JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

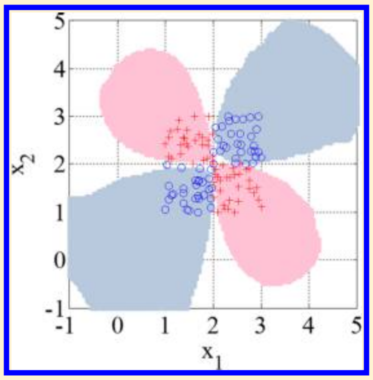
pubs.acs.org/jcim

Applicability Domain Based on Ensemble Learning in Classification and Regression Analyses

Hiromasa Kaneko and Kimito Funatsu*

Department of Chemical Systems Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT: We discuss applicability domains (ADs) based on ensemble learning in classification and regression analyses. In regression analysis, the AD can be appropriately set, although attention needs to be paid to the bias of the predicted values. However, because the AD set in classification analysis is too wide, we propose an AD based on ensemble learning and data density. First, we set a threshold for data density below which the prediction result of new data is not reliable. Then, only for new data with a data density higher than the threshold, we consider the reliability of the prediction result based on ensemble learning. By analyzing data from numerical simulations and quantitative structural relationships, we validate our discussion of ADs in classification and regression analyses and confirm that appropriate ADs can be set using the proposed method.



INTRODUCTION

Chemoinformatics[1] aims to solve chemistry problems using an informatics method. There is a great deal of previous research in this field, revolving around topics such as quantitative structure–activity relationships (QSARs), quantitative structure–property relationships (QSPRs), reaction design, and drug design.

Multivariate techniques such as the *k*-nearest neighbor algorithm (*k*-NN),[2] random forest (RF),[3] support vector machine (SVM),[4] partial least-squares (PLS),[5] and support vector regression (SVR)[6] are powerful tools for handling several problems in chemoinformatics. These multivariate techniques can be divided into classification methods and regression analysis methods. Several studies on variable selection,[7,8] sample selection,[9,10] and ensemble learning,[11,12] among others, have been carried out to improve the results of classification and regression analyses.

A constructed classification model (classifier) or regression model is used to predict class values or values of some activity or property of new data (e.g., new chemical structures); however, the reliability of the predicted values depends on each datum (e.g., each structure). In other words, accuracy in a classification analysis and error ranges in a regression analysis differ for each datum or each structure. When new data are similar to the data used to construct the classifier or regression model, the reliability of the predicted values of the new data is high, which means that the accuracy is high in a classification analysis and the error ranges are small in a regression analysis. Conversely, when new data of explanatory variables (**X**) differ greatly from the data used to construct the model, reliability of the predicted values of the new data is low, which means that the accuracy is low in a classification analysis and the error ranges are large in a regression analysis. Thus, the applicability domain (AD)[8,13,14] of a model must be taken into account in the prediction of new data.

In general, methods for setting the AD using training data are based on range,[15] similarity,[13,16,17] data density,[18,19] and ensemble learning.[8,17,20] These AD parameters can be used in combination. Sheridan combined similarity, ensemble learning, and the prediction itself and proposed a new method for defining the AD in a regression analysis.[21] Then he gave in-depth consideration to the AD using RF and concluded that ensemble learning and the prediction itself were sufficient for setting the AD in a regression analysis.[22] Both the local AD defined by molecular descriptors specializing in training data and the universal AD defined by diverse descriptors are important.[18] ADs have been applied to process control[14,15,19,20] as well as to QSARs and QSPRs.

Previously, the "distance to model" (DM)[8,19] was known as the index of the reliability of predicted values. However, DM is a slightly confusing term because the "model" is not always a point, line, or hyperplane in coordinate space and it is difficult to define "distance". Thus, we use an index for monitoring prediction reliability (IMPR), or more simply, prediction reliability (PR).[23] PR values can be used to give the estimated accuracy in a classification analysis and the estimated prediction errors in a regression analysis.

PR values can be calculated on the basis of ensemble learning. In a regression analysis, multiple subregression models can be constructed using the training data by changing the training samples, descriptors or explanatory variables (**X**), and hyperparameters in a regression model. Then the standard deviation (STD) of the multiple values predicted by the submodels is used as the PR only when there is no bias with

Received: April 15, 2014
Published: August 13, 2014

ACS Publications © 2014 American Chemical Society

respect to the experimental value, for example. If new data of **X** differ greatly from the training data, it is possible that the predicted **y** value is far from the normal range for each model (the predicted **y** value also has the possibility to be within the normal range, of course), and thus, the predicted values in the submodels can differ. Then the distribution of the predicted values is large, which means that the STD value is high. The point is that prediction errors can be small even when STD values are large. This is the case because a predicted value follows the distribution (ideally the normal distribution) where the mean is the final predicted value and the variance is the square of the STD value. However, the prediction errors can be large since the distribution is wide. As the distribution of predicted values becomes wider, the prediction results are less reliable, and the error range becomes larger. In addition, Sheridan gave in-depth consideration to the AD based on ensemble learning in a regression analysis.[21,22]

In a classification analysis, the ratio of the classification results of multiple subclassifiers is one of the PRs, and for classification methods that can give a quantitative value as a classification result, indices such as STD and STD-PROB (see ref 17) are used as PRs. Several classification methods are combined in ensemble learning, which can be used to construct consensus models. However, in a binary classification, for example, when the new data of **X** differ greatly from the training data, the ratio of the classification results does not necessarily converge to 50%, which is the minimum value. This is the case because even when new data of **X** differ greatly from the training data and are far from the boundary hyperplanes of the subclassifiers, the classification results of the subclassifiers can be the same, which means that the final classification result has 100% reliability. However, the actual regions of class 1 and those of class −1 can be complicated in a space defined by **X** variables. Even when the data are on the same side of the boundary hyperplane as the training data of class 1, these data do not necessarily belong to the same class, especially in the case that the data exist far from the region of the training data. Thus, ensemble learning methods do not work consistently well as PRs.

Therefore, in this study we first discuss ADs based on ensemble learning in a regression analysis and a classification analysis. In a regression analysis, the AD (i.e., the error range of a predicted value) can be appropriately set using ensemble learning methods, although attention must be paid to the bias of the predicted values. However, because the set AD can be too wide in a classification analysis, we propose an AD based on ensemble learning and data density for classification analysis. If new data are present in the domain where the density of the training data is high and most of the subclassifiers support the same class, the classification result can be reliable.

First, we use numerical simulation data to show that the AD based on ensemble learning is too wide in a classification analysis. Then we conduct regression and classification analyses using ensemble learning methods with QSPR and QSAR data. In this study, we used PLS and SVR as regression analysis methods and *k*-NN, RF, and SVM as classification analysis methods.

■ METHOD

In this section, we explain how to set ADs based on ensemble learning in regression and classification analyses, an AD based on data density, and the proposed AD combining ensemble learning and data density. We specifically deal with bootstrap aggregating (bagging) as an ensemble learning method, but the essence of setting the AD does not differ when another ensemble method such as boosting is used.

AD Based on Ensemble Learning in a Regression Analysis.

In ensemble approaches for regression, multiple subregression models are constructed from the training data. The submodels differ with respect to the combination of **X** variables, samples, and/or hyperparameters used in their construction. In this paper, combinations of **X** variables are changed without replacement and subtraining samples are changed with replacement in ensemble learning. These methods are called variable jackknife aggregating (jagging) and sample bagging, respectively.

