ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Genetic Algorithm-Based Partial Least-Squares with Only the First Component for Model Interpretation

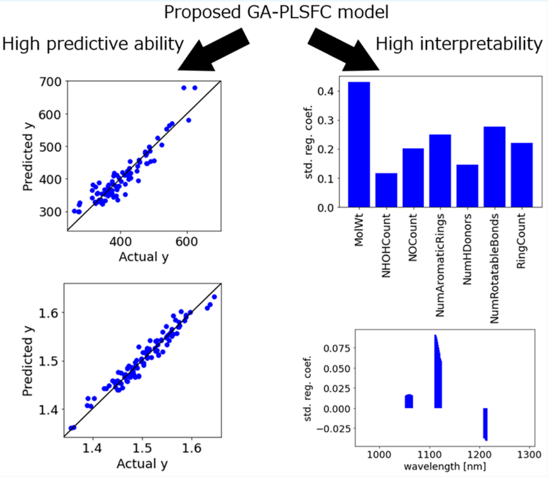
Hiromasa Kaneko*

Cite This: *ACS Omega* 2022, 7, 8968−8979

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: In the fields of molecular design, material design, process design, and process control, it is important not only to construct models with high predictive ability between explanatory variables X and objective variables y but also to interpret the constructed models to clarify phenomena and elucidate mechanisms in the fields. However, even in linear models, it is dangerous to use regression coefficients as contributions of X to y due to multicollinearity among X. Thus, the focus of this study is the model of partial least-squares with only the first component (PLSFC). It is possible to use regression coefficients as contributions of X to y for the PLSFC model. In addition, selecting the combination of X that can construct a predictive PLSFC model using a genetic algorithm (GA) is proposed, which is called GA-based PLSFC (GA-PLSFC). The constructed model would have both high predictive ability and high interpretability with regression coefficients that can be defined as contributions of X to y. The effectiveness of the proposed PLSFC and GA-PLSFC is verified using numerically simulated data sets and real material data sets. The proposed method was found to be capable of constructing predictive models with high interpretability. The Python codes for GA-PLSFC are available at https://github.com/hkaneko1985/dcekit.



1. INTRODUCTION

In molecular design, material design, process design, and process control, it is common to utilize regression model $y = f(X)$ constructed between objective variables y and explanatory variables X using a data set. One of the important things in regression analysis is to construct models with high predictive ability. Examples of linear modeling methods include partial least-squares regression (PLS),[1] ridge regression, and the least absolute shrinkage and selection operator (LASSO),[2] and nonlinear regression methods include support vector regression,[3] Gaussian process regression,[4] decision tree (DT),[5] random forests (RF),[6] gradient boosting (GB),[7−9] and deep neural networks.[10] Since there is no optimal regression analysis method, it is necessary to select a method that is appropriate for each data set.

However, it is also important to interpret the constructed regression models and clarify relationships between y and X, to elucidate the mechanism of expression of properties and activities, and to explain the phenomena. For example, the DT model can be interpreted as a relationship between y and X by using the combination of thresholds of X. In addition, feature importance in RF and GB can be used to determine the importance of each X in predicting y. Variable importance in projection[1] can be used for linear PLS. Although this importance is calculated considering the entire value of y, the importance of each X can vary depending on whether the y value is high, middle, or low. Shimizu and Kaneko proposed the DT and RF hybrid model that successfully interpreted global and local relationships between y and X.[11]

In addition to RF, there are other contribution indexes such as the local interpretable model-agnostic explanations (LIME)[12] and Shapley additive explanations (SHAP)[13] that can be combined with any regression analysis methods. In LIME and SHAP, by obtaining an approximation of the shape of the model at a certain sample point, the slope of X with respect to y around that sample point is obtained.

When relationships between y and X are linear, robust models can be constructed with linear regression methods. Furthermore, weights of X to y or regression coefficients are provided in linear models. However, it is dangerous to use regression coefficients as the contributions of X to y because there is multicollinearity among X. When X variables are highly correlated, the regression coefficient of an X variable becomes positively high and that of another correlated variable becomes

Received: December 30, 2021
Accepted: February 25, 2022
Published: March 4, 2022

negatively high, which has nothing to do with the true relationship between y and X. For example, when y is $\mathbf{y}^T = [2\ 4\ 6\ 8]$, X is $\mathbf{x}_1{}^T = [1\ 2\ 3\ 4]$, and $\mathbf{x}_2{}^T = [2\ 4\ 6\ 8]$, and multiple linear regression is conducted, there are infinite solutions of the regression coefficients $\mathbf{b}$, such as $\mathbf{b}^T = [2\ 0]$, $\mathbf{b}^T = [0\ 1]$, and $\mathbf{b}^T = [-10\ 4]$. It is inaccurate to consider the regression coefficient as the magnitude of the contribution of each variable to a model. When X is correlated, regression coefficients cannot be trusted.

Although active learning was applied to select informative samples and construct predictive PLS models,[14,15] other PLS methods such as orthogonal PLS,[16] biorthogonal PLS,[17] and kernel PLS[18] and sparse modeling methods such as LASSO, sparse PLS,[19] and envelope-based sparse PLS[20] were proposed. These contributed to the improvement of prediction performance and variable selection; the regression coefficients of X could not be interpreted. Therefore, regression coefficients can be used as the contributions of X to y only when there is no multicollinearity among X and when X is a single variable. Since the former is not realistic, this study focuses on regression analysis with one variable.

A PLS model with only the first component is applied. A regression coefficient of one component can be used as the contribution of the component to y because there is only one component in PLS. Therefore, the regression coefficients of X, calculated by using loading and weight vectors, can be the contributions of X to y. This method is called PLS with only the first component (PLSFC). In this study, the contributions of X to y are interpreted with regression coefficients using PLSFC.

However, as only one component is used in PLSFC, it is difficult to construct models with high predictive ability, especially when the number of X is large and noise variables are included in X. It is pointless to interpret a model with low predictive ability. Therefore, this study combines PLSFC and a genetic algorithm (GA),[21] selecting only important X variables, to construct a PLSFC model with high predictive ability. By setting the fitness in GA as the predictive accuracy of the PLSFC model verified with cross-validation, a set of X variables from which the predictive PLSFC model is constructed can be obtained, and their regression coefficients can be handled as contributions of X to y, indicating that the model is interpretable. When multiple linear regression, that is, PLS with all the components and PLS with more than one component are used instead of PLSFC, the regression coefficients of the constructed model cannot be interpreted. In addition, although the combination of GA and multiple linear regression has a high risk of overfitting, it is much lower in PLSFC, compared to a case where the number of components is optimized using cross-validation. This is because the PLSFC model, which has only one component, is simple. The proposed method is called a GA-based PLSFC (GA-PLSFC).

We focus on GA-based wavelength selection with PLS (GAWLS-PLS),[22] which selects variables as units of wavelength region for spectral analysis, and GA-based process variables and their dynamics selection with PLS (GAVDS-PLS),[23] which selects variables while considering dynamics in a process or time delays of process variables to y for soft sensor analysis as variable selection methods based on GA and PLS. The variable selection methods using PLSFC in GAWLS and GAVDS are called GAWLS-PLSFC and GAVDS-PLSFC, respectively. In this study, predictive ability and interpretability of the proposed methods are verified using numerical simulation data sets, molecular and material data sets, spectral data sets, and time series data sets.

This study makes the following contributions to the literature. First, it shows that regression coefficients of PLSFC model can be interpreted as the contributions of X to y. Second, GA-PLSFC allows us to construct a highly predictive model whose regression coefficients can be interpreted by selecting only the important X variables with GA. Third, GAWLS-PLSFC can be used to construct an interpretable model with high predictive ability by selecting the combination of wavelength regions in spectral analysis. Fourth, GAVDS-PLSFC can be used to construct an interpretable model with high predictive ability by selecting the important process variables and their time delays simultaneously in a time series data analysis or soft sensor analysis.

