Development of Prediction Models for the Self-Accelerating Decomposition Temperature of Organic Peroxides

Toshiharu Morishita and Hiromasa Kaneko*

ABSTRACT: Thermal risk assessment is very important in the primary stages of chemical compound development. In this study, a model to estimate the self-accelerated decomposition temperature of organic peroxides was developed. The structural information of compounds was used to calculate descriptors, on which partial least-squares (PLS) regression and support vector regression were applied for temperature prediction. Molecular mechanics and density functional theory calculations were performed before descriptor calculations, for structure optimization, using a genetic algorithm for variable selection. Structure optimization and variable selection immensely improved the prediction accuracy. Thus, a PLS model, with $R^2 = 0.95$, root mean square error = 5.1 °C, and mean absolute error = 4.0 °C, exhibiting higher accuracy than existing self-accelerating decomposition temperature prediction models, was constructed.

INTRODUCTION

Thermal risk assessment is extremely crucial in the development of chemical compounds. Self-accelerating decomposition temperature (SADT) is a key parameter characterizing the thermal risk of organic peroxides. It is the lowest temperature for self-accelerating decomposition in organic peroxides and self-reactive substances (used in transportation packaging). Thus, it determines the optimum temperature-control to avoid thermal hazards during material storage and transport. Several experimental methods measure SADT for thermal risk assessment; however, the associated cost, risk, and chemicals make early-stage evaluation very difficult. Therefore, it is beneficial to develop a simple and high-accuracy SADT prediction method.

Previous studies have proposed quantitative structure–property relationship models to predict SADT. Wang et al. performed density functional theory (DFT) calculations [6-31G(d)/B3LYP] in Gaussian 09 to obtain descriptors. Geometrical descriptors (bond length, bond angle, and dipole moment) and quantum chemical descriptors [highest occupied molecular orbital (HOMO)/lowest unoccupied MO (LUMO), bond dissociation energy] were used to construct prediction models with multiple linear regression (MLR) and support vector regression (SVR) to estimate SADT. He et al. used the semiempirical molecular orbital technique (AM1) for preprocessing before geometry optimization and frequency calculations. Descriptors (excluding quantum chemical descriptors) were calculated in DRAGON 6.0, and the genetic algorithm (GA) was applied for variable selection, followed by MLR and SVR construction, to estimate SADT. The first and second methods suffered from the limitations of high computational load and low accuracy, respectively.

In this study, a model with high accuracy and low calculation load was developed to estimate the SADT of organic peroxides. Descriptors were calculated using the optimal molecular conformation, determined by molecular mechanics (MM) and DFT calculations. GA was used for variable selection, followed by the application of partial least-squares (PLS) regression, as a linear regression method, and SVR, as a nonlinear regression method, to predict SADT. Prediction models, including and excluding structural optimization and variable selection, were developed and analyzed.

METHODS

Data Set. The data set included 65 organic compounds, with 90.14–571.00 molecular weights, and −5.0 to 196.5 °C SADTs, obtained from the literature, determined using different calorimetric methods (TG-DSC and C80). However, as previously reported, SADT is independent of the determination method used. Compounds included commonly used organic peroxides, such as dialkyl peroxide, diacyl peroxide, hydroperoxide, peroxyester, ketone peroxide, peroxy carbonate, and...
Table 1. Compounds and Their Experimental SADTs