We define a training data set as $\mathbf{X} \in R^{n\times m}$ and an objective variable as $\mathbf{y} \in R^{n\times 1}$, where n is the number of data values and m is the number of **X** variables. When subtraining data sets are prepared with varying random combinations of **X** variables (i.e., in variable jagging), a data set for model construction is represented as $\mathbf{X}_i \in R^{n\times p}$ and $\mathbf{y} \in R^{n\times 1}$, where i (ranging from 1 to k) is the number of the subregression model, p is the number of **X** variables in the subtraining data set, and k is the number of subtraining data sets. A subregression model is constructed for $\mathbf{X}_i$ and **y** as follows:

$$\mathbf{y} = f_i(\mathbf{X}_i) + \mathbf{e}_i \tag{1}$$

where f_i is the ith subregression model and the $\mathbf{e}_i \in R^{n\times 1}$ are the **y** residuals. For new data $\mathbf{x}_{\mathrm{new},i}$, a y value is predicted using f_i as follows:

$$y_i = f_i(\mathbf{x}_{\mathrm{new},i}) \tag{2}$$

where y_i is the ith predicted y value. The final predicted y value is the average or median of the multiple y values ($y_1, y_2, ..., y_k$) predicted by the submodels. In addition, the STD of the multiple y values predicted by the submodels is given as

$$\mathrm{STD} = \sqrt{\frac{\sum_{i=1}^{k}(y_i - \bar{y})^2}{k-1}} \tag{3}$$

This index is a PR and is used to set the AD. If the predicted **y** values are close together and the STD is small, the prediction error is assumed to be small, that is, the actual difference between the average prediction value and the experimental value will be small if there is no bias. Conversely, when the predicted **y** values vary greatly and the STD is large, the prediction error is assumed to be large. Thus, the STD can be used as an index of prediction errors. For example, 3 times the STD value can be an error range of a predicted **y** value as the AD.[14] When a molecule is outside the AD, the STD value is high, and accordingly, the error range is also high.

Meanwhile, if subtraining data sets are prepared with varying random combinations of samples used in model construction (i.e., in sample bagging), a data set for model construction is represented as $\mathbf{X}_j \in R^{q\times m}$ and $\mathbf{y}_j \in R^{q\times 1}$, where j (ranging from 1 to k) is the number of the subregression model, q is the size of the data set for model construction, and k is the number of subtraining data sets. A subregression model is constructed using $\mathbf{X}_j$ and $\mathbf{y}_j$ as follows:

$$\mathbf{y}_j = f_j(\mathbf{X}_j) + \mathbf{e}_j \tag{4}$$

where f_j is the jth subregression model and the $\mathbf{e}_j \in R^{q\times 1}$ are the **y** residuals. For new data $\mathbf{x}_{\mathrm{new},j}$, a y value is predicted using f_j as follows:

$$y_j = f_j(\mathbf{x}_{\text{new},j}) \tag{5}$$

where y_j is the jth predicted y value. Equations 4 and 5 for sample bagging are analogous to eqs 1 and 2, respectively, for variable jagging. As is the case with variable jagging, the final predicted y value is the average or median of the multiple y values predicted by the submodels, and the STD is calculated as in eq 3. Larger STD values mean lower PR values of their predicted y values. A practical error range for a predicted y value is set as c times the STD, where c is a constant value that is optimized using training data or validation data.

AD Based on Ensemble Learning for Classification Analysis. In a classification analysis, $\mathbf{y}$ is a label variable taking only integer numbers. In this study, a binary classification is considered, and $\mathbf{y}$ can be set only to 1 (class 1) or −1 (class −1).

Ensemble learning in a classification analysis basically differs little from that in a regression analysis. The subtraining data set selection of $\mathbf{X}_i$ for variable jagging and that of $\mathbf{X}_j$ and $\mathbf{y}_j$ for sample bagging are the same as those in a regression analysis. However, in a classification analysis, classification models (classifiers) are constructed instead of regression models as follows:

$$\mathbf{y} = g_i(\mathbf{X}_i) + \mathbf{f}_i \tag{6}$$

$$\mathbf{y}_j = g_j(\mathbf{X}_j) + \mathbf{f}_j \tag{7}$$

where g_i and g_j are the ith and jth subclassifiers and $\mathbf{f}_i \in R^{n\times 1}$ and $\mathbf{f}_j \in R^{k\times 1}$ are the classification errors, respectively. The classification errors $\mathbf{f}_i$ and $\mathbf{f}_j$ have values of 2, 0, or −2. For new data $\mathbf{x}_{\text{new},i}$ and $\mathbf{x}_{\text{new},j}$, y values (i.e., class values) are predicted using g_i and g_j, respectively, as follows:

$$y_i = g_i(\mathbf{x}_{\text{new},i}) \tag{8}$$

$$y_j = g_j(\mathbf{x}_{\text{new},j}) \tag{9}$$

For sample bagging and variable jagging, the k classification results can be obtained using k classifiers. The final predicted y value is set according to the majority of 1's or −1's. In addition, the ratios of 1's and −1's are given as

$$\text{RT}_1 = \frac{N_{\text{subclassifiers}(1)}}{k} \tag{10}$$

$$\text{RT}_{-1} = \frac{N_{\text{subclassifiers}(-1)}}{k} \tag{11}$$

where $N_{\text{subclassifiers}(1)}$ is the number of subclassifiers whose prediction result is 1 and $N_{\text{subclassifiers}(-1)}$ is the number of subclassifiers whose prediction result is −1. RT_1 can be used as the PR for data whose final predicted y value is 1, while RT_{-1} can be used as the PR for data whose final predicted value is −1. Larger RT_1 or RT_{-1} values, which mean that a greater number of subclassifiers support the prediction results, indicate higher PR values. The relationship between RT_1 and RT_{-1} is given as

$$\text{RT}_{-1} = \frac{N_{\text{submodels}(-1)}}{k} = \frac{k - N_{\text{submodels}(1)}}{k} = 1 - \text{RT}_1 \tag{12}$$

If y_i and y_j in eqs 8 and 9 are quantitative values, the final predicted y value is 1 in the case that the average or median of multiple y values is greater than 0 and −1 in the case that the average or median of multiple y values is less than 0. In addition, the STD (see eq 3) and STD-PROB values, among others, can be used as PRs.[17] When the predicted y values are close together and the STD value is low, the prediction accuracy will be high. Conversely, when the predicted y values vary widely and the STD value is high, the prediction accuracy will be low.

AD Based on Data Density. Data density is used as a PR. If the number of training data neighboring new data $\mathbf{x}_{\text{new}}$ is large and the data density is high, the PR value is also high, and the prediction result for $\mathbf{x}_{\text{new}}$ is as reliable as the result for the training data. Conversely, if the number of training data neighboring $\mathbf{x}_{\text{new}}$ is small and the data density is low, the PR value is also low, and the prediction result for $\mathbf{x}_{\text{new}}$ is unreliable. For example, a one-class SVM and k-NN were employed to calculate data density.[19]

AD Based on Ensemble Learning and Data Density. The reliability of the prediction result for $\mathbf{x}_{\text{new}}$ will be high if both the PR value based on ensemble learning and that based on data density are high. We therefore propose an AD based on both ensemble learning and data density. First, the threshold is set for data density, and if the data density for $\mathbf{x}_{\text{new}}$ is lower than the threshold, the prediction result is not reliable. Then, only for those $\mathbf{x}_{\text{new}}$ whose data density is greater than the threshold, the reliability of the prediction result is considered using the PR based on ensemble learning.

■ RESULTS AND DISCUSSION

First, we confirm that the domains of the ADs based on ensemble learning in a classification analysis are too large. Then we consider ADs for both regression and classification analyses using QSAR and QSPR data. In this study, we use the PLS and SVR methods for regression analyses and the k-NN, RF, and SVM methods for classification analyses. For the ensemble learning, the number of $\mathbf{X}$ variables in each submodel, p, is the square root of the number of $\mathbf{X}$ variables and the number of samples in each submodel, q, is the number of training data, and we accept overlapping in sample bagging. The numbers of submodels in the regression and classification analyses are 300 and 301, respectively.