2. METHOD

2.1. Partial Least Squares with Only the First Component (PLSFC).

The basic equations for PLS are given as follows

$$\begin{aligned} X &= \sum_{i=1}^{a} t_i p_i^T + E \\ &= \mathrm{TP}^T + E \end{aligned} \tag{1}$$

$$\begin{aligned} \mathbf{y} &= \sum_{i=1}^{a} \mathbf{t}_i q_i + \mathbf{f} \\ &= \mathbf{T}q + \mathbf{f} \end{aligned} \tag{2}$$

where $\mathbf{X}$ and $\mathbf{y}$ are data sets of X and y after autoscaling, respectively, $\mathbf{t}_i$ and $\mathbf{p}_i$ are score and loading vector of ith component, respectively, $\mathbf{E}$ and $\mathbf{f}$ denote the error matrix of X and error vector of y, respectively, q_i is a regression coefficient of ith component, and a is the number of components. Assuming that $\mathbf{t}_1$ is represented with a linear combination of X, $\mathbf{t}_1$ is given as follows

$$\mathbf{t}_1 = \sum_{i=1}^{m} w_1^{(i)} \mathbf{x}_i = \mathbf{X}\mathbf{w}_1 \tag{3}$$

where m is the number of X, $\mathbf{x}_i$ is the vector of the ith X, $w_1{}^{(i)}$ is the weight of $\mathbf{x}_i$ to $\mathbf{t}_1$, and $\mathbf{w}_1$ is a vector whose elements are $w_1{}^{(i)}$. As a constraint, the magnitude of $\mathbf{w}_1$ is set to one as follows:

$$\sum_{i=1}^{m} (w_1^{(i)})^2 = 1 \tag{4}$$

While satisfying eq 4, $\mathbf{w}_1$ is calculated to maximize the covariance or inner product of $\mathbf{y}$ and $\mathbf{t}_1$, which is given as follows

$$\mathbf{w}_1 = \frac{\mathbf{X}^T\mathbf{y}}{\|\mathbf{X}^T\mathbf{y}\|} \tag{5}$$

where $\mathbf{w}_1$ is calculated from $\mathbf{X}$ and $\mathbf{y}$ and then $\mathbf{t}_1$ is calculated with eq 3.

$p_1{}^{(i)}$, which corresponds to ith X in $\mathbf{p}_1$, is obtained by minimizing the sum of the squares of the errors in $\mathbf{x}_1$, and q_1 is obtained by minimizing the sum of the squares of the errors in $\mathbf{y}$, as follows:

$$\mathbf{p}_1 = \frac{\mathbf{X}^T\mathbf{t}_1}{\mathbf{t}_1^T\mathbf{t}_1} \tag{6}$$

$$q_1 = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{t}_1}{\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1} \tag{7}$$

The parts of X and y that cannot be represented with the first component are denoted by $\mathbf{X}_1$ and $\mathbf{y}_1$, respectively, which are given as follows:

$$\mathbf{X}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} \tag{8}$$

$$\mathbf{y}_1 = \mathbf{y} - q_1\mathbf{t}_1 \tag{9}$$

By using $\mathbf{X}_1$ and $\mathbf{y}_1$ as $\mathbf{X}$ and $\mathbf{y}$ in the previous equations, respectively, $\mathbf{w}_2$, $\mathbf{t}_2$, $\mathbf{p}_2$, q_2, $\mathbf{X}_2$, and $\mathbf{y}_2$ can be calculated for the second component as well, and then $\mathbf{w}_i$, $\mathbf{t}_i$, $\mathbf{p}_i$, q_i, $\mathbf{X}_i$, and $\mathbf{y}_i$ are calculated for the third and subsequent components in turn.

Standardized regression coefficient $\mathbf{b}$ for PLS is given as follows

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{q} \tag{10}$$

where $\mathbf{W}$ is a matrix including $\mathbf{w}_1$, $\mathbf{w}_2$, and so on.

It is necessary to determine the number of components used. In general, cross-validation is performed for each number of components in PLS, and the number of components that maximize the coefficient of determination r^2 between actual y values and y values predicted in cross-validation is used. However, in the proposed PLSFC, only the first component is used.

In this study, the PLS calculation uses cross_decomposition.PLSRegression[24] from the scikit-learn library.

2.2. Genetic Algorithm-Based PLSFC (GA-PLSFC).

Given that only the first component is used in PLSFC, it is not possible to calculate the informative component to explain y when multiple X-variables exist and noise variables are included in X. A model is proposed to select a set of important X variables that can construct a PLSFC model with high predictive ability using GA, which is called GA-PLSFC. GA is a metaheuristic inspired by the process of natural selection. GA prepares several chromosomes representing candidates for selected X numbers as genes and searches for solutions by selecting chromosomes with high fitness values preferentially and repeating operations such as crossover and mutation.

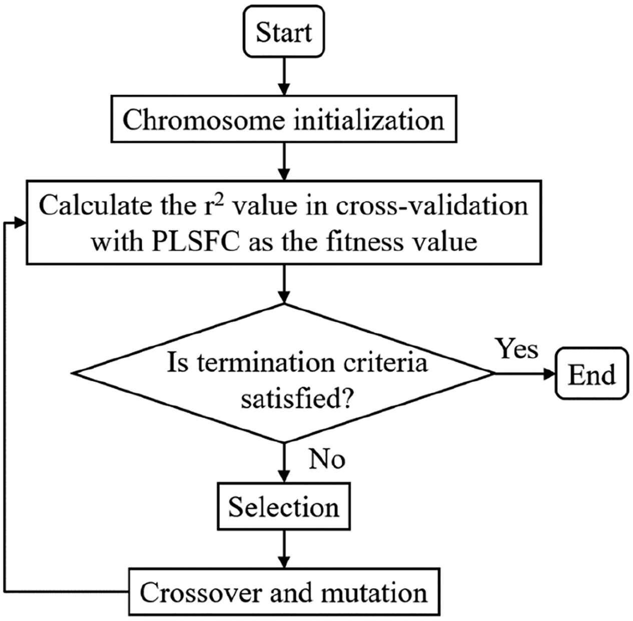

Start
Chromosome initialization
Calculate the r² value in cross-validation with PLSFC as the fitness value
Is termination criteria satisfied?
Yes
End
No
Selection
Crossover and mutation

Figure 1. Flowchart of GA-PLSFC modeling.

Figure 1 shows the flowchart of GA-PLSFC modeling. While the fitness in GA for GA-PLS is the maximum value of r^2 between actual y values and y values predicted with cross-validation in adding components of PLS, the fitness in GA for the proposed GA-PLSFC is the r^2 value between the actual y values and y values predicted in cross-validation with PLSFC. GA-PLSFC provides a set of X variables with high r^2 after cross-validation, and the contributions of X to y can be obtained by interpreting the regression coefficients of the PLSFC model with the X variables selected with GA-PLSFC.

DEAP[25] was used to calculate the GA. The Python code for GA-PLSFC is available at https://github.com/hkaneko1985/dcekit.

2.3. GAWLS-PLSFC and GAVDS-PLSFC.

GAWLS-PLS is a GA-based wavelength selection method for spectral analysis that can optimize the combination of wavelength regions. When the number of wavelength regions to be selected is k, a chromosome in GA is represented by k sets, where the first wavelength number and width to be selected comprise one set. Accordingly, a chromosome is a sequence of $2k$ values.

The fitness in GA for GAWLS-PLS is the maximum value of r^2 between the actual y values and y values predicted in cross-validation while adding components of PLS. The fitness in GA for the proposed method GAWLS-PLSFC is the r^2 value between the actual y values and y values predicted in cross-validation with PLSFC. GAWLS-PLSFC provides a set of wavelengths with high r^2 after cross-validation, and the contributions of X to y can be obtained by interpreting the regression coefficients of the PLSFC model with the X variables selected with GAWLS-PLSFC.

The GAVDS is a variable selection method based on the GA, which can optimize the process variables and their time delays simultaneously. Furthermore, time-delayed variables are selected continuously in time. For each process variable, time-delayed variables are prepared for each measurement time; when m process variables and time-delayed variables are considered within a unit time of n, the number of all variables is equal to $m \times (n + 1)$. When the time width that is selected for a process variable indicates a region, the number of selected regions and maximum value of regions is set.