| no. | compound name                          | CAS no. | MW    | SADT [°C] |
|-----|----------------------------------------|---------|-------|-----------|
| 1   | tert-butyl hydroperoxide                | 75-91-2 | 90.14 | 120.4     |
| 2   | cumyl hydroperoxide                     | 80-15-9 | 152.21| 79.0      |
| 3   | dicumyl peroxide                        | 80-43-3 | 270.40| 77.8      |
| 4   | p-menthane hydroperoxide                | 80-47-7 | 172.30| 73.5      |
| 5   | dibenzyl peroxide                       | 94-36-0 | 242.24| 80.0      |
| 6   | diisopropyl peroxycarbonate             | 105-64-6| 206.22| 5.0       |
| 7   | tert-butyl peroxyacetate                | 107-71-1| 132.18| 65.0      |
| 8   | tert-butyl peroxyisobutyrate            | 109-13-7| 160.24| 30.0      |
| 9   | di-tert-butyl peroxide                  | 110-05-4| 118.30| 80.9      |
| 10  | diacetyl peroxide                       | 110-22-5| 118.10| 35.0      |
| 11  | disuccinic acid peroxide                | 123-23-9| 234.18| 25.0      |
| 12  | bis(2,4-dichlorobenzoyl)peroxide        | 133-14-2| 380.00| 60.0      |
| 13  | tert-butyl peroxybenzoate               | 133-14-2| 380.00| 60.0      |
| 14  | tert-butyl peroxyacetate                | 137-71-1| 132.18| 65.0      |
| 15  | tert-butyl peroxyisobutyrate            | 137-71-1| 132.18| 65.0      |
| 16  | tert-butyl peroxy-2-ethylhexanoate      | 130097-36-8| 278.38| 10.0      |
| 17  | tert-butyl cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 18  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 19  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 20  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 21  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 22  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 23  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 24  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 25  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 26  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 27  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 28  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 29  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 30  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 31  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 32  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 33  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 34  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 35  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 36  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 37  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 38  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 39  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 40  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 41  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 42  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 43  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 44  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 45  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 46  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 47  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 48  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 49  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 50  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 51  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 52  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 53  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 54  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 55  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 56  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 57  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 58  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 59  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 60  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
| 61  | tert-butyl-cumyl peroxide               | 3457-61-2| 208.33| 77.1      |
Table 1. continued

| no. | compound name                                      | CAS no.     | MW  | SADT [°C] |
|-----|---------------------------------------------------|-------------|-----|-----------|
| 62  | 2,5-bis-(2-ethylhexanoylperoxy)-2,5-dimethylhexane | 13052-09-0  | 430.70 | 38.6      |
| 63  | tert-amyl peroxy-2-ethylhexanoate                 | 686-31-7    | 230.39 | 35.0      |
| 64  | 2,2-bis(tert-butylperoxy)butane                  | 2167-23-9   | 234.38 | 70.0      |
| 65  | di-tert-amyl peroxide                             | 10508-09-5  | 174.32 | 70.4      |

diperoxide. Organic peroxides used here, and their experimental SADTs, are listed in Table 1. The data set was divided into two subsets: a training set (52 samples) and a test set (13 samples), following the method reported in a previous publication. 6

**Geometry Optimization.** Molecular structures of the 65 compounds were prepared in the molfile format. Some descriptors were molecular-structure-dependent, making geometry optimization very important. Auto geometry optimization on Avogadro 3, a molecular modeling software for quantum chemical calculations, was performed for each mol file. The basic MM potential energy function includes bonded terms (for covalently bonded atomic interactions) and nonbonded terms (for long-range electrostatic and van der Waals forces). Here, UFF (universal force field) was used to improve bond lengths and bond angles to obtain a minimum-energy conformation. Prior MM calculations improved DFT convergence, with less computational load. DFT calculations using Firefly (PC GAMESS),7 via MoCalc2012,8 optimized the molecular structures. “Geometry optimization” job type, using hybrid density functional B3LYP with 6-31G basis set, was used. B3LYP hybrid functional incorporates approximations to the exchange–correlation energy functional in DFT (combination of exact exchange from Hartree–Fock theory, and from other sources). Polarization-incorporated 6-31G, 6-31G(d), is commonly used for organic compounds. Here, 6-31G was chosen to reduce the computational load and verify the effect of d-orbitals.

**Descriptor Calculation and Selection.** After structural optimization, 5666 and 552 molecular descriptors of organic peroxides were calculated by alvaDesc 2.0.8 and CODESSA 3,9 respectively, most of which were irrelevant to this study. These methods calculate molecular descriptors and fingerprints from structural information. descriptors with small standard deviations and strong multicollinearity were eliminated. CODESSA was used to calculate the quantum chemical descriptors, such as HOMO/LUMO, and GA was used to find a descriptor set for model construction.