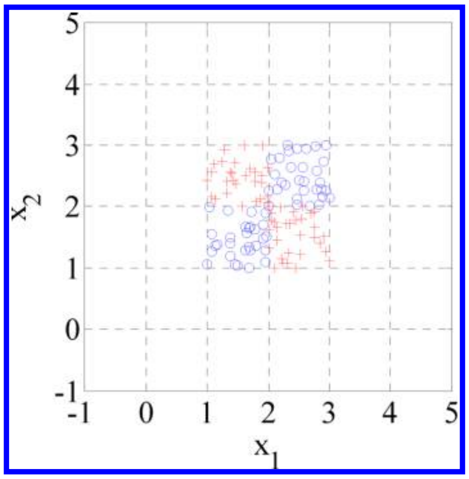
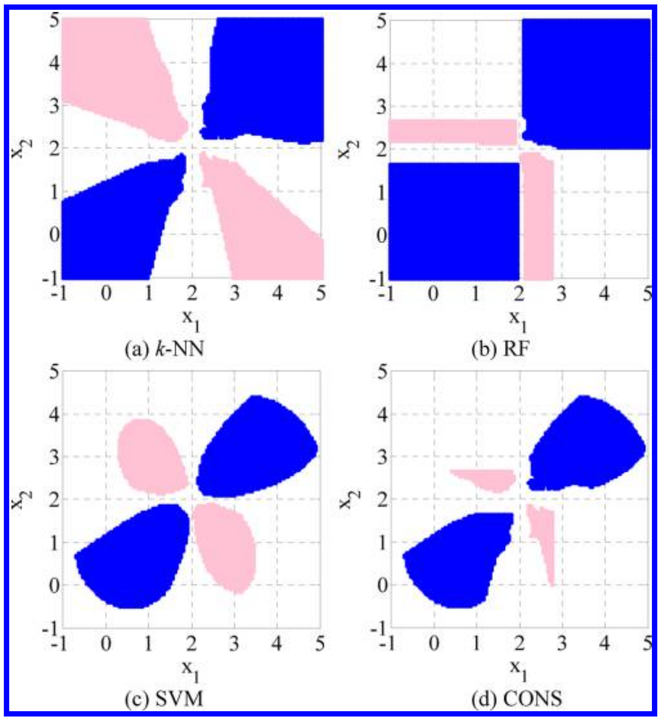
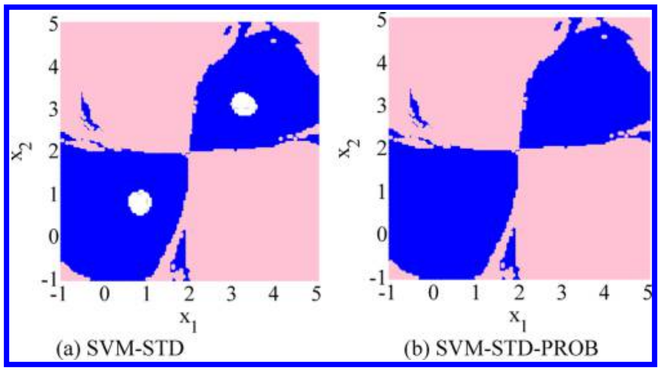


Figure 1. Plot of $\mathbf{x}_1$ vs $\mathbf{x}_2$ for the numerical simulation data. Circles denote the training data of class 1, while crosses denote the training data of class −1.



(a) *k*-NN (b) RF (c) SVM (d) CONS

Figure 2. ADs defined by each classification method. The blue domains are the ADs for class 1, and the pink domains are those for class −1.



(a) SVM-STD (b) SVM-STD-PROB

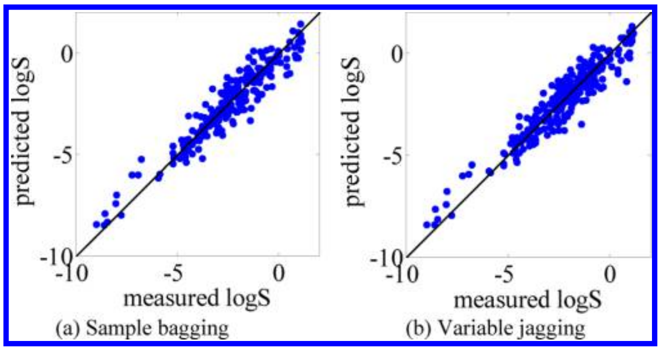
Figure 3. ADs defined by each continuous classification method. The blue domains are the ADs for class 1, and the pink domains are those for class −1.

Table 1. Training Data, Test Set 1, and Test Set 2 for QSPR Analysis

	training data	test data
compounds consisting only of H, C, N, and O	567	257 (test set 1)
other compounds	0	466 (test set 2)

Table 2. RMSE Values for Test Set 1 and Test Set 2 with the QSPR Data

	test set 1	test set 2
PLS (sample)	0.650	1.05
PLS (variable)	0.653	1.12
SVR (sample)	0.586	1.22
SVR (variable)	0.571	1.28



(a) Sample bagging (b) Variable jagging

Figure 4. Measured and predicted logS values for test set 1 using SVR modeling with the QSPR data.

The specifications of the computer used in this study are given below: OS, Windows 7 Professional (64 bit); CPU, Intel Xeon X5690 3.47 GHz; RAM, 48.0 GB. The R2013a version of MATLAB was used.

AD in a Classification Analysis Using Numerical Simulation Data. The data set was randomly generated as shown in Figure 1. Circles and crosses denote training data for class 1 and class −1, respectively. The number of training data in each class is 60. Only sample bagging was performed because there are two **X** variables, $\mathbf{x}_1$ and $\mathbf{x}_2$. Test data were generated on a grid from −1 to 5 for $\mathbf{x}_1$ and $\mathbf{x}_2$. The ranges of $\mathbf{x}_1$ and $\mathbf{x}_2$ extend beyond those of the training data to investigate whether the defined ADs are within the training data domains.

The domains for $RT_1 = 1$ and $RT_{-1} = 1$ are depicted in Figure 2. The colored domains mean that all 301 subclassifiers yielded the same classification results and that the prediction results for the *k*-NN, RF, and SVM models are reliable. The blue domains indicate $RT_1 = 1$ and the pink domains $RT_{-1} = 1$, and these are the ADs. In each plot, although the regions close to the boundaries of class 1 and class −1 are outside the ADs, the ADs are too large compared with the training data shown in Figure 1. The results of the consensus model (CONS) combined with the *k*-NN, RF, and SVM models are shown in Figure 2d. The number of subclassifiers is 903 (301 × 3). Even the consensus model that yielded good results in ref 17 resulted in a too-large AD according to Figures 1 and 2d.

Figure 3 shows the results of the ADs when the output of the SVM is quantitative values. The STD-PROB value can be calculated by assuming a normal distribution with mean equal to a predicted value and variance equal to the square of an STD value and integrating from 0 to $+\infty$ for a positive predicted value or from $-\infty$ to 0 for a negative value. For the details of STD-PROB, the reader is referred to ref 17. The STD values for the training set are computed using the predicted values after 5-fold cross-validation. The minimum value of STD and the maximum value of STD-PROB in the training data were set as the thresholds for each PR, that is, the data domain where the STD values were lower than the threshold was the AD for SVM−STD and the data domain where the STD-PROB values were higher than the threshold was the AD for SVM−STD-PROB. These domains are shown in Figure 3a,b, respectively. The ADs are too large compared with the data domain of the training data shown in Figure 1.

In this situation, the data domain of the training data could be represented using data density.

QSPR Study of Aqueous Solubility. We analyzed the aqueous solubility data reported by Hou et al.[24] Aqueous solubility is expressed as logS, where S is the solubility (in

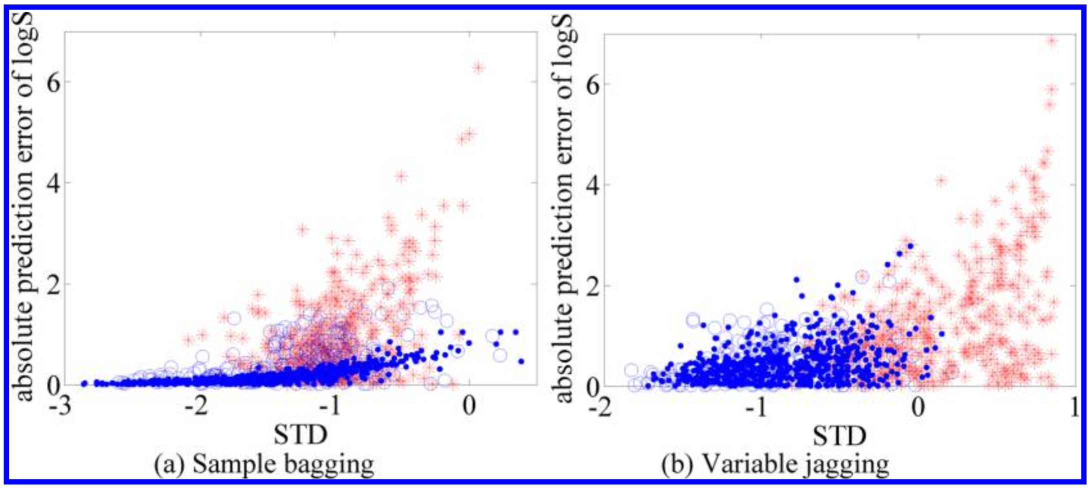


Figure 5. Relationship between STD and absolute prediction error of logS using SVR modeling with the QSPR data. The *x* axes are logarithmic axes. Blue points denote training data; blue circles denote test set 1 data; and red asterisks denote test set 2 data.