The fitness in GA for GAVDS-PLS is the maximum value of r^2 between actual y values and y values predicted in cross-validation while adding components of PLS, and the fitness in GA for the proposed method GAVDS-PLSFC is the r^2 value between the actual y values and y values predicted in cross-validation with PLSFC. GAVDS-PLSFC provides a set of process variables and their time delays with high r^2 after cross-validation, and the contributions of X to y can be obtained by interpreting the regression coefficients of the PLSFC model with the X variables selected with the GAVDS-PLSFC.

The GAWLS-PLS, GAWLS-PLSFC, GAVDS-PLS, and GAVDS-PLSFC programs in this study were written in Python[26] using DEAP.[25]

3. RESULTS AND DISCUSSION

To verify the predictive ability of the proposed methods, two numerical simulation data sets were used first. To compare PLS and PLSFC, the first data set (SIM1) used was a data set of 100 mutually correlated X and y that is linearly related to X. The ith X, X_i, was generated as follows:

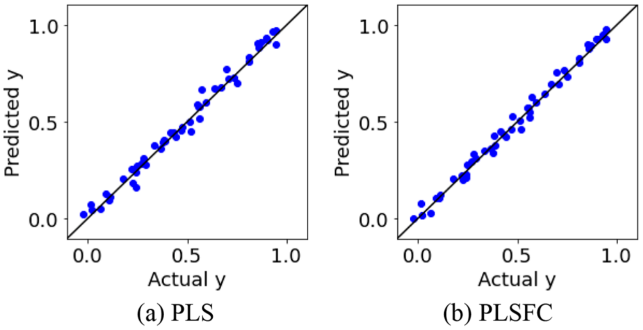
$$X_i = z + 0.2 \times \mathrm{std}(z) \times N(0{,}1) \tag{11}$$

Here, z denotes uniform random numbers between 0 and 1, std(z) is the standard deviation of z, and $N(0, 1)$ denotes standard normal random numbers. y is set to z plus standard normal random number with 10% of its standard deviation as noise. From the above operations, the contributions of X for y are equivalent for all X, and 50 samples of training data and 50 samples of test data were generated.

The prediction results for test data in PLS and PLSFC are shown in Table 1 and Figure 2. r^2_{TEST} is r^2 for test data, and RMSE$_{\mathrm{TEST}}$ is the root-mean-squared error for test data. There is no significant difference between PLS and PLSFC in r^2_{TEST} and RMSE$_{\mathrm{TEST}}$. The RMSE$_{\mathrm{TEST}}$ of PLSFC was slightly lower than that of PLS because the PLS model was overfitted by optimizing the number of components with cross-validation. The plots of actual y values and predicted y values indicate that the samples of both PLS and PLSFC are clustered around the diagonal and the y values could be predicted with high accuracy. It was confirmed that PLSFC model had high predictive ability.

Table 1. Prediction Results in the Test Data of SIM1

	PLS	PLSFC
r^2_{TEST}	0.984	0.989
RMSE$_{\mathrm{TEST}}$	0.0354	0.0294



(a) PLS (b) PLSFC

Figure 2. Actual y vs predicted y in the test data of SIM1.

The standardized regression coefficients for PLS and PLSFC are shown in Figure 3. Although the standardized regression coefficients in the PLS model were varied positively and negatively, they are not appropriate because the data set of SIM11 was generated such that the contributions of X for y were the same for all X. This would occur due to the collinearity among X. Thus, it is dangerous to interpret the standardized regression coefficients of the general PLS model as the contributions of X to y. Conversely, Figure 3b shows that the standardized regression coefficients of the PLSFC model are almost the same for all X, which is consistent with the fact that the data set of SIM1 was generated such that the contributions of X for y were the same for all X. Hence, it was confirmed that the standardized regression coefficients of PLSFC could be used to properly determine the contributions of X to y.

To compare PLS, PLSFC, GAWLS-PLS, and GAWLS-PLSFC, a second data set (SIM2) was prepared assuming spectral data. X was generated as follows:

$$X_i = \begin{cases} U(0,\ 1) & (i = 1) \\ 0.95 \times X_{i-1} + 0.05 \times U(0,\ 1) & (i = 2, 3, ..., 500) \end{cases} \tag{12}$$

Here, $U(0, 1)$ denotes uniform random numbers between 0 and 1. The coefficients of the linear combination of X with respect to y were generated, as shown in Figure 4, using the probability density function of the normal distribution. Standard normal random numbers with 10% of the standard deviation of y was added to y as noise, and 200 samples of training data and 100 samples of test data were generated.

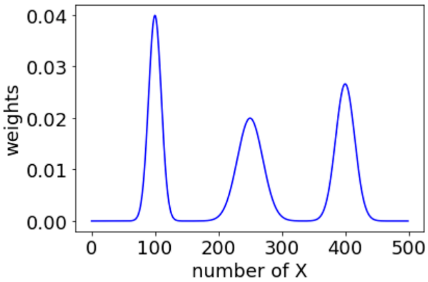
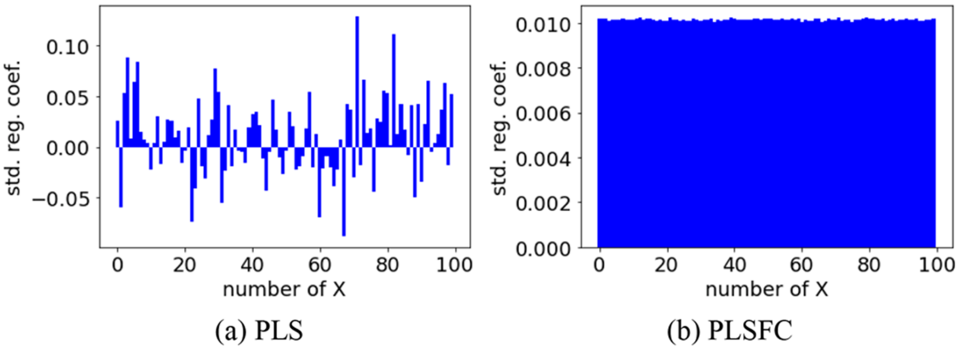


Figure 4. Weights of X to y in SIM2.

The prediction results of test data in PLS, GAWLS-PLS, PLSFC, and GAWLS-PLSFC are shown in Table 2 and Figure 5.

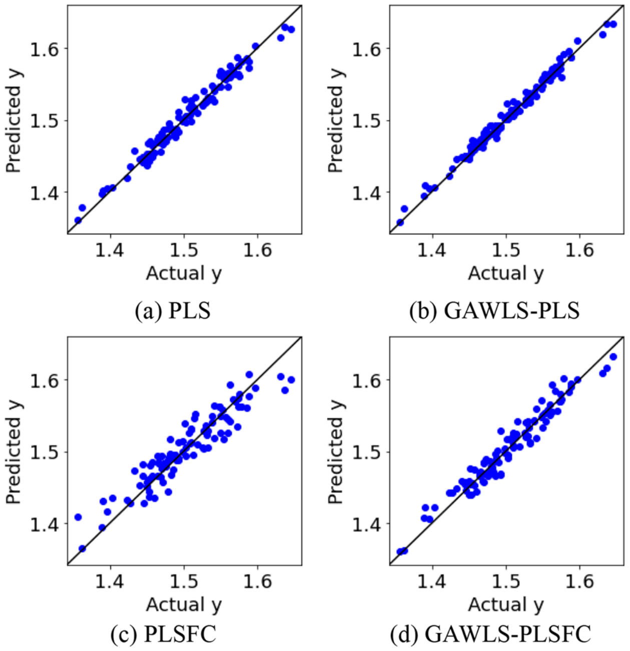
Table 2. Prediction Results in the Test Data of SIM2

	PLS	GAWLS-PLS	PLSFC	GAWLS-PLSFC
r^2_{TEST}	0.970	0.981	0.872	0.945
RMSE$_{\mathrm{TEST}}$	0.0098	0.0078	0.0203	0.0133



(a) PLS (b) PLSFC

Figure 3. Standardized regression coefficients in SIM1.



