$$

Regression based on convex clustering is simple, making it easy to analyze the behavior of the prediction process. We discuss such an analysis in relation to radical reactions in Section 3.

Here, we discuss the similarities and differences between the (modified) convex clustering method and the k-nearest neighbor (*k*-NN) algorithm. The *k*-NN algorithm predicts an unknown data point by selecting the *k* nearest neighbors from the

View Article Online
DOI: 10.1039/D5CP03946K

training dataset and determining the outcome based on this local information. For classification tasks, the class is assigned by majority voting among the neighbors, whereas for regression tasks, the prediction typically involves averaging the neighbors' values or applying distance-based weighting. The k-NN method usually requires storing all training data and computing distances to every data point during prediction, which increases computational cost and model complexity as the dataset grows. In contrast, the convex clustering method represents the dataset using representative points selected from the training data, modeled as a mixture of multiple Gaussian distributions. Only these representative points, along with their mixing ratios, are retained in the model and used during prediction, thereby reducing complexity. Although the convex regression approach differs theoretically from the non-parametric *k*-NN method, both share the characteristic of leveraging training data for prediction, contributing to an intuitive understanding of the prediction process.

### 2-4. Computational conditions

In this study, we analyzed chemical data related to polymer radical reactions using ML. The radical reaction data were generated through DFT calculations. The structures of the reactants, products, and transition states were optimized using the B3LYP functional with Grimme's empirical dispersion correction and the 6-31+G* basis set (B3LYP-D3/6-31+G*).[38-40] All DFT calculations were performed using the Gaussian16 software package.[41] For convenience, a summary of the dataset is provided in Section S2 of the Supporting Information, and more detailed descriptions can be found in the literature.[6] We implemented the modified convex clustering (regression) method in

View Article Online
DOI: 10.1039/D5CP03946K

Python.[42] The machine learning analysis based on the random forest (n_estimators = 30), kernel ridge (alpha = 1.0), and k-NN algorithms[36] was carried out using the scikit-learn library (version 1.5.1).[43]