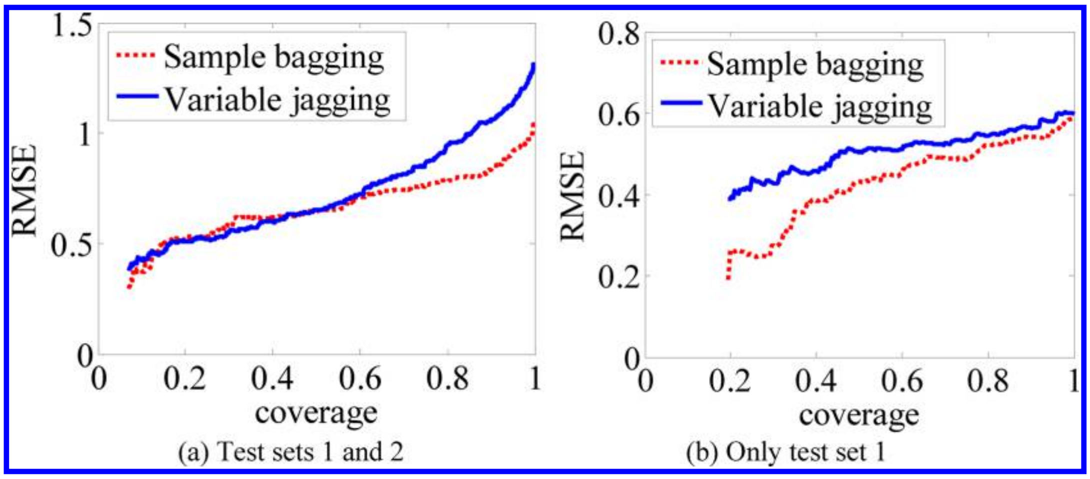


Figure 6. Relationships between the coverage and the RMSE for the QSPR data.

moles per liter) at a temperature of 20−25 °C. The Hou data set includes 1290 diverse compounds and has been analyzed by several groups.[25−31]

Some 2232 molecular descriptors were calculated on this data set using the Dragon 6.0 software.[32] Descriptors with the same values for 100% of the molecules were removed, leaving 1691 descriptors.

The method for dividing the training and test data sets is given in Table 1. The training data were the compounds consisting only of H, C, N, and O in the training data in ref 24. The other training data in ref 24 were included in test set 1. Test set 2 included the test data in ref 24. The training data set contained a total of 567 molecules, and the remaining 723 molecules (257 + 466) were included in the two test data sets. In the classification analysis, the threshold of logS was set so that the numbers of training data for class 1 and class −1 were the same. Compounds with logS values greater than the threshold belonged to class 1, and those with logS values lower than the threshold belonged to class −1. To compare the ADs for the regression and classification analyses using ensemble learning, the training data and the data for test sets 1 and 2 were the same in the regression and classification analyses.

The results of the regression analysis are discussed first. Table 2 shows the root-mean-square error (RMSE) values of the test data sets for each regression analysis method and each ensemble learning method. The lower the RMSE value, the greater is the predictive accuracy. The molecules in test set 1 consisted of the same atom types as those in the training data, and accordingly, each molecule in test set 1 has a possibility of being within the AD of a regression model constructed using the training data. However, the AD is a complex hyperspace related to the regression method, ensemble method, and structural similarity between molecules, and thus, further investigations of the AD were required and are shown later. Because the RMSE values for SVR are smaller than those for PLS, the predictive models were constructed using the SVR method. The nonlinear relationship between **X** and **y** could be handled by the SVR method.

The RMSE values of test set 2 are very high compared with those of test set 1. The reason for this is that the molecules in test set 2 consist of atoms that are not included in the molecules in the training data and the molecules in test set 2 are basically outside the AD, considering the universal AD.[18] The large RMSE values are thus reasonable. However, as mentioned above, the AD is a complex hyperspace, and further investigations are shown later. The results for sample bagging and those for variable jagging did not differ greatly in this case study.

Figure 4 shows the relationships between the measured and predicted logS values for test set 1 using SVR modeling. The

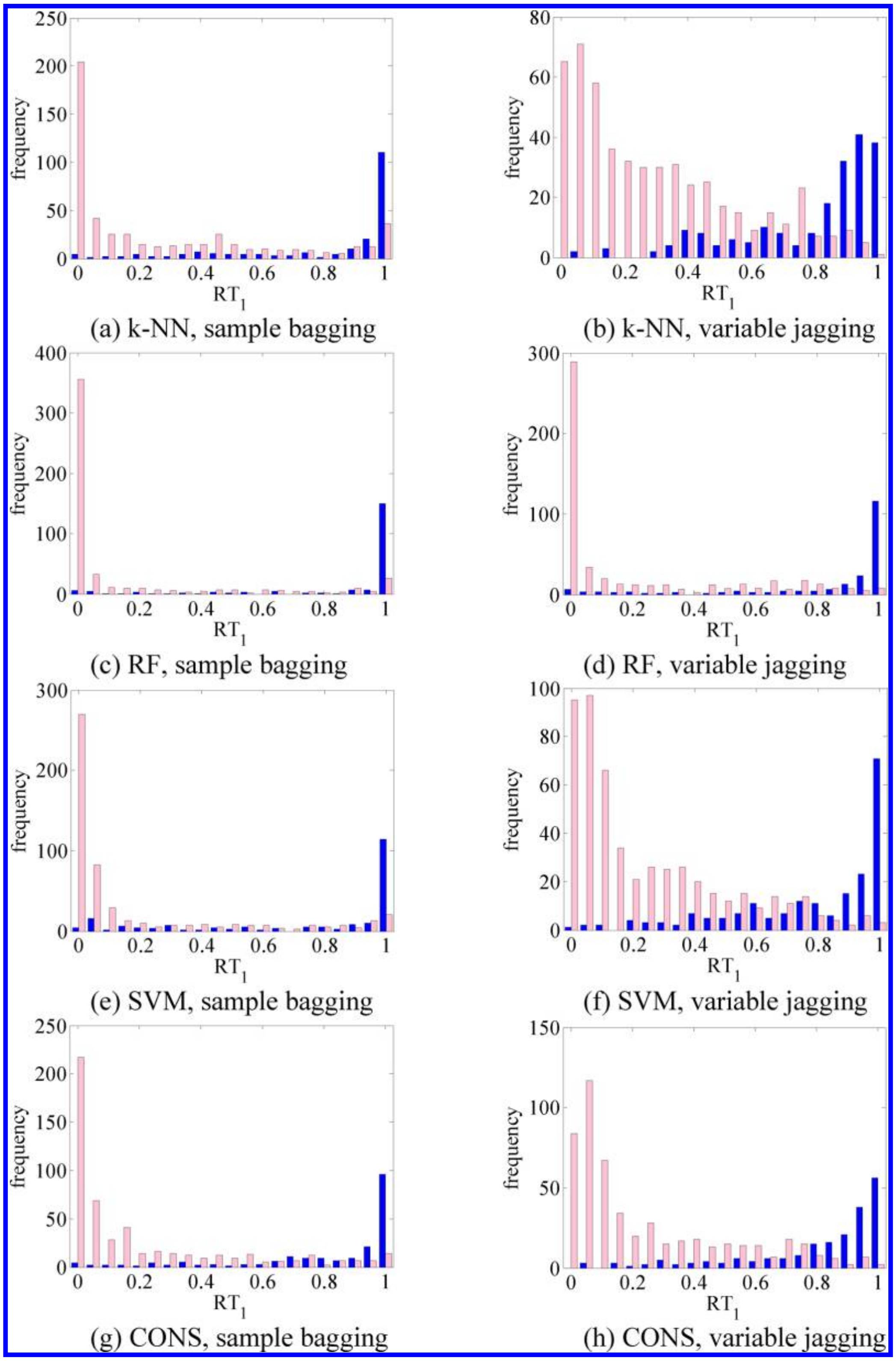


Figure 7. Histograms of RT$_1$ with the QSPR data. Blue bars indicate frequencies of test data of class 1, and pink bars indicate frequencies of test data of class −1.

SVR models based on sample bagging and variable jagging can both accurately predict the molecules of test set 1, most of which would be within the AD.