(a) PLS (b) GAWLS-PLS
(c) PLSFC (d) GAWLS-PLSFC

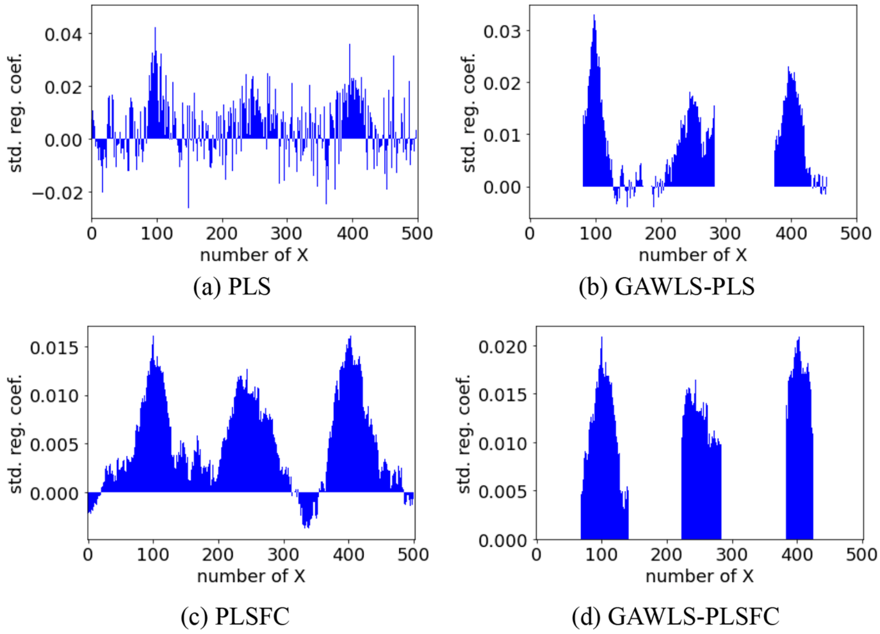
Figure 5. Actual *y* vs predicted *y* in the test data of SIM2.

The r^2_{TEST} of GAWLS-PLSFC is much higher than that of PLSFC and the $\text{RMSE}_{\text{TEST}}$ of GAWLS-PLSFC is much lower than that of PLSFC, indicating that the predictive ability of PLSFC model is greatly improved by selecting wavelengths with GAWLS. Although the r^2_{TEST} of GAWLS-PLSFC is lower than that of PLS and that of GAWLS-PLS, the samples in Figure 5d are distributed close to the diagonal line in the whole *y* values. Thus, it was confirmed that the predictive ability of PLSFC model can be successfully improved by selecting variables using GAWLS.

Figure 6 shows the standardized regression coefficients for PLS, GAWLS-PLS, PLSFC, and GAWLS-PLSFC. While the actual weights of *X* for *y* change continuously for each *X* number as shown in Figure 4, the standardized regression coefficients of the PLS model were discontinuous values and those of GAWLS-PLS also varied positively and negatively in each wavelength region, which could be due to the effect of collinearity among *X*. This confirmed that the standardized regression coefficients of the general PLS model were not appropriate as the contributions of *X* to *y*. However, Figure 6c shows that standardized regression coefficients are continuous to a certain extent with respect to the number of *X* in PLSFC, which is similar to the weights in Figure 4. Although original weights of *X* are zero or nearly zero and their *X* variables do not affect *y*, *y* values change due to noise, which means that *y* values increase in spite of little change of the *X* values, and thus, the standardized regression coefficients of the *X* variables become negative. In addition, GAWLS-PLSFC accurately selects only the regions of *X* with large weights in Figure 4, and all the values of the standardized regression coefficients are appropriately given as positive. Therefore, it is confirmed that the proposed methods can construct models with high predictive ability and that standardized regression coefficients for PLSFC and GAWLS-PLSFC can be used to interpret the relationship between *X* and *y* appropriately.

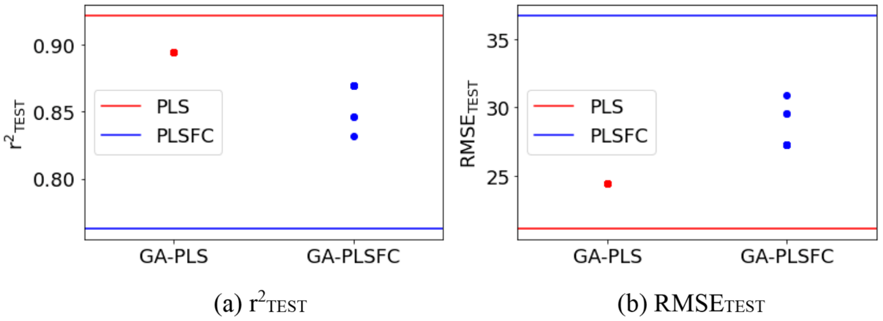
Next, to verify the predictive ability of the proposed methods, data sets of boiling points (BP),[27] solubility in water (logS),[28] and environmental toxicity (Tox)[29] were used as data sets for compounds or materials. The tablet data set of Shootout2002[30] (API1) and the tablet data set of Shootout2012[31] (API2) were used as spectral data sets. The time series data sets of a debutanizer column (DEB)[32] and a sulfur recovery unit (SRU)[32] were used as process data sets. These data sets are real data.

In the data sets for compounds, RDKit[33] was used to calculate the molecular descriptors. To test the interpretability of the model, another BP data set with only interpretable descriptors was prepared, which is referred to as BP(selected). Time-delayed *X*-variables were added to *X* from 1−60 unit time for DEB and SRU. The data were randomly split so that the training set contained 70% of samples and the testing set contained the remaining 30% for BP(selected), BP, logS, Tox, API1, and API2,



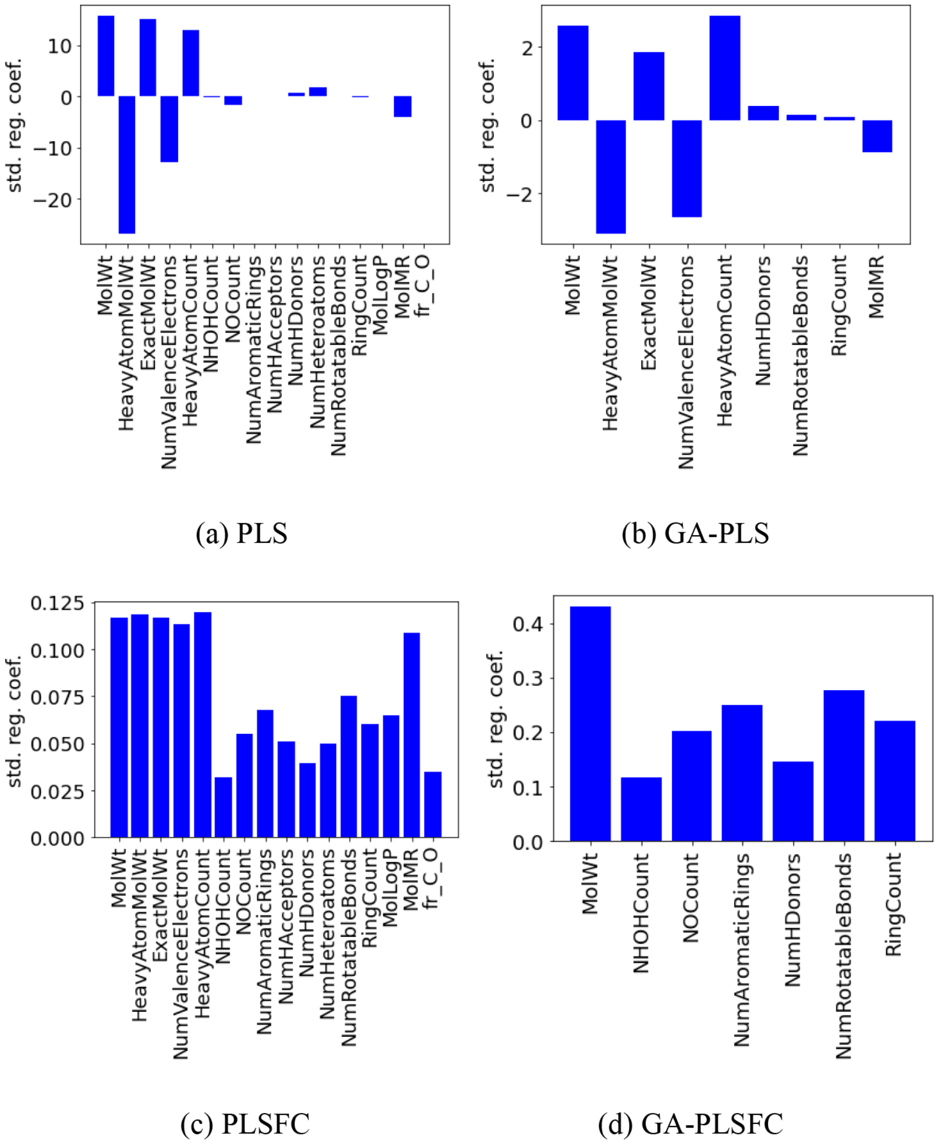
(a) PLS (b) GAWLS-PLS
(c) PLSFC (d) GAWLS-PLSFC

Figure 6. Standardized regression coefficients in SIM2.