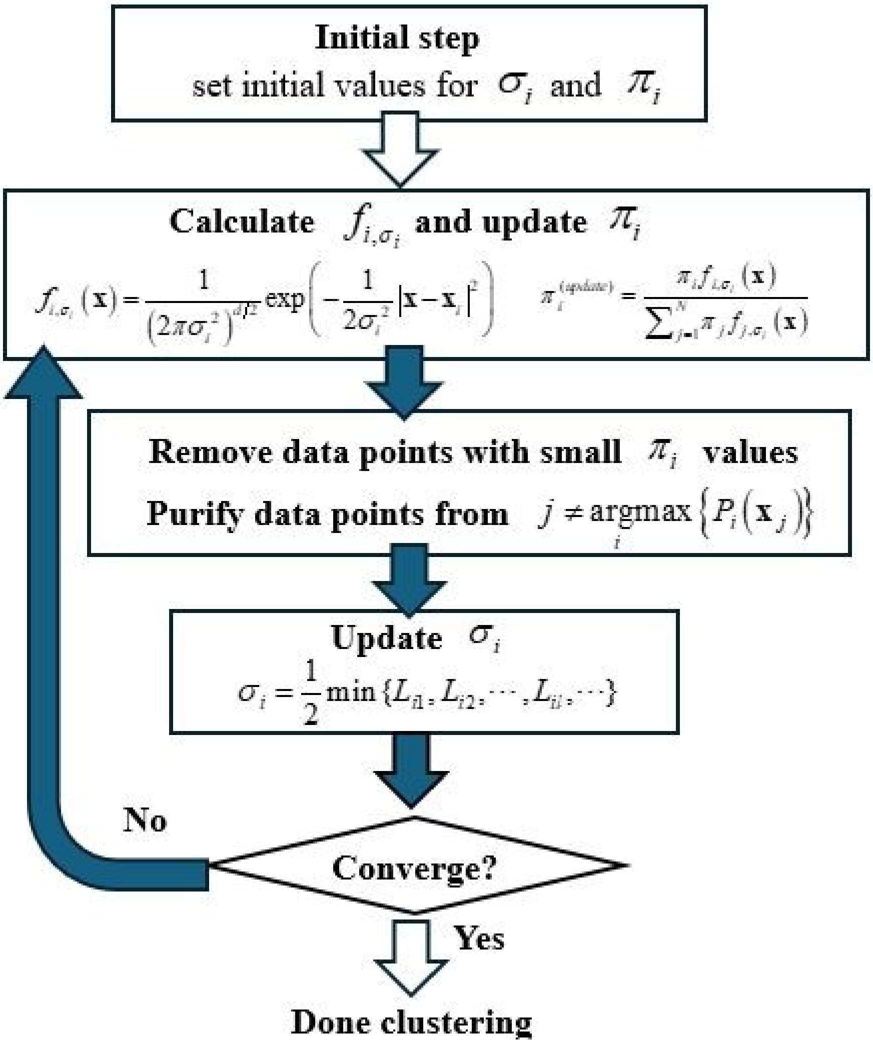

Initial step
set initial values for $\sigma_i$ and $\pi_i$
Calculate $f_{i,\sigma_i}$ and update $\pi_i$
$f_{i,\sigma_i}(\mathbf{x}) = \frac{1}{(2\pi\sigma_i^2)^{d/2}} \exp\left(-\frac{1}{2\sigma_i^2}|\mathbf{x}-\mathbf{x}_i|^2\right)$
$\pi_i^{(update)} = \frac{\pi_i f_{i,\sigma_i}(\mathbf{x})}{\sum_{j=1}^{N} \pi_j f_{j,\sigma_i}(\mathbf{x})}$
Remove data points with small $\pi_i$ values
Purify data points from $j \neq \underset{i}{\operatorname{argmax}}\{P_i(\mathbf{x}_j)\}$
Update $\sigma_i$
$\sigma_i = \frac{1}{2}\min\{L_{i1}, L_{i2}, \cdots, L_{ii}, \cdots\}$
No
Converge?
Yes
Done clustering

**Figure 1.** Computational flow of the modified convex clustering method.

View Article Online
DOI: 10.1039/D5CP03946K

## 3. Results and discussions

### 3.1 Basic behavior of the modified convex clustering method

We analyze the behavior of the modified convex clustering method using a simple two-dimensional dataset with 1,000 points (**Figure 2a**). The points were randomly generated from five normal distributions with means at (1, 6), (5, 2), (6, 9), (8, 5), and (10, 8) and the standard deviations of 0.9, 0.8, 1.0, 1.1, and 1.2, respectively. We applied the modified convex clustering method to this dataset. We employed a maximum threshold value of $\sigma^{\max} = 3$. We show clustering results with $\delta^{\min}$ = 1.0, 0.8, and 0.5 in **Figures 2b**, **2c**, and **2d**, respectively. In these figures, star markers indicate representative points, and orange-dotted circles indicate the class distributions ($\sigma_i$). For $\delta^{\min} = 1.0$, five representative points are obtained from the convex clustering method (**Figure 2b**). The means of these representatives are at (0.91, 5.86), (4.94, 2.10), (6.08, 9.03), (8.23, 4.97), and (10.417, 8.30). These representative points are included in the training dataset in **Figure 2a**. When smaller $\delta^{\min}$ values are used, data points are partitioned into more narrowly defined classes. For example, $\delta^{\min}$ = 0.8 and 0.5 result in 7 and 12 classes, respectively. In the convex clustering method, the number of clusters is automatically determined based on $\delta^{\min}$. This hyperparameter sets the lower bound for the class distribution; thus, smaller values lead to finer-grained clustering. In

View Article Online
DOI: 10.1039/D5CP03946K

other words, a smaller $\delta^{\min}$ value allows the dataset to be covered by more compact clusters. Conversely, a larger $\delta^{\min}$ value results in a coarser representation of the data. Thus, we can control the density (granularity) of the clusters by adjusting $\delta^{\min}$.