GA, applied as a variable selection method, is an iterative procedure that continuously improves the fitness function, from which the fitness score (indicating probability of descriptor set selection) is calculated. An initial population of descriptors (a few hundred sets) was selected at random or heuristically. Each iteration step calculated and assigned a fitness value to the descriptor sets, and proportional probabilities were used to select a new descriptor population. This selection procedure cannot independently generate a new point in the search space; thus, crossover and mutation were additionally used by GA to generate new descriptor sets. Here, the coefficient of determination (R²), after fivefold cross validation in PLS and SVR modeling, labeled GA-PLS and GA-SVR, respectively, was used as the fitness function. Number of components in the population was 100, crossover probability was 0.5, mutation probability was 0.2, and maximum number of generations was 200.

**Regression and Validation.** Here, PLS and SVR were used to develop the prediction models. PLS is a statistical linear regression method used to find fundamental relations between explanatory variables (X) and response variables (Y). It is widely used when the number of explanatory variables is significantly larger than the number of samples. Decompositions of X and Y, as shown below, were constructed to maximize the covariance between the latent variables (T) and response variables (Y).

\[ X = \sum_{a=1}^{A} t_p^a T + E = TP^T + E \]  
\[ Y = \sum_{a=1}^{A} t_q^a + f = Tq^T + f \]

where A is the number of latent variables, \( t_p^a \) is the a\(^{th} \) latent variable, \( p_a \) is the a\(^{th} \) loading, \( q_a \) is the weight on a\(^{th} \) latent variable, and \( E \) and \( f \) are the error terms that cannot be explained by \( X \) and \( Y \). \( p_a \) and \( q_a \) were calculated to minimize the sum of squares of errors. The number of latent variables with the highest \( R^2 \) obtained via fivefold cross validation, was used in the prediction model.

Variable importance in projection (VIP) 10 scores of each descriptor were calculated to identify variables that contributed significantly to the prediction. VIP scores, as shown below, were defined for each X variable and \( j \) as the sum of latent variables of its PLS-weight value \( (w_j) \), weighted by the percentage of explained Y variance.

\[ VIP_j = \sqrt{\frac{\sum_{i=1}^{h} R^2(y, t_j)(w_j/||w||)^2}{(1/p)\sum_{i=1}^{h} R^2(y, t_j)}} \]

where h is the number of latent variables, \( w \) is the weight vector, and \( R^2(y, t_j) \) is the percentage of explained Y variance.

SVR is also a regression method, performing linear and nonlinear regression using the kernel trick, implicitly mapping inputs into high-dimensional feature spaces. If \( f(x) \) is the explanatory variable for the i\(^{th} \) sample, the response variable, \( f(x(i)) \), is expressed as follows

\[ f(x(i)) = \phi(x(i))w + b \]

\[ w = [w_1 w_2 ... w_k] \]

where b is a constant term, w is the weight vector, and k is the number of dimensions. The error function is expressed as follows

\[ E = C \sum_{i=1}^{n} \max(0, |y(i) - f(x(i))| - \epsilon) + \frac{1}{2}||w|| \]

where C is the regularization parameter, n is the number of training samples, and \( \epsilon \) specifies the epsilon tube within which no penalty is associated with the loss function.

The RBF kernel, a kernel function used in machine learning, shown below, was used.
These parameters were calculated as follows:

\[ K(x^{(i)}, x^{(j)}) = \exp(-\gamma \|x^{(i)} - x^{(j)}\|^2) \]  

where \( \gamma \) is the RBF kernel parameter. SVR includes three hyperparameters (\( C, \varepsilon \), and \( \gamma \)) to be provided before model construction.

Model development and descriptor selection were performed using Python 3.7. Scikit-learn, a machine learning library in Python, was used for the PLS and SVR calculations. GridSearchCV, a parameter estimator in Python, was used for hyperparameter optimization. Fast optimization of hyperparameters was implemented following the procedure adopted by a previous publication.