(a) r^2_{TEST} (b) $\text{RMSE}_{\text{TEST}}$

Figure 7. Prediction results in the test data of BP (selected). Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GA-PLS, and GA-PLSFC, respectively.

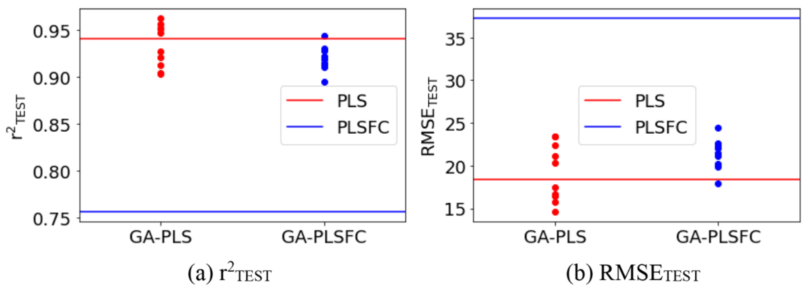


(a) PLS (b) GA-PLS

(c) PLSFC (d) GA-PLSFC

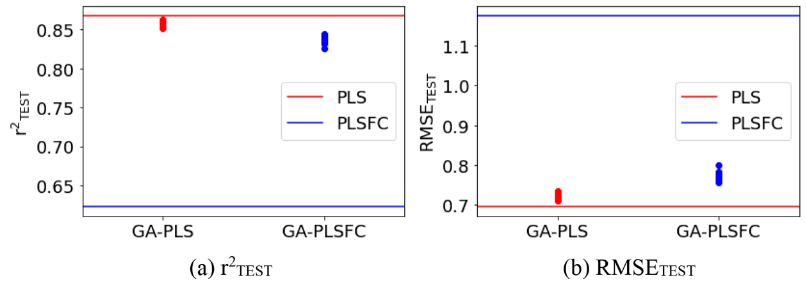
Figure 8. Standardized regression coefficients in BP(selected).

and the data were randomly split so that the training set contained 1000 samples and the testing set contained the remaining samples for DEB and SRU. Those X variables for which the ratio of samples with the same values in the training data accounted for 80% or more were deleted. One of the pairs of X variables for which the absolute value of the correlation coefficient attained a value of 1 was subsequently deleted. Each variable selection of the GA-PLS and GA-PLSFC was conducted ten times. In the GAWLS-PLS, GAWLS-PLSFC, GAVDS-PLS, and GAVDS-PLSFC, the number of regions was set to 1−8, and each variable selection was conducted 10 times.



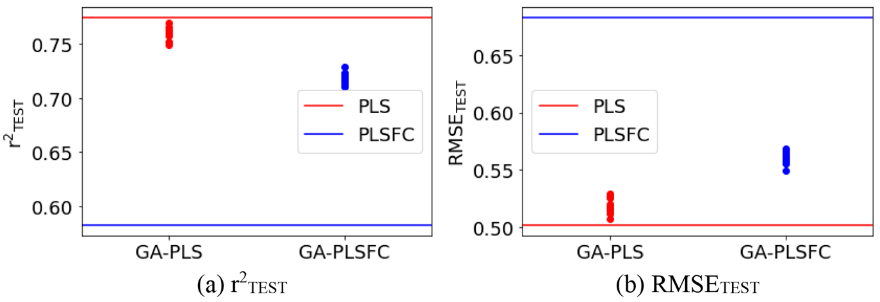
(a) r^2_{TEST} (b) $RMSE_{TEST}$

Figure 9. Prediction results in the test data of BP. Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GA-PLS, and GA-PLSFC, respectively.



(a) r^2_{TEST} (b) $RMSE_{TEST}$

Figure 10. Prediction results in the test data of logS. Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GA-PLS, and GA-PLSFC, respectively.



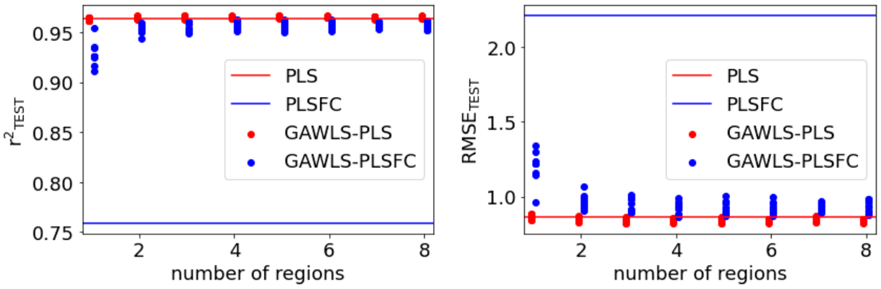
(a) r^2_{TEST} (b) $RMSE_{TEST}$

Figure 11. Prediction results in the test data of Tox. Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GA-PLS, and GA-PLSFC, respectively.

Figure 7 shows the prediction results of test data for PLS, GA-PLS, PLSFC, and GA-PLSFC in BP(selected). Although there were ten results for GA-PLS and GA-PLSFC, some results had the same r^2_{TEST} and $RMSE_{TEST}$; GA-PLS and GA-PLSFC appeared to have one and three points, respectively. Figure 7 indicates that the r^2_{TEST} of GA-PLSFC was higher than that of PLSFC and the $RMSE_{TEST}$ of GA-PLSFC was lower than that of PLSFC, and predictive ability was improved by using GA-PLSFC. Although the r^2_{TEST} of GA-PLS was lower than that of PLS and the $RMSE_{TEST}$ of GA-PLS was higher than that of PLS, indicating overfitting due to GA-PLS, GA-PLSFC improved the prediction accuracy over PLSFC without overfitting. In GA-PLS, X variables would be selected to fit the cross-validation result, which led to the overfitting of training data, compared to PLS. The plots of actual y values and predicted y values for test data in each method are available in the Supporting Information. Hence, it was confirmed that GA-PLSFC can appropriately select X variables to improve predictive ability of the PLSFC model.

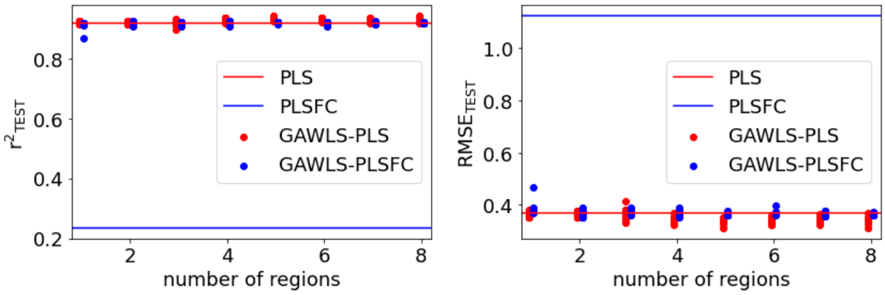
Figure 8 shows the standardized regression coefficients for PLS, GA-PLS, PLSFC, and GA-PLSFC. In the PLS and GA-PLS, the standardized regression coefficient of HeavyAtomMolWt (molecular weight including only heavy atoms) was negative, for example, even though molecular weight should contribute positively to boiling point. This inconsistence would occur due to the collinearity among X. It was confirmed that it is dangerous to use the standardized regression coefficients of PLS and GA-PLS models as the contributions of X to y. Additionally, absolute standardized regression coefficients were high for PLS and GA-PLS. When a regression coefficient is high, the predicted y value for the same value of X corresponding to the regression coefficient becomes high, and the extrapolation regions of X indicate high values. Thus, the predicted y value can be an outlier. It difficult to predict y values for extrapolation regions of X using the PLS and GA-PLS models.