Figure 5 shows plots of STD and the absolute errors of logS using SVR modeling for sample bagging and variable jagging. Blue points denote training data, blue circles represent test set 1, and red asterisks represent test set 2. The *x* axes are logarithmic axes. According to the results for both sample bagging and variable jagging, as the STD value increases, so too does the prediction error and the variance in the errors, which implies a typical tendency for PRs. Although a wide range of absolute errors can be seen for a given STD, the predicted **y**

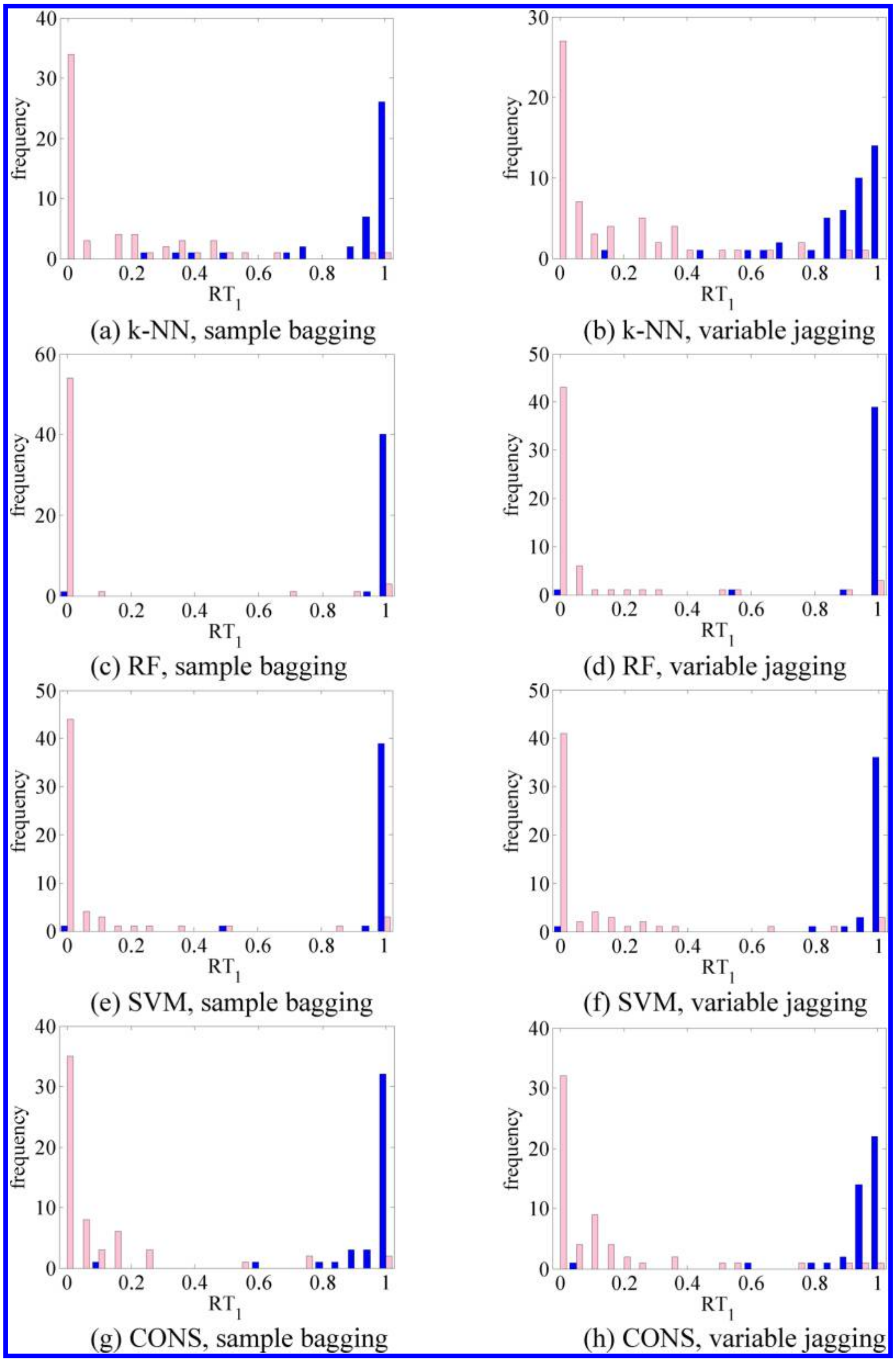


Figure 8. Histograms of RT_1 only for the test data within the AD based on data density with the QSPR data. Blue bars indicate frequencies of test data of class 1, and pink bars indicate frequencies of test data of class −1.

values have their distribution. The STD represents the variation of the distribution. Therefore, even when an STD value is high, prediction errors may be low by chance because the STD does not represent the amount of prediction errors but instead represents the variation of the distribution of predicted values. (If the STD represented the amount of prediction errors, the predicted values would be corrected using the STD values.) The point is that the prediction errors can be high when the STD values are high. In other words, the possibility of high prediction errors increases according to the increase of the STD values. However, in the case of sample bagging, the plots of the training data, test set 1, and test set 2 overlap completely and

Table 3. Training Data, Test Set 1, and Test Set 2 in QSAR Analysis

	training data	test data
compounds consisting only of H, C, and O	300	186 (test set 1)
other compounds	0	607 (test set 2)

Table 4. RMSE Values for Test Set 1 and Test Set 2 with the QSAR Data

	test set 1	test set 2
PLS (sample)	0.359	0.839
PLS (variable)	0.374	0.833
SVR (sample)	0.344	0.820
SVR (variable)	0.368	0.821

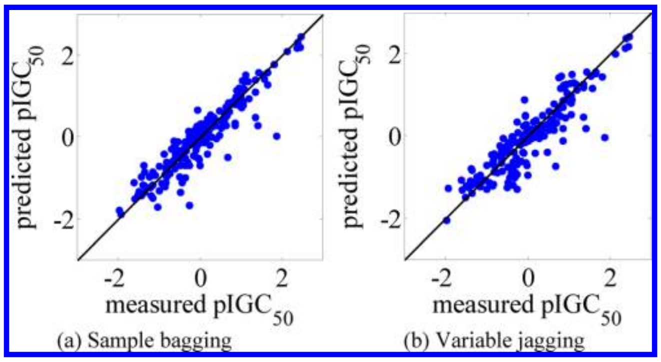
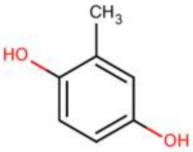
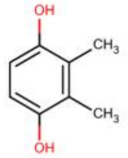
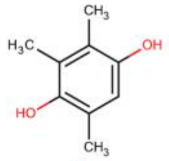
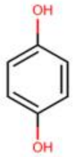
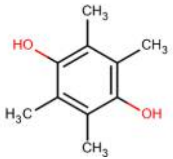


Figure 9. Measured and predicted pIGC$_{50}$ values for data in test set 1 using SVR modeling with the QSAR data.

Table 5. Measured and Predicted pIGC$_{50}$ Values for Selected Molecules

Data set	Structure	Measured pIGC$_{50}$	Predicted pIGC$_{50}$
Test set 1		1.86	-0.0316
Test set 1		1.41	0.20
Test set 1		1.34	0.47
Test set 1		0.47	-0.40
Training data		1.28	0.57

the STD values cannot distinguish the training data from test set 2 or test set 1 from test set 2. Although most of the data in test set 2 are outside the AD because the molecules in test set 2 consisted of different atom types than those in the training data, the STD of sample bagging shows that the data for test set 2 are within the AD and that the prediction errors are small. However, the actual prediction errors are large. The PR based on sample bagging does not work because similar subregression models were constructed and the diversity of the submodels was too low to evaluate the diverse chemical structures. The results also indicate that the STD for sample bagging is not appropriate as a PR when structures that are more diverse than those of training data are used as input.