Hyperparameter optimization and descriptor selection were performed only on the training data set and then evaluated the performance of the model for the test data set. Model accuracy was evaluated on the common statistical parameters: root mean square error (RMSE), mean absolute error (MAE), and \( R^2 \). These parameters were calculated as follows:

\[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,\text{measured}} - y_{i,\text{predicted}})^2}{n}} \]  

\[ \text{MAE} = \frac{\sum_{i=1}^{n} |y_{i,\text{measured}} - y_{i,\text{predicted}}|}{n} \]  

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{i,\text{measured}} - y_{i,\text{predicted}})^2}{\sum_{i=1}^{n} (y_{i,\text{measured}} - y_{i,\text{mean}})^2} \]  

### RESULTS AND DISCUSSION

Prediction models, including and excluding structural optimization and variable selection, were developed and analyzed. Additionally, prediction accuracies of existing models were compared (Table 2).

#### Table 2. Comparison of Predictive Performance of Existing Models for Test Data

|                     | Wang\(^2\) | He\(^3\) |
|---------------------|------------|----------|
| geometry optimization | DFT        | MM+/MO PM1 |
| frequency calculation | B3LYP/6-31G(d) | GA-MLR   |
| variable selection   | 8          | 9        |
| number of descriptors| 40         | 57       |
| modeling method      | MLR        | SVR      |
| number of training data | 40      | 57       |
| number of test data  | 10         | 14       |
| RMSE                 | 12.0       | 6.43     |

**Case 1 (Geometry Optimization: MM).** In this case, only the MM calculation was performed before model development. After preprocessing, the model was developed by PLS, with five latent variables, and SVR (\( C = 8.0, \varepsilon = 0.00098, \) and \( \gamma = 0.00024 \)). A parity plot of actual values versus calculated values is shown in Figure 1. Statistical parameters of the model for the training set were as follows: \( R^2 = 0.99, \) RMSE = 0.98, and MAE = 0.76 for PLS, and \( R^2 = 0.99, \) RMSE = 2.15, and MAE = 0.34 for SVR, and those for the test set were as follows: \( R^2 = 0.82, \) RMSE = 9.80, and MAE = 7.70 for PLS, and \( R^2 = 0.77, \) RMSE = 11.07, and MAE = 8.68 for SVR. Prediction performance was significantly higher than case 1 and comparable to existing models. Prediction accuracy of both the models could be improved by DFT calculations, and the addition of quantum chemical descriptors, before model building. Changing the number of latent variables in PLS, and hyperparameter values in SVR, did not improve the prediction accuracy.

**Case 2 (Geometry Optimization: MM/DFT).** In this case, the MM and DFT calculations were performed before model development. After preprocessing, the model was developed by PLS using 13 latent variables and SVR (\( C = 4.0, \varepsilon = 0.00098, \) and \( \gamma = 0.00049 \)). A parity plot of actual values versus calculated values is shown in Figure 2. The statistical parameters of the model for the training set were as follows: \( R^2 = 0.99, \) RMSE = 0.98, and MAE = 0.76 for PLS, and \( R^2 = 0.99, \) RMSE = 2.15, and MAE = 0.34 for SVR, and those for the test set were as follows: \( R^2 = 0.82, \) RMSE = 9.80, and MAE = 7.70 for PLS, and \( R^2 = 0.77, \) RMSE = 11.07, and MAE = 8.68 for SVR. Prediction performance was significantly higher than case 1 and comparable to existing models. Prediction accuracy of both the models could be improved by DFT calculations, and the addition of quantum chemical descriptors, before model building. Changing the number of latent variables in PLS, and hyperparameter values in SVR, did not improve the prediction accuracy, similar to case 1.