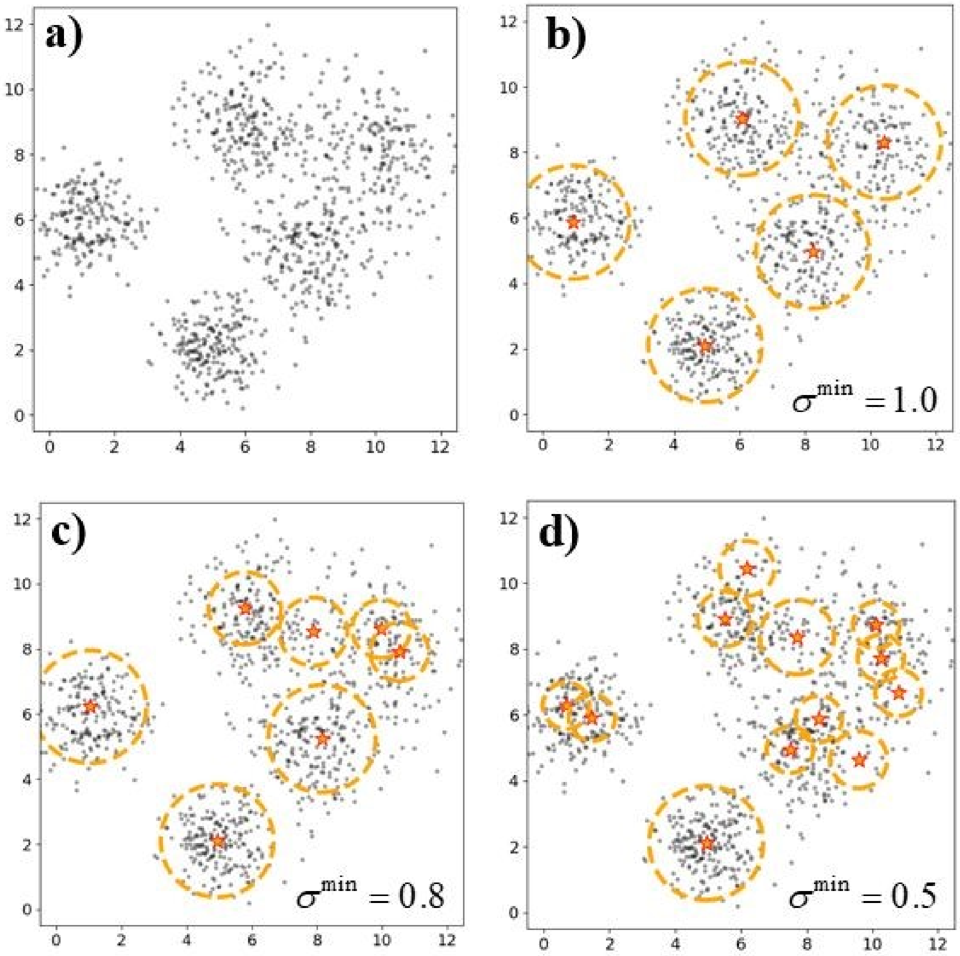


**Figure 2.** Clustering results for a simple dataset based on the modified convex clustering algorithm.

View Article Online
DOI: 10.1039/D5CP03946K

### 3.2 Chemical reaction data analysis based on the convex regression method

In this section, we analyze ACR and MA radical reactions (**Figure 3a**) using the modified convex regression method. Previously, we presented radical reactions involving various combinations of ACRs and MAs (**Figure 3b**) and calculated their reaction barriers using the DFT method.[6] The chemical dataset was analyzed using a modified convex clustering approach. The reaction energy barriers ($\Delta E_{\mathrm{TS}}$) for radical reactions were predicted using the convex regression method based on Eq. (13), with product–reactant energy difference ($\Delta E_{RP}$) and a dummy parameter (DP(X)) as explanatory variables (**Figure 3c**). Here, the dummy parameter DP(m) represents either ACR or MA for the radical reaction X· + Y → XY·, where m specifies X or Y: DP(X) = 0 (1) indicates X = ACR (MA). The values are also summarized in SI. We standardized the input features to train ML models. **Table 1** presents the regression results obtained by varying the hyperparameter $\sigma^{\min}$. To evaluate the performance of the regression models, we used the 5-fold cross-validation to calculate the mean absolute error (MAE).[36] The table also includes the coefficient of determination ($R^2$). Notably, smaller MAE values can be obtained when the training dataset is used directly as input. **Figure 4** shows a comparison of the DFT-calculated reaction energy barriers with prediction results from the regression model for $\sigma^{\min} = 0.0075$. The results confirm that the regression model tends to perform better when smaller values of $\sigma^{\min}$ are used. As discussed in the previous section, $\sigma^{\min}$ controls the density of clusters (or

View Article Online
DOI: 10.1039/D5CP03946K

classes) covering the data, with the number of clusters increasing as the value decreases, and hence the regression model can more accurately predict reaction barriers. For example, MAE values of 0.39 and 0.52 kcal/mol were obtained for $\sigma^{\min} = 0.015$ and 0.03, respectively. Conversely, there is no established method for determining the optimal value of sigma in convex clustering. Therefore, similar to standard hyperparameter tuning, the parameter needs to be adjusted, and the model's behavior evaluated to identify a suitable value. According to the results in **Table 1**, the regression performance tends to level off as the sigma value decreases, indicating that sigma should be chosen with consideration for model complexity to avoid overfitting. Developing a more systematic approach to selecting sigma in convex clustering may be an important topic for future research.

Here, we compare the results obtained from convex clustering regression with kernel ridge, random forest, and k-NN algorithms. For kernel ridge regression, the MAE and $R^2$ were 0.44 kcal/mol and 0.83, respectively, when using a linear kernel. With a radial basis function kernel, the MAE and $R^2$ improved to 0.36 kcal/mol and 0.89. The random forest method achieved an MAE of 0.36 kcal/mol and an $R^2$ of 0.88. For the k-NN method with 3, 5, and 7 neighbors, the MAEs were 0.31, 0.30, and 0.33 kcal/mol, and the corresponding $R^2$ values were 0.92, 0.91, and 0.89. In comparison, the convex regression method with sigma = 0.0075 yielded an MAE of 0.30 kcal/mol and an $R^2$ of 0.93. These results indicate that the convex regression method provides predictive performance comparable to, and in some cases slightly better than, other machine learning approaches for the dataset analyzed in this study.

In the modified convex regression method, representative points are selected from the training dataset and stored as internal variables (or states) within the model.