Meanwhile, from Figure 8c, the standardized regression coefficients of PLSFC indicate that the coefficients of descriptors that are considered to contribute positively to boiling point, including molecular weight-related descriptors, are appropriately positive. In addition, the standardized regression coefficients of the descriptors selected by GA-PLSFC are also appropriately positive. This confirmed that the proposed methods can properly handle the standardized regression coefficient as the contributions of X to y.



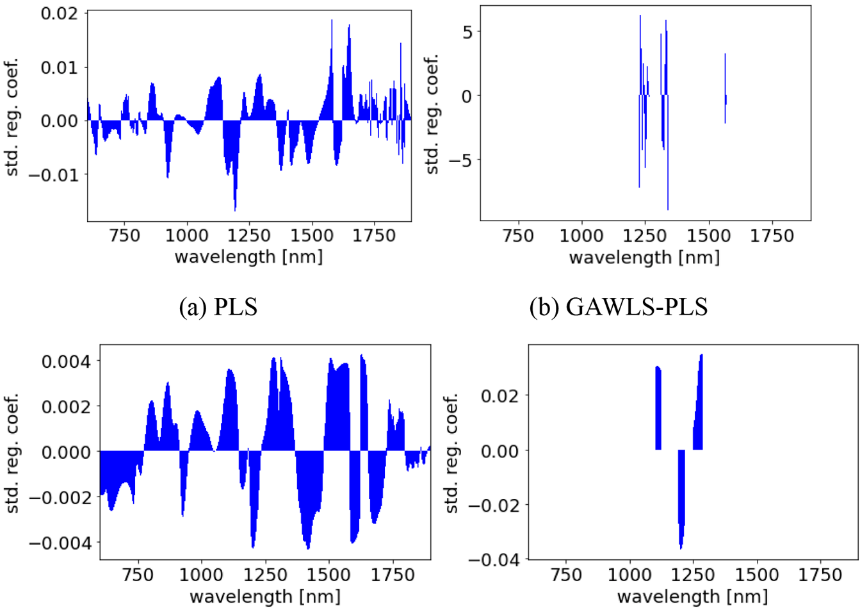
(a) r^2_{TEST} (b) $RMSE_{TEST}$

Figure 12. Prediction results in the test data of API1. Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GAWLS-PLS, and GAWLS-PLSFC, respectively.



(a) r^2_{TEST} (b) $RMSE_{TEST}$

Figure 13. Prediction results in the test data of API2. Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GAWLS-PLS, and GAWLS-PLSFC, respectively.

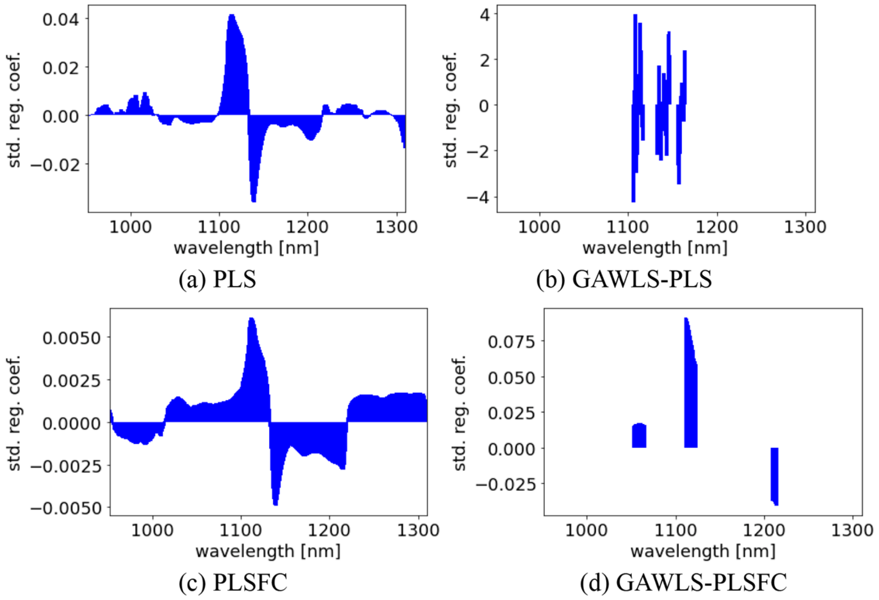


(a) PLS (b) GAWLS-PLS

© PLSFC (d) GAWLS-PLSFC

Figure 14. Standardized regression coefficients in API1. There are three regions in GAWLS-PLS and GAWLS-PLSFC.

The prediction results of test data for BP, logS, and Tox are shown in Figures 9, 10, and 11, respectively. Ten GA calculations were performed, and accordingly, there were 10 results for each data set for GA-PLS and GA-PLSFC. In all of the data sets, the r^2_{TEST} of GA-PLSFC was higher than that of PLSFC and the $RMSE_{TEST}$ was lower than that of PLSFC, confirming that variable selection with GA could appropriately improve the predictive ability of PLSFC models. Although the r^2_{TEST} of GA-PLS was higher or lower than that of PLS depending on the data set, GA-PLSFC showed consistently higher r^2_{TEST} compared to PLSFC, confirming that GA-PLSFC could select variables stably that improved the predictive ability. Furthermore, for BP and logS, the results indicated that GA-PLSFC outperformed PLS and GA-PLS in terms of predictive accuracy. The plots of actual y values and predicted y values for test data in each method are available in the Supporting



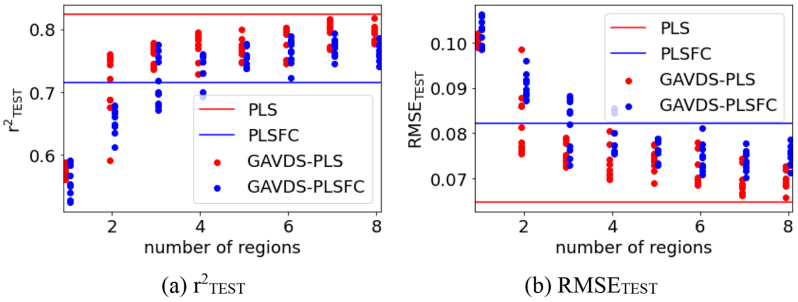
(a) PLS

(b) GAWLS-PLS

(c) PLSFC

(d) GAWLS-PLSFC

Figure 15. Standardized regression coefficients in API2. There are three regions in GAWLS-PLS and GAWLS-PLSFC.



(a) r^2_{TEST}

(b) $\mathrm{RMSE}_{\mathrm{TEST}}$

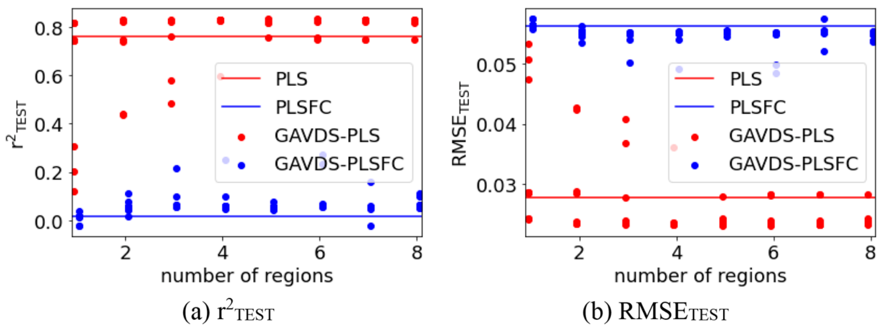
Figure 16. Prediction results in the test data of DEB. Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GAVDS-PLS, and GAVDS-PLSFC, respectively.

Information. Therefore, it was confirmed that the proposed methods could construct linear model with both high predictive ability and high interpretability.