**Case 3 (Geometry Optimization: MM/DFT, Variable Selection: GA).** In this case, in addition to preprocessing (similar to case 2), the descriptors were selected using GA-PLS and GA-SVR, before model building. After variable selection, the fitness function of GA was improved from \( R^2 = 0.69 \) to \( R^2 = 0.91 \) for PLS and from \( R^2 = 0.38 \) to \( R^2 = 0.90 \) for SVR. The model was developed by PLS with 11 latent variables and SVR (\( C = 466, \varepsilon = 0.02343, \) and \( \gamma = 0.00000714 \)). A parity plot of actual values versus calculated values is shown in Figure 3. The statistical parameters of the model for the training set were as follows: \( R^2 = 0.99, \) RMSE = 1.57, and MAE = 1.14 for PLS, and \( R^2 = 0.99, \) RMSE = 3.69, and MAE = 1.78 for SVR, and those for the test set were as follows: \( R^2 = 0.95, \) RMSE = 5.11, and MAE = 4.03 for PLS, and \( R^2 = 0.91, \) RMSE = 6.87, and MAE = 5.15 for SVR. Prediction performance dramatically improved compared to
Comparison of Prediction Accuracy with Existing Models. The model in the literature\(^2\) had small RMSE and high prediction accuracy; however, the computational load was very high due to DFT calculation. On the other hand, the model in the literature,\(^3\) which used the semiempirical molecular orbital method (AM1), had relatively larger RMSE, although the computational load was low.

The prediction accuracy of the proposed model was significantly improved by the addition of quantum chemical descriptors and variable selection using GA. Changing the basis set for DFT calculation to 6-31G could reduce the computational load while maintaining the prediction accuracy.

The addition of quantum chemical descriptors and appropriate optimization of molecular conformation before calculating the descriptors were important for a model with high accuracy. However, improvement in prediction accuracy reached a ceiling at some point, so improvement in prediction accuracy should be balanced with a computational load to create an effective model (Table 4).

Descriptors with High Impact on Prediction Accuracy. Top 15 descriptors with highest VIP scores are shown in Table 5. Various descriptors related to oxygen bonding (bond order, valence, and charge), and quantum chemical descriptors (LUMO and repulsion/attraction energy), are included in the table. The result shows that these descriptors obviously influenced the prediction accuracy of SADT.

Analyzing Effects of Preprocessing Including and Excluding DFT/GA. In no. 53, 54, 55, 59, 64, and 65, the prediction accuracy was improved by performing the DFT calculations before descriptor calculations, but in no. 57, 58, and 60, preprocessing by DFT calculations did not significantly improve the prediction accuracy. Additionally, no. 56, 62, and 63 showed high prediction accuracy from the beginning. In all the cases, except no. 62, variable selection via GA improved the prediction accuracy. Additionally, no. 56, 62, and 63 showed high prediction accuracy from the beginning. In all the cases, except no. 62, variable selection via GA improved the prediction accuracy. Additionally, no. 56, 62, and 63 showed high prediction accuracy from the beginning. In all the cases, except no. 62, variable selection via GA improved the prediction accuracy. Additionally, no. 56, 62, and 63 showed high prediction accuracy from the beginning.

No. 53, 54, 56, and 58 were compared as representatives (Figures 4–7). Conformation (bond length and angle) near the O–O bond of each molecule significantly changed after DFT calculation, as presented in Table 6. No. 53 and 54 exhibited relatively larger bond length changes than no. 56 and 58 due to the greater difference between the initial and optimal states for no. 53 and 54 than between no. 56 and 58. No. 56 exhibited a low prediction error for case 1, despite only 18 iterations, which could be due to the optimal initial conformation, where only MM calculations were performed. A large number of iterations yielded a high prediction accuracy, with some exceptions. Prediction error for no. 58 could not be reduced significantly.

Table 4. Comparison with Predictive Performance of Existing Models for Test Data

| geometry optimization | proposed method | Wang\(^2\) | He\(^3\) |
|-----------------------|-----------------|-----------|---------|
| frequency calculation | MM/DFT 6-31G/B3LYP | DFT 6-31G(d)/B3LYP | MM+/MO PM1 |
| variable selection    | GA-PLS          | GA-SVR    | GA      |
| number of descriptors | 559             | 521       | 8       |
| modeling method       | PLS             | SVR       | MLR     |
| number of training data | 52              | 40        | 57      |
| number of test data   | 13              | 10        | 14      |
| RMSE                  | 5.11            | 6.87      | 9.91    |

Figure 3. Actual vs calculated values of SADT in case 3 (left: PLS, right: SVR).