View Article Online
DOI: 10.1039/D5CP03946K

**Figure 5** shows several representative points with relatively large $\pi_i$ values, stored in the prediction model for $\sigma^{\min} = 0.0075$, using a chemical reaction representation. ML models often behave like black boxes, making it difficult to interpret how predictions are made. However, in the modified convex regression method, predictions are based on representative points that originate from the training dataset. This characteristic allows for straightforward analysis of model behavior. For instance, to understand the prediction of reaction energy barriers, we can examine the contribution of each representative point. **Figure 6** shows the contributions for several predictions. For example, the DFT-based TS calculation yielded a barrier of 6.55 kcal/mol for the radical reaction between methyl MA (compound **4** in **Figure 3b**) and γ-butyrolactone MA (compound **8** in **Figure 3b**), whereas the convex regression model predicted a barrier of 6.39 kcal/mol. This prediction was primarily influenced by two reactions: the radical reaction between methyl MA (compound **4**) and methacrylic acid (compound **2**) yielded the contribution of 76.4%, and the reaction between *t*-butyl MA (compound **6**) and MA (compound **4**) yielded the contribution 23.4%, **(Figure 6a**). Similarly, for the reaction between ethyl-cyclohexyl ACR (compound **9**) and ethyl-cyclohexyl MA (compound **10**), the DFT calculation yielded a barrier of 4.40 kcal/mol, whereas the ML model predicted 4.84 kcal/mol. In this case, the reaction between ethyl-cyclohexyl ACR (compound **9**) and methyl MA (compound **4**) contributed 93.8%, and the reaction between γ-butyrolactone ACR (compound **7**) and methacrylic acid (compound **2**) contributed 3.89% (**Figure 6b)**. Thus, the convex regression method enables intuitive interpretation of ML prediction by analyzing the contributions of representative points.

In these reaction predictions, the machine learning model clearly focuses on similarities among reactant monomers. For example, methyl MA (compound **4**) in

View Article Online
DOI: 10.1039/D5CP03946K

**Figure 6a** and ethyl-cyclohexyl ACR (compound **9**) in **Figure 6b** appear both in the target reactions and in the reactions that contributed most to the prediction. The model also assigned importance to whether the monomer belongs to MA or ACR when making predictions, as shown in **Figure 6**. This distinction is critical in radical reactions because it determines the stability of the reactant and product radicals. For reactions involving relatively large side chains, the model tends to reference reactions with similarly large side chains, as these can influence radical behavior through steric effects. The machine learning model appears to account for these effects as well, which aligns with chemical intuition. Conversely, it is worth noting that machine learning using the convex clustering method does not predict reactions based on chemical understanding or causal relationships as researchers do, but rather relies solely on the similarity of reaction data. Nevertheless, analyzing the model's behavior in this way may help researchers extract chemical insights from the data. Improving prediction transparency could facilitate uncovering chemical insights from machine learning analyses of chemical data.

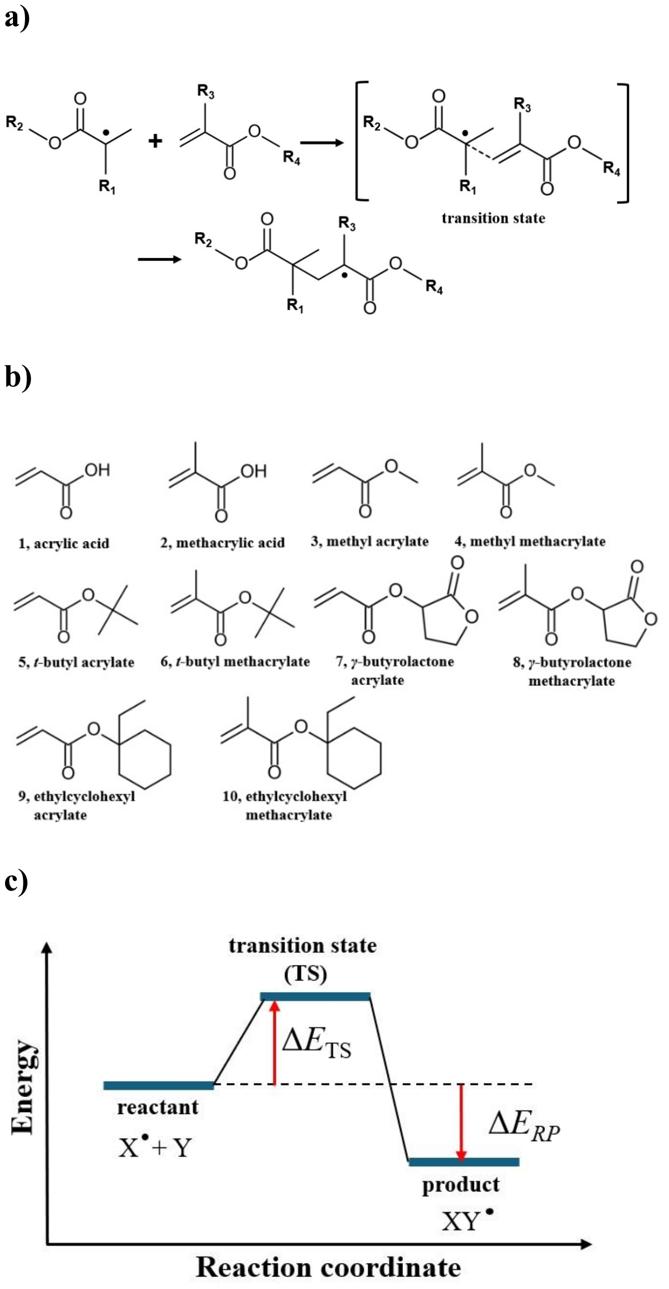
The convex clustering method selects representative points along with their class proportions, making the model simpler and more efficient. A simple prediction process also helps improve understanding of the model's behavior. As shown in Eqs. (12) and (13), even when the dataset is small and an unseen data point is far from the representative points, the regression process remains influenced by the nearest representatives. This helps prevent extreme predictions and ensures stable behavior. Conversely, as with other machine learning methods, improving predictive performance requires expanding the training dataset. In addition, similar to distance-based models such as GMM and k-NN, feature selection is also critical for convex clustering.