Meanwhile, the STD values for variable jagging are able to distinguish the training data and data for test set 2 with large prediction errors and the data for test set 1 and test set 2 with large prediction errors. Test set 2 with the same STD values as the training data actually shows the smallest prediction errors, as is the case for the training data and test set 1, which indicates that these test data are within the AD. Test set 2 with large STD values actually has the possibility to have large prediction

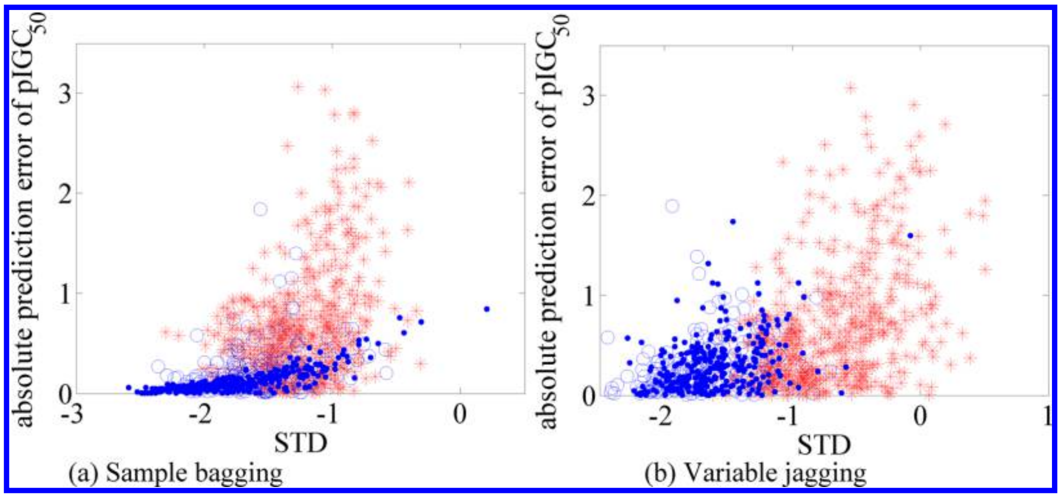
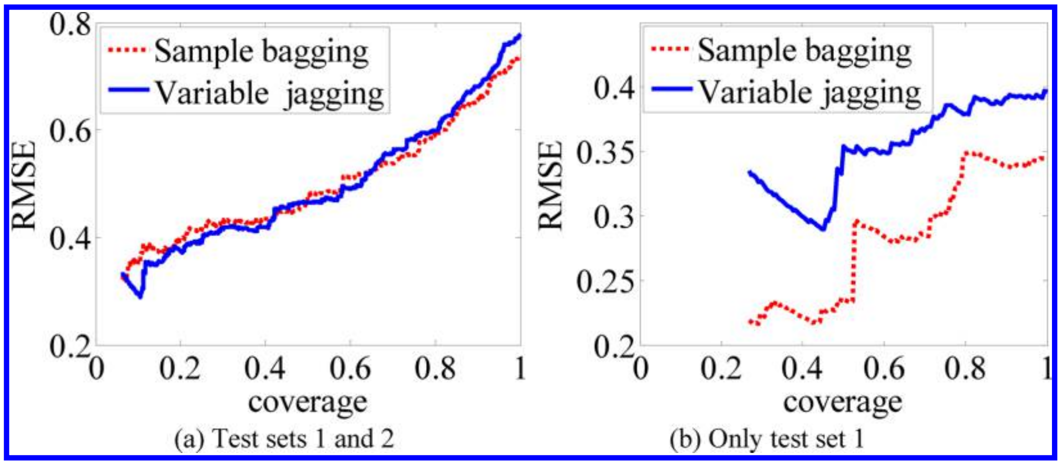


Figure 10. Relationships between STD and absolute prediction error of pIGC$_{50}$ using SVR modeling with the QSAR data. The *x* axes are logarithmic axes. Blue points denote training data; blue circles denote test set 1 data; and red asterisks denote test set 2 data.



(a) Test sets 1 and 2

(b) Only test set 1

Figure 11. Relationships between the coverage and the RMSE for the QSAR data.

errors and a large variance in the errors. The diversity of subregression models is appropriate and a suitable AD can be set using variable jagging.

Figure 6 shows the relationships between the coverage and the RMSE. First, the test data were sorted in ascending order for each STD, and we calculated the coverage,[19] that is, the ratio of the number of data within each AD to the total number of data N_{all}. The coverage of the ith data point is defined as follows:

$$\text{coverage}_i = \frac{i}{N_{\text{all}}} \tag{13}$$

The ith RMSE value was calculated using the i data points, which have STD values that are not greater than that of the ith data point, as follows:

$$\text{RMSE}_i = \sqrt{\frac{\sum_{j=1}^{i} (y_{\text{obs},j} - y_{\text{pred},j})}{i}} \tag{14}$$

where $y_{\text{obs},j}$ is the measured $\mathbf{y}$ value and $y_{\text{pred},j}$ is the predicted $\mathbf{y}$ value for each of the i data points in the test data. It is desirable that the smaller the values of the coverage are, the smaller the RMSE values become, and vice versa. As shown in Figure 6a, the curves for both sample bagging and variable jagging are desirable for PRs, and there does not seem to be a significant difference between the results for sample bagging and variable jagging. Although the RMSE values for sample bagging are lower than those for variable jagging for coverages of 0.6 or more, the molecules with the high STD values must be outside the ADs completely, considering that 466 of 723 (=257 + 466) molecules (64.5%) consist of atoms that are not included in the molecules in the training data. The point is that an AD is defined appropriately and highly predictive ability is performed inside the AD, and thus, the prediction accuracy for the molecules outside the AD does not mean that much. However, when only the molecules in test set 1 consisting of the same atom types as those in the training data were used, the RMSE values calculated with sample bagging were smaller than those with variable jagging, reflecting the fact that a more desired AD could be set by using sample bagging. When the input molecules were similar to the molecules in the training data, the STD for sample bagging was appropriate as a PR.

Next, we discuss the results of the classification analysis. The subclassifiers were constructed using the k-NN, RF, and SVM methods for sample bagging and variable jagging. Figure 7 shows the RT_1 histograms for each method. Blue bars denote the frequencies of the test data for class 1, while pink bars denote the frequencies of the test data for class −1. Figure 7a−f shows that many molecules with RT_1 values close to 1 and 0 (i.e., RT_{-1} = 1; see eq 12) actually belong to class 1 and class −1, respectively, using k-NN, RF, and SVM for both sample bagging and variable jagging. However, for several molecules with RT_1 values close to 1 and 0 (i.e., RT_{-1} = 1), the actual classes are −1 and 1, respectively. For example, RT_1 = 1 (i.e., RT_{-1} = 0) means that the possibility of class 1 is 100% (and the possibility of class −1 is 0%), although the true class is −1. Although the ADs show that the prediction results are mostly reliable, the results are incorrect, which is an undesirable situation. As confirmed in the numerical simulation data analysis, the ADs based on ensemble learning are too large. According to Figure 7g,h, the same is true for the consensus model (CONS), which produced one of the best results in ref 17.

This tendency of the ADs is stronger for sample bagging than for variable jagging. From the results for sample bagging, the frequency of class −1 molecules with RT_1 values close to 1 is higher than that of class −1 molecules with RT_1 values relatively far from 1. Conversely, compared with variable jagging, the frequencies of correctly classified molecules with RT_1 values close to 1 and 0 are high, which is a good situation.

We then combined the AD based on ensemble learning and that based on data density. In this paper, first we reduced the dimensions of the $\mathbf{X}$ variables by principal component analysis (PCA),[33] and then, for new data, we employed the average of the distances of the five nearest neighbors in the training data as the threshold based on data density. The number of principal components was determined so that the cumulative contribution ratio was above 0.9999. The threshold of the average distance was set so that 99.7% of the training data were within the AD and new data whose average distances were below the threshold were also within the AD. About 99.7% of the values lie within three standard deviations of the mean in a normal distribution. This 99.7% was used in the 3σ rule. The average distance did not always follow a normal distribution, and therefore, the normal distribution was not assumed and the threshold was set on the basis of the training data directly. The loss in coverage due to the density filter was 86%, and most of the molecules were judged to be outside the AD, which is reasonable in this case study because most of the molecules consisted of atoms that are not included in the molecules in the