Table 3. Model Development Condition and Validation Results

|          | case 1          | case 2          | case 3          |
|----------|-----------------|-----------------|-----------------|
| descriptors | 5889 to >2659 | 1216 + 553 to >1586 | 1216 + 553 to >1586 |
| calculated by | alvaDesc 2 | alvaDesc 2 + CO-DESSA 3 | alvaDesc 2 + CODES- SA 3 |
| geometry optimization | MM (UFF)     | MM (UFF), DFT (6-31G/B3LYP) | MM (UFF), DFT (6-31G/B3LYP) |
| variable selection | GA-PLS | GA-SVR | 1586 to >559 |
| modeling method | PLS SVR | PLS SVR | PLS SVR |
| RMSE | 22.4 24.7 9.8 11.1 5.1 6.9 | 17.5 20.0 7.7 8.7 4.0 5.2 | 0.26 0.23 0.82 0.77 0.95 0.91 |
This could be due to the DFT calculation errors because 6-31G, instead of 6-31G(d), was used as the basis function.

Thus, appropriate optimization of molecular conformation and addition of quantum chemical descriptors immensely influence SADT prediction (Tables 7−10). Double Cross-Validation. Here, following a previous study, the data set was divided into training and test data to conduct holdout validation of the prediction accuracy. The temperature range of the test data was 15−70 °C, and it was unclear whether the model was applicable to a wider temperature range. To verify the extrapolation (generalization performance) for a high temperature range, using a small data set, double cross-validation was conducted using case 3 data, variables, and model building conditions. In inner cross-validation, hyperparameters were determined via fivefold cross-validation, whereas in outer cross-validation, leave-one-case-out cross-validation was performed.

### Table 5. Top 15 Descriptors with Highest VIP Scores

| no. | VIP | descriptor calculated by | explanation |
|-----|-----|---------------------------|-------------|
| 1   | 2.248 | AvgBondOrd_O | CODESSA | average bond order for all atoms of O type |
| 2   | 2.231 | MaxOneCent-ElecElecRepEn | CODESSA | maximum one-center electron−electron repulsion energy |
| 3   | 2.182 | SM02_EA(dm) | alvaDesc | spectral moment of order 2 from edge adjacency mat. weighted by dipole moment |
| 4   | 2.176 | SM08_EA(dm) | alvaDesc | spectral moment of order 8 from edge adjacency mat. weighted by dipole moment |
| 5   | 2.133 | MinOneCent-CoreElecAttrEn | CODESSA | minimum one-center core-electron attraction energy |
| 6   | 2.082 | SM07_EA(dm) | alvaDesc | spectral moment of order 7 from edge adjacency mat. weighted by dipole moment |
| 7   | 2.040 | MaxTwoCent-TotEn_AB | CODESSA | maximum two-center total energy, all bonds |
| 8   | 2.021 | SpMax_B(s) | alvaDesc | leading eigenvalue from Burden matrix weighted by I-state |
| 9   | 2.004 | AvgVal_O | CODESSA | average valence for atoms of O type |
| 10  | 1.978 | MaxTwoCent-CoreElecResEn_AP | CODESSA | maximum two-center core-electron resonance energy, all pairs |
| 11  | 1.971 | B02[O−O] | alvaDesc | presence/absence of O−O at topological distance 2 |
| 12  | 1.958 | AvgBondOrd_O_O | CODESSA | average bond order among all bonds between atoms of type O and O |
| 13  | 1.949 | qpmax | alvaDesc | maximum positive charge |
| 14  | 1.907 | MinTwoCent-CoreElecAttrEn_AP | CODESSA | minimum two-center core-electron attraction energy, all pairs |
| 15  | 1.861 | LUMOEn | CODESSA | energy of lowest energy unoccupied molecular orbital |