View Article Online
DOI: 10.1039/D5CP03946K

Dimensionality reduction and careful feature selection are essential for building effective models with this approach. To address this challenge, we are currently investigating a method that combines feature refinement with convex clustering. The results of this research will be reported elsewhere.

The complexity of ML algorithms leads to difficulties in understanding their predictive process by humans. To alleviate the incomprehensibility of ML, we focused on the selection process of representative points. In the convex clustering approach, representative points are selected from the training dataset and used to make predictions. By analyzing the contributions of these representative points, we can gain some insight into a model's behavior. Selecting representative points directly from the training dataset plays a valuable role in enhancing the transparency and interpretability of ML models. We considered the convex regression method as an example, which may serve as a guideline for developing ML algorithms with improved analyzability. In particular, chemical datasets often contain rich information, and incorporating this information directly into ML models can further enhance their interpretability and transparency.

View Article Online
DOI: 10.1039/D5CP03946K



View Article Online
DOI: 10.1039/D5CP03946K

**Figure 3.** a) Scheme of the radical reaction between ACR and MA. b) Reactant monomers of acrylic acid, ACR, methacrylic acid, and MA. We categorize acrylic and methacrylic acids as ACR and MA, respectively. c) Energy diagram for radical reaction $X^{\cdot} + Y \rightarrow XY^{\cdot}$, where $X^{\cdot}$ represents the radical monomer.

**Table 1. Predictive performance of the convex regression method for energy reaction barrier based on cross-validation.**

| $\sigma^{\min}$ | **MAE [kcal/mol]** | **R$^2$** |
|---|---|---|
| 0.04 | 0.58 | 0.67 |
| 0.03 | 0.52 | 0.74 |
| 0.02 | 0.49 | 0.80 |
| 0.015 | 0.39 | 0.86 |
| 0.0075 | 0.30 | 0.93 |