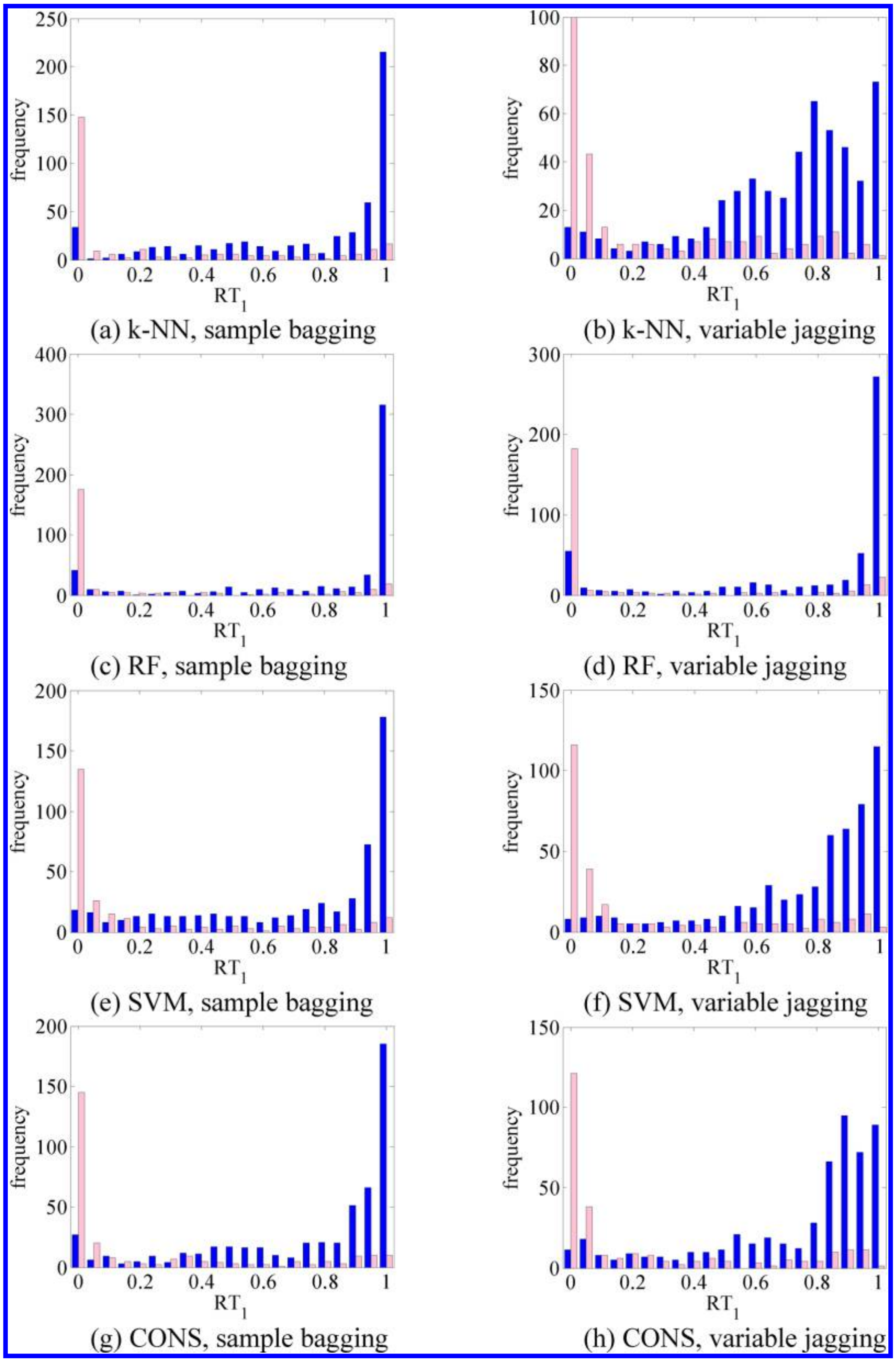
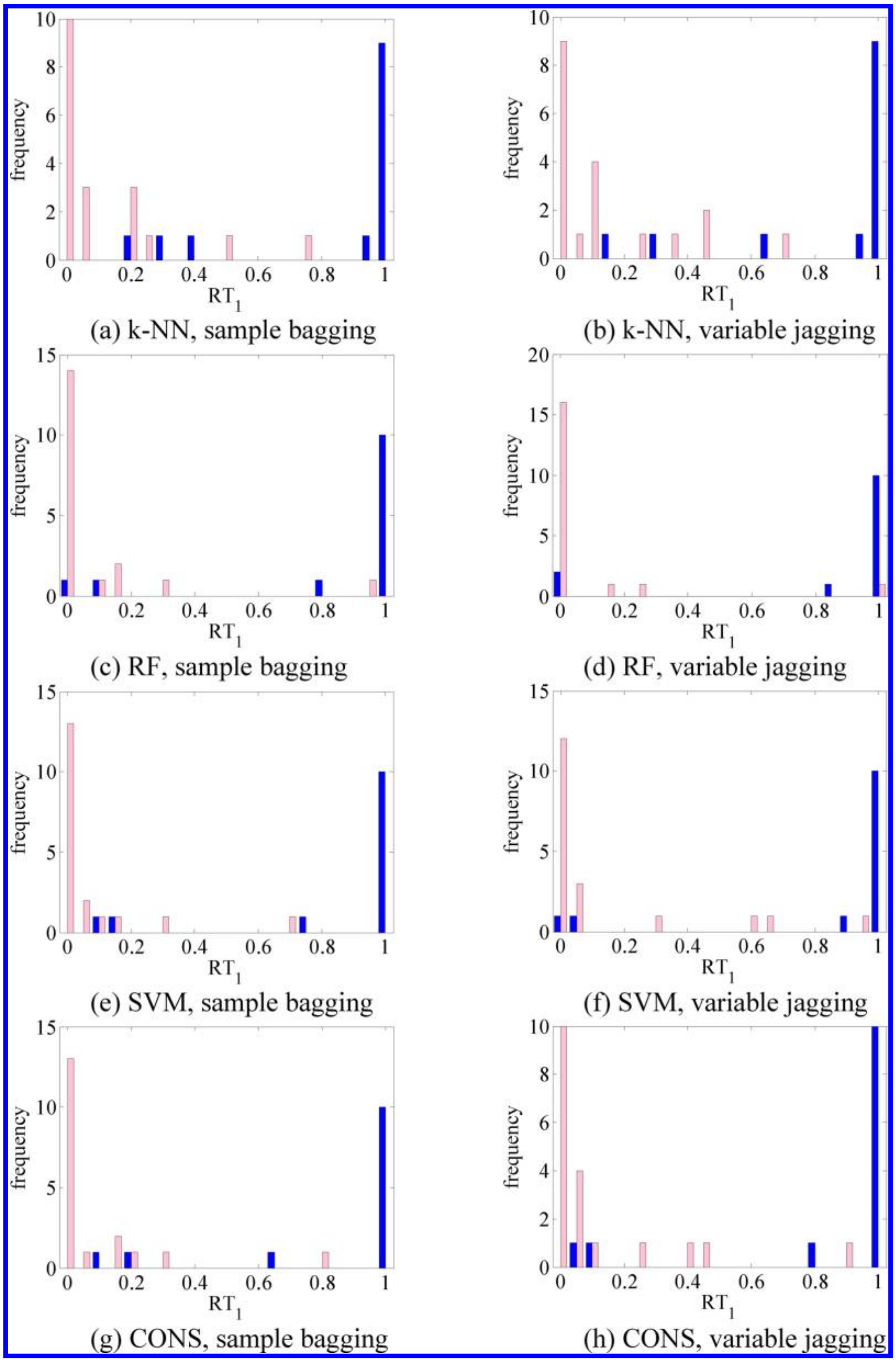


Figure 12. Histograms of RT$_1$ with the QSAR data. Blue bars indicate frequencies of test data of class 1, and pink bars indicate frequencies of test data of class −1.

training data. The RT$_1$ histograms only for molecules within the AD based on data density are shown in Figure 8. Compared with Figure 7, the numbers of class 1 and class −1 molecules with RT$_1$ values close to 1 and 0, respectively, are reduced. However, there are few class −1 molecules with RT$_1$ values close to 1 and few class 1 molecules with RT$_1$ values close to 0. Although molecules with RT$_1$ values around 0.5 are present, the AD based on ensemble learning (i.e., using sample bagging or variable jagging) is able to determine that these molecules should be outside the AD. This confirms that the AD in a classification analysis can be set appropriately by combining ensemble learning and data density. Moreover, Figure 8 shows



(a) k-NN, sample bagging

(b) k-NN, variable jagging

(c) RF, sample bagging

(d) RF, variable jagging

(e) SVM, sample bagging

(f) SVM, variable jagging

(g) CONS, sample bagging

(h) CONS, variable jagging

Figure 13. Histograms of RT$_1$ only for the test data within the AD based on data density with the QSAR data. Blue bars indicate frequencies of test data of class 1, and pink bars indicate frequencies of test data of class −1.

that when ensemble learning and data density are combined, sample bagging is more suitable than variable jagging because the numbers of the correctly classified molecules with RT$_1$ values close to 1 and 0 are higher in sample bagging compared with variable jagging. Most of the molecules were judged to be outside the AD using density-based similarity. One of the reasons is that the molecules in test set 2 consisted of atoms that are not included in the molecules in the training data and should be judged to be outside the AD, at least considering the universal AD.[18] In addition, not all of the molecules in test set 1, consisting of only atoms that are included in the molecules in the training data, are inside the AD because the AD is a complex hyperspace. The important thing is the result that only the reliable molecules are selected to be inside the AD when both ensemble learning and data density are used. Indices of data density and their thresholds are subjects of future investigation.