### Table 6. Comparison Results for Four Representative Molecules

| element | unit | no. 53 | no. 54 | no. 56 | no. 58 |
|---------|------|--------|--------|--------|--------|
| improving of prediction accuracy | °C | 19.9 | 52.9 | 1.0 | 3.1 |
| | % | 79.6 | 86.2 | 17.3 | 16.6 |
| iterations | times | 21 | 51 | 18 | 25 |
| length (O1−O2) | angstrom | +0.263 | +0.237 | +0.162 | +0.174 |
| length (C1−O1) | angstrom | +0.059 | +0.064 | +0.020 | +0.024 |
| length (C2−O2) | angstrom | +0.059 | +0.063 | +0.065 | +0.046 |
| angle(∠C1O1O2) | degree | −9.9 | −12.5 | −7.6 | −13.5 |
| angle(∠O1O2C2) | degree | −9.9 | −9.9 | −2.9 | −12.5 |

Figure 4. Comparison of prediction accuracy between PLS models.
out cross-validation was performed. Statistical parameters of the PLS model were as follows: \( R^2 = 0.76, \) RMSE = 17.74, and MAE = 13.29, and those of the SVR model were as follows: \( R^2 = 0.74, \) RMSE = 18.11, and MAE = 13.50. No significant difference in the prediction error at low and high temperatures was observed, and both the models evenly predicted a wide range of SADTs, with good accuracy (Figure 9).

### CONCLUSIONS

In this study, we constructed a model to estimate the SADT of organic peroxides. PLS regression and SVR were applied on the descriptors calculated using the structural information of the

---

**Table 7. Representative Bond Lengths and Angles before/after Optimization for No. 53**

| element unit | before optimization | after optimization | difference |
|--------------|---------------------|--------------------|------------|
| O1−O2        | 1.268               | 1.531              | +0.263     |
| C1−O1        | 1.359               | 1.418              | +0.059     |
| C2−O2        | 1.359               | 1.418              | +0.059     |
| ∠C1O1O2      | 124.5               | 114.6              | −9.9       |
| ∠O1O2C2      | 124.5               | 114.6              | −9.9       |

**Table 8. Representative Bond Lengths and Angles before/after Optimization for No. 54**

| element unit | before optimization | after optimization | difference |
|--------------|---------------------|--------------------|------------|
| O1−O2        | 1.273               | 1.510              | +0.237     |
| C1−O1        | 1.352               | 1.416              | +0.064     |
| C2−O2        | 1.353               | 1.416              | +0.063     |
| ∠C1O1O2      | 122.3               | 109.8              | −12.5      |
| ∠O1O2C2      | 120.5               | 110.4              | −9.9       |

**Table 9. Representative Bond Lengths and Angles before/after Optimization for No. 56**

| element unit | before optimization | after optimization | difference |
|--------------|---------------------|--------------------|------------|
| O1−O2        | 1.297               | 1.459              | +0.162     |
| C1−O1        | 1.360               | 1.380              | +0.020     |
| C2−O2        | 1.417               | 1.482              | +0.065     |
| ∠C1O1O2      | 125.0               | 117.4              | −7.6       |
| ∠O1O2C2      | 108.2               | 105.3              | −2.9       |

**Table 10. Representative Bond Lengths and Angles before/after Optimization for No. 58**

| element unit | before optimization | after optimization | difference |
|--------------|---------------------|--------------------|------------|
| O1−O2        | 1.272               | 1.446              | +0.174     |
| C1−O1        | 1.354               | 1.378              | +0.024     |
| C2−O2        | 1.352               | 1.398              | +0.046     |
| ∠C1O1O2      | 121.7               | 108.2              | −13.5      |
| ∠O1O2C2      | 121.1               | 108.6              | −12.5      |
compounds, to predict the SADTs. MM and DFT calculations were performed before calculating the descriptors, and GA was used for variable selection. In DFT calculation, B3LYP with the 6-31G basis set was used instead of 6-31G(d), significantly improving prediction accuracy and reducing computational load. Thus, a model with higher accuracy than the existing SADT prediction models was developed.

Appropriate preprocessing and variable selection were important for a model with high accuracy, and optimizing compound conformation before descriptor calculations improved SADT prediction accuracy. However, the improvement in prediction accuracy should be balanced with a computational load to create an effective model.