View Article Online
DOI: 10.1039/D5CP03946K

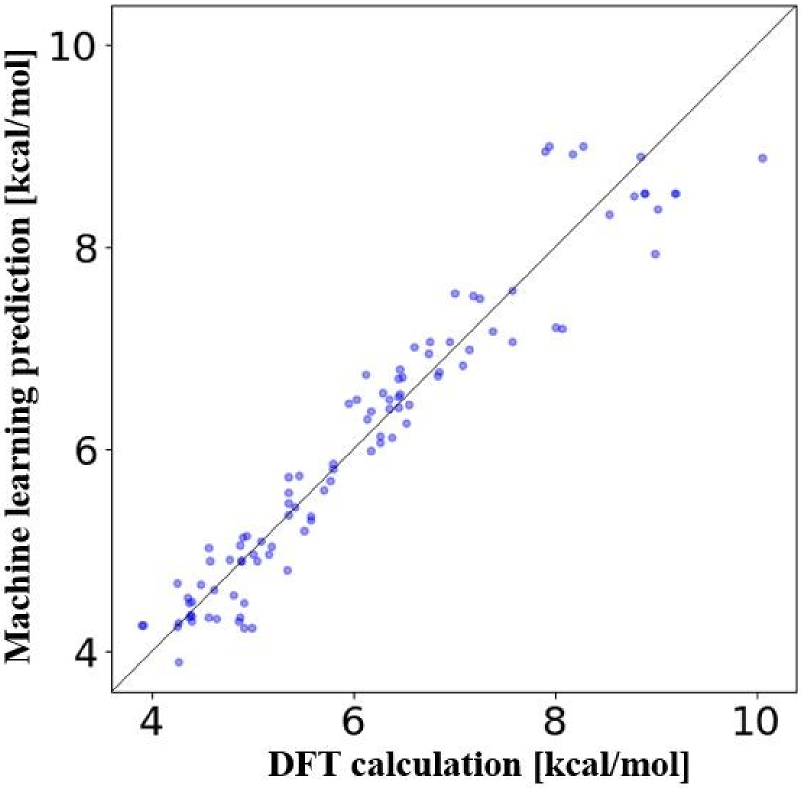
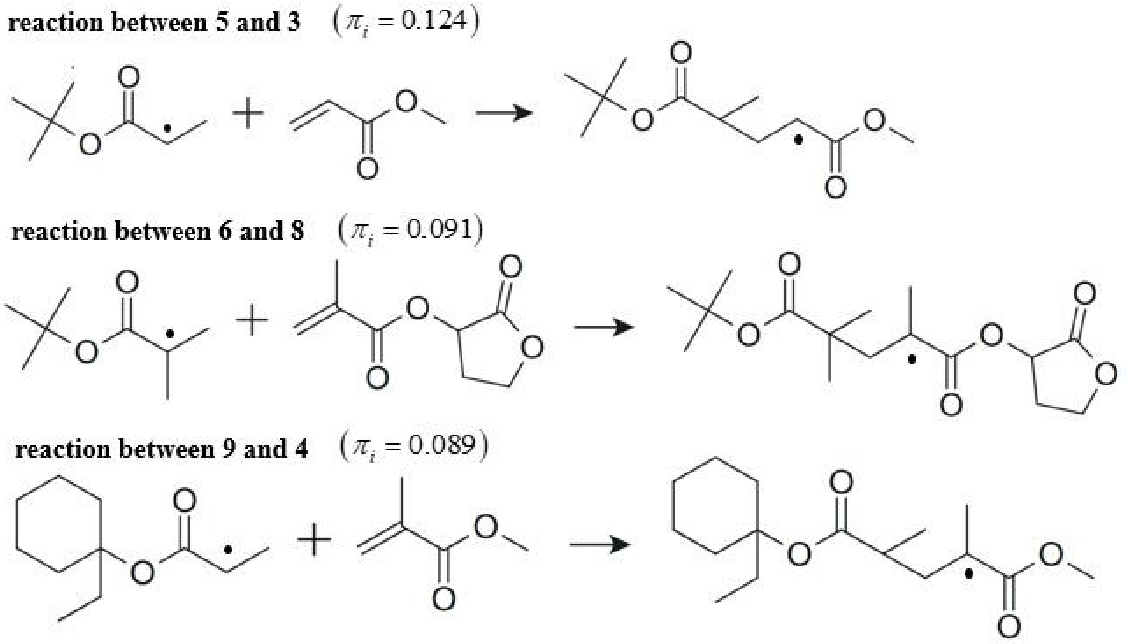


**Figure 4.** Comparison between DFT calculations and ML predictions for reaction energy barriers [kcal/mol].

View Article Online
DOI: 10.1039/D5CP03946K



reaction between 5 and 3 $(\pi_i = 0.124)$

reaction between 6 and 8 $(\pi_i = 0.091)$

reaction between 9 and 4 $(\pi_i = 0.089)$

**Figure 5.** Representative data stored in the convex regression model. We show some representative points with relatively large $\pi_i$ values using a chemical reaction representation.