The prediction results of test data in PLS, GAWLS-PLS, PLSFC, and GAWLS-PLSFC for API1 and API2 are shown in Figures 12 and 13, respectively. Since the GA was calculated ten times for each number of regions, there were 10 results for GAWLS-PLS and GAWLS-PLSFC for each region number. The r^2_{TEST} was improved and $\mathrm{RMSE}_{\mathrm{TEST}}$ was reduced greatly by selecting wavelengths with GAWLS-PLSFC, compared to PLSFC. In particular, the average of prediction *y* errors was reduced to less than half of those in PLSFC. Furthermore, GAWLS-PLSFC produced results comparable to those of PLS and GAWLS-PLS. The plots of actual *y* values and predicted *y* values for test data in each method are available in the Supporting Information. This confirmed that the predictive model could be constructed by using the proposed GAWLS-PLSFC.

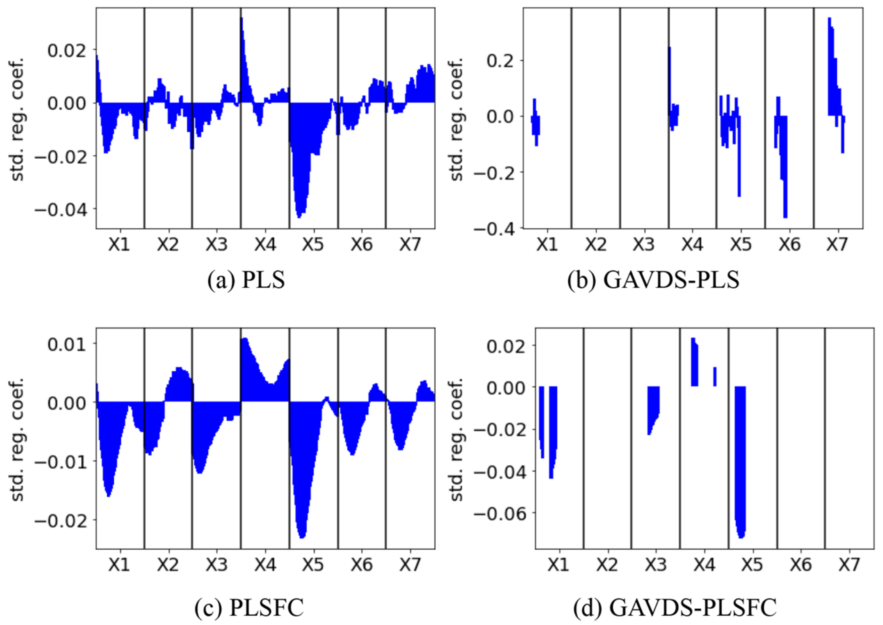
Figures 14 and 15 show standardized regression coefficients for PLS, GAWLS-PLS, PLSFC, and GAWLS-PLSFC in API1 and API2, respectively. GAWLS-PLS and GAWLS-PLSFC depict the results when the number of regions is three. In both API1 and API2, compared to the standardized regression coefficients of PLS, those of PLSFC change smoothly according to wavelength numbers, and the smoothness is appropriate when considering the spectral shape. Additionally, the absolute standardized regression coefficients were lower in PLSFC than in PLS, and overfitting is less likely to occur. Although there were several wavelengths where the standardized regression coefficients were negative, it would not be a problem. In the pure spectrum of the target component to be predicted as *y*, even when the intensity of a certain wavelength is zero or nearly zero, the *X* values of mixture spectra will become large for low *y* values, that is, a small amount of the target component, since the amount of the other components is large. Conversely, when *y* values are high, the *X* values of mixture spectra become small since the amount of the other components is small. These mean that there is a negative correlation between *y* and *X* at the corresponding wavelength, and the standardized regression coefficient can be negative.

Figures 14b and 15b show that when wavelengths were selected with GAWLS-PLS, the standardized regression coefficients were not consistent between positive and negative values in the same wavelength region, and the absolute standardized regression coefficients were exceptionally large, which could be because of the influence of collinearity among *X*. Hence, it is difficult to use the standardized regression coefficient as the contributions of *X* to *y*. Furthermore, the large absolute standardized regression coefficients make prediction unstable, especially for new samples in extrapolation of *X*. Meanwhile, from Figures 14d and 15d, the positive and negative standardized regression coefficients for GAWLS-



(a) r^2_{TEST} (b) $RMSE_{TEST}$

Figure 17. Prediction results in the test data of SRU. Red lines, blue lines, red points, and blue points indicate the results of the PLS, PLSFC, GAVDS-PLS, and GAVDS-PLSFC, respectively.



(a) PLS (b) GAVDS-PLS

(c) PLSFC (d) GAVDS-PLSFC

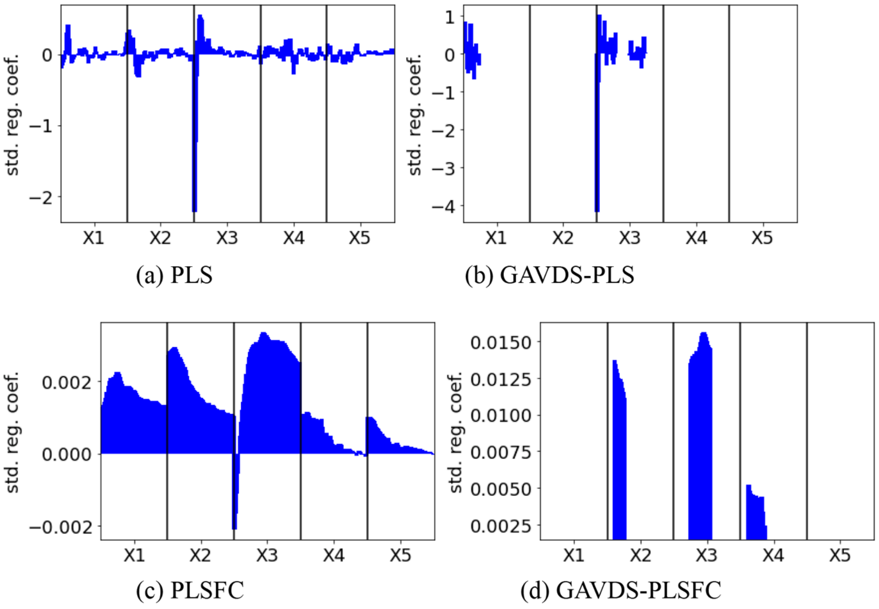
Figure 18. Standardized regression coefficients in DEB. For X1−X7, standardized regression coefficients of time-delayed variables are shown in order from left to right. There are six regions in GAVDS-PLS and GAVDS-PLSFC.

PLSFC were consistent for each selected wavelength region, suggesting that the weights of X to y can be calculated appropriately. Furthermore, the absolute standardized regression coefficients were not large, which suggests that new samples in extrapolation of X can be predicted stably. Thus, it was confirmed that the proposed methods can construct models in spectral analysis with both high predictive ability and high interpretability.

The prediction results of test data for PLS, GAVDS-PLS, PLSFC, and GAVDS-PLSFC in DEB and SRU are shown in Figures 16 and 17, respectively. Since 10 GA calculations were performed for each number of regions, there were 10 results for each region in GAVDS-PLS and GAVDS-PLSFC. By selecting process variables and their dynamics with GAVDS-PLSFC, the r^2_{TEST} was improved and $RMSE_{TEST}$ was reduced, compared to PLSFC. In DEB, GAVDS-PLSFC achieved predictive accuracy comparable to GAVDS-PLS; however, in SRU, the r^2_{TEST} of GAVDS-PLSFC was much lower than that of PLS and GAVDS-PLS. In SRU, y values varied due to various effects of process variables, and the variations could be handled with the one-component model even after the selection of process variables and their dynamics. The plots of actual y values and predicted y values for test data in each method are available in the Supporting Information. Thus, it was confirmed that the prediction accuracy of the PLSFC model can be improved by appropriately selecting the process variables and their time delays simultaneously.

Figures 18 and 19 show the standardized regression coefficients for PLS, GAVDS-PLS, PLSFC, and GAVDS-PLSFC in DEB and SRU, respectively. For X1, X2, and so on, standardized regression coefficients of time-delayed variables are shown in order from left to right. The number of regions in both GAVDS-PLS and GAVDS-PLSFC is six for DEB and four for SRU. In both DEB and SRU, the standardized regression coefficients change more smoothly according to time delays in PLSFC than in PLS, and the smoothness is appropriate when considering the dynamic characteristics of the process. In SRU, the absolute standardized regression coefficients of PLSFC are lower than those of PLS, which indicates overfitting is less likely to occur for PLSFC than PLS.