In the future, application of machine learning models other than SVR, preprocessing methods for descriptor calculations, and descriptor selection with high contribution to SADT prediction, to further improve prediction accuracy, will be investigated.

**AUTHOR INFORMATION**

**Corresponding Author**

Hiromasa Kaneko — Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan; orcid.org/0000-0001-8367-6476; Email: hkaneko@meiji.ac.jp

**Author**

Toshiharu Morishita — Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan; orcid.org/0000-0003-4150-2675

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.1c06481

**Notes**

The authors declare no competing financial interest.

Data and Software Availability: structures of chemical compounds were downloaded from the Chemical Book database.14 Some of the software used to calculate optimal conformation of molecules can be freely downloaded from the linked site.5–7 Python 3.7 was used to build the models and the source code was referenced to the linked site.15 All data underlying the results are available as part of the article and no additional source data are required.

**ACKNOWLEDGMENTS**

This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI) (grant number 19K15352) from the Japan Society for the Promotion of Science.

**ABBREVIATIONS**

SADT, self-accelerating decomposition temperature; QSPR, quantitative structure–property relationship; DFT, density functional theory; MLR, multiple linear regression; SVR, support vector regression; GA, genetic algorithm; MM, molecular mechanics; PLS, partial least squares; HOMO, highest occupied molecular orbital; LUMO, lowest unoccupied molecular orbital; VIP, variable importance in projection; RMSE, root mean square error; MAE, mean absolute error

**REFERENCES**

(1) Chen, W.-T.; Chen, W.-C.; You, M.-L.; Tsai, Y.-T.; Shu, C.-M. Evaluation of thermal decomposition phenomenon for 1,1-bis-(tertbutylperoxy)-3,3,5-trimethylcyclohexane by DSC and VSP2. J. Therm. Anal. Calorim. 2015, 122, 1125–1133.
(2) Wang, B.; Yi, H.; Xu, K.; Wang, Q. Prediction of the self-accelerating decomposition temperature of organic peroxides using QSPR models. J. Therm. Anal. Calorim. 2017, 128, 399–406.
(3) He, P.; Pan, Y.; Jiang, J.-c. Prediction of the self-accelerating decomposition temperature of organic peroxide based on support vector machine. Procedia Eng. 2018, 211, 215–225.
(4) Sun, J.; Li, Y.; Hasegawa, K. A study of self-accelerating decomposition temperature (SADT) using reaction calorimetry. J. Loss Prev. Process Ind. 2001, 14, 331–336.
(5) Avogadro Home Page. https://avogadro.cc/ (accessed 2021-10-11).
(6) Firefly Computational Chemistry Program Home Page. http://classic.chem.msu.su/gran/gamess/ (accessed 2021-10-11).
(7) SourceForge MoCalc2012 Download Page. https://sourceforge.net/projects/mocalc2012/ (accessed 2021-10-11).
(8) AlvaDesc Home Page. https://www.alvascience.com/alvadesc (accessed 2021-10-11).
(9) Semichem Page for Codessa III. http://www.semichem.com/codessa/codessa-new.php (accessed 2021-10-11).
(10) Akarachantachote, N.; Chadcham, S.; Saithanu, K. Cutoff Threshold of Variable Importance in Projection for Variable Selection. Int. J. Pure Appl. Math. 2014, 94, 307–322.
(11) Scikit-learn Home Page. https://scikit-learn.org/stable/ (accessed 2021-10-11).
(12) Scikit-learn GridSearchCV Page. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed 2021-10-11).
(13) Kaneko, H.; Funatsu, K. Fast optimization of hyperparameters for support vector regression models with highly predictive ability. *Chemom. Intell. Lab. Syst.* **2015**, *142*, 64–69.

(14) *Chemical Book Home Page*. [https://www.chemicalbook.com/ProductIndex_JP.aspx](https://www.chemicalbook.com/ProductIndex_JP.aspx) (accessed 2021-10-11).

(15) *GitHub Home Page*. [https://github.com/hkaneko1985/dcekit](https://github.com/hkaneko1985/dcekit) (accessed 2021-10-11).