View Article Online
DOI: 10.1039/D5CP03946K

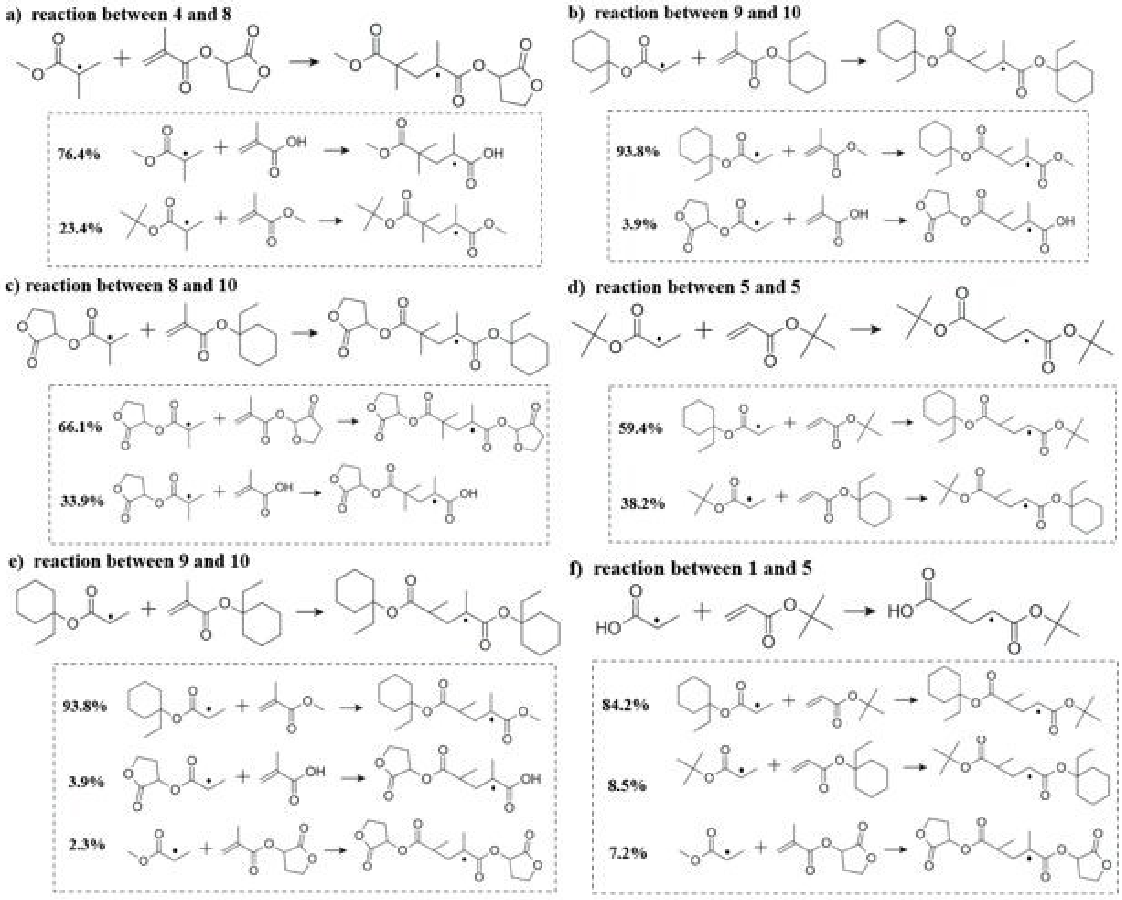

a) reaction between 4 and 8
76.4%
23.4%
b) reaction between 9 and 10
93.8%
3.9%
c) reaction between 8 and 10
66.1%
33.9%
d) reaction between 5 and 5
59.4%
38.2%
e) reaction between 9 and 10
93.8%
3.9%
2.3%
f) reaction between 1 and 5
84.2%
8.5%
7.2%

**Figure 6.** Contributions from representative points stored in ML models to predict energy barriers.

View Article Online
DOI: 10.1039/D5CP03946K

## 4. Summary

In this study, we explored a modified convex clustering (regression) method, in which representative points used to describe classes are selected directly from the training dataset. In this approach, each class is associated with a parameter that defines its size (distribution), allowing for a flexible data representation. We demonstrated that data are described more coarsely when the number of representative points is smaller. Conversely, increasing the number of classes (clusters) enables a more fine-grained representation, which can enhance the predictive performance of ML models. However, this increased granularity induces greater model complexity. The number of representative points provides a means to control both model granularity and complexity. We applied the modified method to ACR and MA radical reaction data and constructed ML models to predict reaction energy barriers. Our results showed that prediction accuracy improves with the number of representative points. We also analyzed the prediction process by examining the contributions of individual representative points, where the energy barrier is estimated as a weighted sum of contributions from radical reactions. The model's behavior can be easily interpreted because the representative points are selected from the training dataset and directly used in predictions. We concluded that selecting representative points from the training dataset is a useful strategy for improving the interpretability and transparency of ML models. The simplicity and analyzability of the modified convex clustering (regression) method make it a promising tool for deeper investigation of chemical and scientific data.

View Article Online
DOI: 10.1039/D5CP03946K

**Supporting Information**.

We described S1) the K-near clustering method and S2) the radical reaction dataset in another PDF file.

**Notes**

There are no conflicts of interest to declare.

**Acknowledgment**

This work is partly supported by Grand-in-Aid for Scientific Research (KAKENHI) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grand Number 21H00026 and 22K05038. This work was partly supported by MEXT as “Program for Promoting Researches on the Supercomputer Fugaku” (JPMXP1020230318). This study used computational resources from Supercomputer Center, the Institute for Solid State Physics, the University of Tokyo and Research Center for Computational Science, Okazaki, Japan.