Figures 18b and 19b show that when process variables and their dynamics were selected with GAVDS-PLS the standardized regression coefficients were not consistent in terms of positive and negative values for each of the same process



(a) PLS

(b) GAVDS-PLS

(c) PLSFC

(d) GAVDS-PLSFC

Figure 19. Standardized regression coefficients in SRU. For X1, X2, ..., and X5, standardized regression coefficients of time-delayed variables are shown in order from left to right. There are four regions in GAVDS-PLS and GAVDS-PLSFC.

variables, and the absolute values became exceptionally large. Because the influence of collinearity among X, it was dangerous to use the standardized regression coefficients as the contributions of X to y. Furthermore, the large absolute coefficients make prediction unstable, especially for new samples in extrapolation of X. Conversely, for GAVDS-PLSFC in Figures 18d and 19d, the positive and negative standardized regression coefficients are consistent for each of the selected process variables, suggesting that the contributions of X to y were appropriately calculated. Furthermore, since the absolute standardized regression coefficients were not large, y values of new samples in extrapolation of X would be predicted stably. In process data analysis or soft sensor analysis, it was hence confirmed that the proposed methods can be used to construct models with both high predictive ability and high interpretability.

4. CONCLUSION

This study focused on the PLSFC model with the aim of constructing linear regression models with high predictive ability and high interpretability. PLSFC has only one component, and standardized regression coefficients can be used as the contributions of X to y, which means high interpretability. However, it is difficult to construct models with high predictive ability. To improve the prediction accuracy of PLSFC, GA-PLSFC, a method to select only the X variables that are important for the PLSFC model with GA, was proposed. By using GA-PLSFC and selecting the combination of X variables that can construct predictive PLSFC model, linear regression models with both high predictive ability and high interpretability can be constructed. The proposed method for spectral analysis is GAWLS-PLSFC, and the proposed method for process data analysis or soft sensor analysis is GAVDS-PLSFC.

The proposed methods were validated using compound data sets, spectral data sets, and time series data sets, and it was confirmed that the prediction accuracy of PLSFC could be improved for all data sets by using the proposed methods. In addition, the proposed methods could perform the predictive ability of linear regression models that were comparable to or surpassed the predictive ability of PLS models that were not limited to a single component, depending on the data set. Furthermore, even when the standardized regression coefficients of the conventional PLS models differed from scientific backgrounds of the data sets, the proposed methods obtained the standardized regression coefficients that were consistent with the scientific backgrounds. It was confirmed that the proposed methods can construct linear regression models with high predictive ability and that the standardized regression coefficients of the constructed models can be used as the contributions of X to y.

However, the proposed method is a linear method and cannot represent nonlinear relationships between X and y. In addition, when relationships between X and y are complex and cannot be represented with the single component-based model even with variable selection, the prediction accuracy will remain low. These are some challenges that need to be addressed in the future.

It is expected that the proposed methods will facilitate the elucidation of mechanisms in molecular design, material design, process design, and process control by constructing models with high predictive ability and interpreting the models.

5. DATA AND SOFTWARE AVAILABILITY

The data that support the findings of this study are available in refs 22−24 and 27.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c07379.

Actual y vs predicted y data (Figures S1−S8) (PDF)

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; Phone: +81-44-934-7197; Email: hkaneko@meiji.ac.jp

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c07379

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI) (Grant Nos. 19K15352, 20H02553, and 20H04538) from the Japan Society for the Promotion of Science and the Center of Innovation (COI) program (Grant No. R1WD12) from the Japan Science and Technology Agency.

REFERENCES

(1) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst* **2001**, *58*, 109−130.

(2) Li, Z. T.; Sillanpaa, M. J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor. Appl. Genet.* **2012**, *125*, 419−435.

(3) Bishop, C. M. *Pattern recognition and machine learning*; Springer: New York, 2006.

(4) Burnaev, E.; Panov, M. Adaptive design of experiments based on Gaussian processes. *Lect. Notes Comput. Sci.* **2015**, *9047*, 116−125.

(5) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model* **2007**, *47*, 219−227.

(6) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model* **2007**, *47*, 150−158.

(7) Natekin, A. K. Gradient boosting machines, a tutorial. *Front Neurobot* **2013**, *7*, 1−21.

(8) Chen, C. G. XGBoost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, DOI: 10.1145/2939672.2939785.

(9) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. Y. LightGBM: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems NIPS***2017**, 3149−3157.

(10) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38* (16), 1291−1307.

(11) Shimizu, N.; Kaneko, H. Constructing regression models with high prediction accuracy and interpretability based on decision tree and random forests. *J. Comput. Chem. Jpn.* **2021**, *20*, 71−87.

(12) Ribeiro, M. T.; Singh, S.; Guestrin, C. Why should I trust you?": Explaining the predictions of any classifier *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics***2016**, 97

(13) Lundberg, S.; Lee, S. I. A unified approach to interpreting model predictions. **2017**, https://arxiv.org/abs/1705.07874.

(14) Liu, Y.; Wu, Q.; Chen, J. Active selection of informative data for sequential quality enhancement of soft sensor models with latent variables. *Ind. Eng. Chem. Res.* **2017**, *56*, 4804−4817.

(15) Deng, H.; Yang, K.; Liu, Y.; Zhang, S.; Yao, Y. Actively exploring informative data for smart modeling of industrial multiphase flow processes. *IEEE T. Ind. Inf* **2021**, *17*, 8357−8366.

(16) Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom* **2002**, *16*, 119−128.

(17) Lestander, T. A.; Rhén, C. Multivariate NIR spectroscopy models for moisture, ash and calorific content in biofuels using bi-orthogonal partial least squares regression. *Analyst* **2005**, *130*, 1182−1189.

(18) Rosipal, R.; Trejo, L. J. Kernel partial least squares regression in Reproducing Kernel Hilbert Space. *JMLR* **2001**, *2*, 97−123.

(19) Chun, H.; Keleş, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B. Stat. Methodol* **2010**, *72*, 3−25.

(20) Zhu, G.; Su, Z. Envelope-based sparse partial least squares. *Ann. Statist* **2020**, *48*, 161−182.

(21) Katoch, S.; Chauhan, S. S.; Kumar, V. A review on genetic algorithm: past, present, and future. *Multimed. Tools Appl.* **2021**, *80*, 8091−8126.

(22) Arakawa, M.; Yamashita, Y.; Funatsu, K. Genetic algorithm-based wavelength selection method for spectral calibration. *J. Chemom* **2011**, *25*, 10−19.

(23) Kaneko, H.; Funatsu, K. A new process variable and dynamics selection method based on a genetic algorithm-based wavelength selection method. *AIChE J.* **2012**, *58*, 1829−1840.

(24) https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html (accessed 2021-09-11).

(25) https://deap.readthedocs.io/en/master/ (accessed 2021-09-11).

(26) https://github.com/hkaneko1985/dcekit (accessed 2021-09-11).

(27) Hall, L. H.; Story, C. T. Boiling point and critical temperature of a heterogeneous data set: qsar with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004−1014.

(28) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.

(29) http://www.cadaster.eu/node/65.html (accessed 2021-09-11).

(30) Wehrens, R. *Chemometrics with R - Multivariate Data Analysis in the Natural Sciences and Life Sciences*; Springer: Heidelberg, 2011.

(31) Dyrby, M.; Engelsen, S. B.; Nørgaard, L.; Bruhn, M.; Nielsen, L. L. Chemometric quantitation of the active substance in a pharmaceutical tablet using near infrared (NIR) transmittance and NIR FT raman spectra. *Appl. Spectrosc.* **2002**, *56*, 579−585.

(32) Fortuna, L.; Graziani, S.; Rizzo, A.; Xibilia, M. G. *Soft sensors for monitoring and control of industrial processes*; Springer-Verlag: London, 2007.

(33) http://www.rdkit.org/ (accessed 2021-09-11).