View Article Online
DOI: 10.1039/D5CP03946K

## References

1. M. Nakata and T. Shimazaki, *J Chem Inf Model*, 2017, **57**, 1300-1308.
2. M. Nakata, T. Shimazaki, M. Hashimoto and T. Maeda, *J Chem Inf Model*, 2020, **60**, 5891-5899.
3. T. Shimazaki and M. Tachikawa, *Acs Omega*, 2022, **7**, 10372-10381.
4. T. Shimazaki and M. Tachikawa, *Chem Phys Lett*, 2023, **829**, 140744.
5. T. Shimazaki and M. Tachikawa, *Chem Phys Lett*, 2025, **861**, 141830.
6. M. Takagi, T. Shimazaki, O. Kobayashi, T. Ishimoto and M. Tachikawa, *Phys Chem Chem Phys*, 2025, **27**, 1772-1777.
7. W. A. Warr, *Mol Inform*, 2014, **33**, 469-476.
8. J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *Npj Computational Materials*, 2019, **5**, 83.
9. J. A. Keith, V. Vassilev-Galindo, B. Q. Cheng, S. Chmiela, M. Gastegger, K. R. Mueller and A. Tkatchenko, *Chem Rev*, 2021, **121**, 9816-9872.
10. K. Jorner, A. Tomberg, C. Bauer, C. Sköld and P. O. Norrby, *Nat Rev Chem*, 2021, **5**, 240-255.
11. A. Iskandarov, T. Tada, S. Iimura and H. Hosono, *Acta Materialia*, 2022, **230**, 117825.
12. R. X. Wang, X. L. Fang, Y. P. Lu and S. M. Wang, *J Med Chem*, 2004, **47**, 2977-2980.
13. J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J Chem Inf Model*, 2012, **52**, 1757-1768.
14. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *Apl Mater*, 2013, **1**, 011002.
15. S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Y. Han, J. E. He, S. Q. He, B. A. Shoemaker, J. Y. Wang, B. Yu, J. Zhang and S. H. Bryant, *Nucleic Acids Res*, 2016, **44**, D1202-D1213.
16. R. Jose and S. Ramakrishna, *Appl Mater Today*, 2018, **10**, 127-132.
17. C. Draxl and M. Scheffler, *Mrs Bull*, 2018, **43**, 676-682.
18. J. S. Smith, O. Isayev and A. E. Roitberg, *Scientific Data*, 2017, **4**, 170193.
19. C. Isert, K. Atz, J. Jimenez-Luna and G. Schneider, *Scientific Data*, 2022, **9**, 273.
20. L. C. Yang, X. Li, S. Q. Zhang and X. Hong, *Org Chem Front*, 2021, **8**, 6187-6195.
21. J. E. Alfonso-Ramos, R. M. Neeser and T. Stuyver, *Digit Discov*, 2024, **3**, 919-931.
22. A. Adadi and M. Berrada, *Ieee Access*, 2018, **6**, 52138-52160.
23. D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf and G. Z. Yang, *Sci Robot*, 2019, **4**, 7120.

View Article Online
DOI: 10.1039/D5CP03946K

24. P. Linardatos, V. Papastefanopoulos and S. Kotsiantis, *Entropy-Switz*, 2021, **23**, 18.

25. S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K. W. Low, S. F. Newman, J. Kim and S. I. Lee, *Nat Biomed Eng*, 2018, **2**, 749-760.

26. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S. I. Lee, *Nat Mach Intell*, 2020, **2**, 56-67.

27. L. Breiman, *Mach Learn*, 2001, **45**, 5-32.

28. D. Lashkari and P. Golland, *Advances in Neural Information Processing Systems*, 2007, **20**, 3181.

29. U. Ali, K. J. B. Abd Karim and N. A. Buang, *Polym Rev*, 2015, **55**, 678-705.

30. S. C. Ligon, K. Seidler, C. Gorsche, M. Griesser, N. Moszner and R. Liska, *J Polym Sci Pol Chem*, 2016, **54**, 394-406.

31. N. Ballard and J. M. Asua, *Prog Polym Sci*, 2018, **79**, 40-60.

32. A. Debuigne, C. Michaux, C. Jérôme, R. Jérôme, R. Poli and C. Detrembleur, *Chem-Eur J*, 2008, **14**, 7623-7637.

33. I. Degirmenci, V. Aviyente, V. Van Speybroeck and M. Waroquier, *Macromolecules*, 2009, **42**, 3033-3041.

34. X. Li, S. Q. Zhang, L. C. Xu and X. Hong, *Angew Chem Int Edit*, 2020, **59**, 13253-13259.

35. K. A. Spiekermann, X. R. Dong, A. Menon, W. H. Green, M. Pfeifle, F. Sandfort, O. Welz and M. Bergeler, *J Phys Chem A*, 2024, **128**, 8384-8403.

36. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learnin: Data Mining, Inference, and Prediction, Second Edition*, Springer, New York, 2009.

37. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

38. A. D. Becke, *Journal of Chemical Physics*, 1993, **98**, 5648.

39. C. T. Lee, W. T. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785-789.

40. S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *Journal of Chemical Physics*, 2010, **132**, 154104.

41. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, J. E. P. Jr., F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K.

View Article Online
DOI: 10.1039/D5CP03946K

Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16, Revision A.03*, Gaussian, Inc., Wallingford CT, , 2016.

42. G. V. Rossum and F. L. Drake, *Python 3 Reference Manual*, CreateSpace, Scotts Valley, 2009.

43. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res,*, 2011, **12**, 2825-2830.

View Article Online
DOI: 10.1039/D5CP03946K

The data supporting this article have been included as part of the Supplementary Information.