QSAR Study Using $pIGC_{50}$. We analyzed data downloaded from the Environmental Toxicity Prediction Challenge 2009 Web site.[34] This is an online challenge that invites researchers to predict the toxicities of molecules against *Tetrahymena pyriformis*, expressed as the logarithm of the 50% growth inhibitory concentration ($pIGC_{50}$). The data set consists of 1093 compounds and has been analyzed for the visualization of molecular fingerprints.[35]

Some 2232 molecular descriptors were calculated on this data set using the Dragon 6.0 software.[36] Descriptors with the same values for 100% of the molecules were removed, leaving 1760 descriptors.

The method used to divide the training and test data sets is shown in Table 3. The training data set comprised 300 molecules, and the remaining 793 (186 + 607) molecules were included in two test data sets. Compounds with $pIGC_{50}$ values greater than the threshold (−0.085) belonged to class 1, while the remaining compounds belonged to class −1. To compare the ADs for the regression and classification analyses using ensemble learning, the compositions of the training data set, test set 1, and test set 2 were the same in both analyses.

As the results of the regression analysis, Table 4 shows the RMSE values for the test data sets using PLS and SVR modeling for each ensemble learning method. There is no significant difference between the results for PLS and SVR. The RMSE values for test set 2 are greater than those for test set 1. This is the case because the molecules in test set 2 included atoms that were not present in the molecules of the training data and should be outside the AD, at least considering the universal AD,[18] and thus, these results are reasonable. For the data in test set 1, most of which are within the AD, the RMSE values for sample bagging are slightly lower than those for variable jagging in this case study. Figure 9 shows the relationships between the measured and predicted $pIGC_{50}$ values using test set 1 in SVR modeling. Although there seem to be some outliers from the diagonal, which are discussed in the next section, each model is able to predict the $pIGC_{50}$ values completely. This means that $pIGC_{50}$ values of molecules within the ADs can be predicted accurately.

Figure 10 shows the relationship between the STD and absolute prediction error of $pIGC_{50}$ for each bagging method when the SVR method is used. The x axes are logarithmic axes. For both sample bagging and variable jagging, the prediction errors increase and their distributions widen with increasing STD value. This is a desirable tendency for PRs. However, the STD values cannot distinguish the training data from the data for test set 2 or the data for test set 1 from those for test set 2 in the case of sample bagging, as was the case in the QSPR data analysis. Although most of the data for test set 2 are outside the AD because the constituent atoms in the data for test set 2 differed from those in the training data, the prediction errors are judged to be small by the AD based on sample bagging because similar submodels with low diversity were constructed. Of course, the actual prediction errors are large. Thus, the STD of sample bagging is not appropriate as a PR.

Meanwhile, as for the QSPR data analysis, the actual prediction errors and their variation increase with increasing STD values using variable jagging. Although there are several molecules in the data for test set 2 with the same STD values as those in the training data, these molecules had equally small prediction errors as the molecules in the training data and test set 1. Several molecules in the data for test set 2 with high STD values had large prediction errors. Thus, the diversity of the submodels is sufficient, and the appropriate AD can be set using variable jagging for test set 2, which consists of diverse molecules. As we mentioned, the predicted **y** values have their distribution, and the STD represents the variance of the distribution. Even when the STD values of molecules are high, there exist molecules whose prediction errors are low (i.e., the **y** values can be accurately predicted by chance). However, the possibility that molecules have large prediction errors increases according to the increase of the STD values.

However, in Figure 10b, there are some molecules with high prediction errors compared with the other molecules in test set 1, although their STD values are low and the molecules were estimated to be within the AD. The top molecule in Table 5 is the molecule with the largest prediction error in test set 1. This molecule was significantly underestimated. The $pIGC_{50}$ values predicted by the submodels are normally distributed with the center of the distribution at −0.0316. The other molecules in Table 5 are similar to those in test set 1 and the training data. All of these were underestimated similarly to the top molecule in Table 5. Molecules including such a skeleton have some negative bias. Although bagging can evaluate the variance of the predicted value and estimate its reliability, its bias cannot be handled. Considerable attention needs to be paid to the bias of the predicted values when using ensemble learning. Marvin View,[34] which is ChemAxon software, was used to visualize the structures.

In Figure 10b, the molecule with the largest STD value in the training data is methanol, which is the smallest molecule in the data set.

The relationships between the coverage and the RMSE are shown in Figure 11. As in the QSPR data analysis, the smaller the values of the coverage, the smaller the RMSE values become, and vice versa, which is desirable for PRs. In addition, there seems to be little difference between the results for sample bagging and variable jagging. However, the RMSE values calculated with sample bagging were smaller than those with variable jagging when only the molecules in test set 1 were used, as shown in Figure 11b. We confirmed that the STD for sample bagging is appropriate as a PR when the input molecules are similar to the molecules in the training data.

Next, we discuss the classification results. The subclassifiers were constructed using sample bagging and variable jagging. Figure 12 shows the histograms for RT_1. For the k-NN, RF, and SVM methods, there are many molecules with RT_1 values close to 1 and 0, where the actual classes were 1 and −1, respectively. However, some molecules of class 1 have RT_1 values close to 0 and some molecules of class −1 have RT_1 values close to 1, as for the QSPR analysis. $RT_1 = 0$ ($RT_{-1} = 1$) means that the possibility of class 1 is 0%, although the actual class is 1. Although the ADs based on ensemble learning show the reliability of the prediction result, it is in fact incorrect, which is not a desirable situation. The AD is too wide, as confirmed by the numerical simulation data analysis. Figure 12g,h shows that the consensus model has the same tendency.

The AD based on data density was combined with that based on ensemble learning. As was the case in the QSPR analysis, the dimensions of the **X** variables were first reduced using PCA, and then the average of the distances from five nearest-neighbor training data was used as the data-density-based AD. The settings for PCA and the threshold of the average distance were the same as those in the QSPR analysis. The loss in coverage due to the density filter was 96%, which is very high. However, this is reasonable in this case study because the rate of the molecules of test set 2 consisting of atoms that are not included in the molecules in the training data is also high. The results for the proposed methods are shown in Figure 13. Compared with Figure 12, the numbers of molecules with RT_1 values close to 1 and 0 where the actual class is 1 and −1, respectively, are reduced. However, there are few molecules with RT_1 values close to 1 and 0 where the actual class is −1 and 1, respectively. This means that the prediction results could be reliable if the AD determined them to be reliable. Although there are some molecules with RT_1 values around 0.5, the PR based on ensemble learning is able to determine that these molecules are outside the AD. As in the QSPR analysis, most of the molecules were judged to be outside the AD using density-based similarity because the molecules in test set 2 consisted of atoms that are not included in the molecules in the training data and should be judged to be outside the AD, at least considering the universal AD,[18] and not all of the molecules in test set 1, consisting of only atoms that are included in the molecules in the training data, are inside the AD because the AD is a complex hyperspace. However, only the reliable molecules are selected to be inside the AD when both ensemble learning and data density are used. Thus, we confirmed that the appropriate AD can be set by combining ensemble learning and data density. In this case study, there is little difference between the results of the combination of sample bagging and data density and those for variable jagging and data density.

CONCLUSION

In this study, we have discussed ADs based on ensemble learning for regression and classification analyses and proposed a new AD based on both ensemble learning and data density. In a regression analysis, the AD can be set and the prediction errors can be estimated appropriately using variable jagging for chemical structures that are more diverse than those of training data. The STD for sample bagging is appropriate as a PR when the input molecules are similar to the molecules in training data. However, attention must be paid to the bias of the predicted values, whereas their variance can be evaluated by bagging.

In a classification analysis, when only ensemble learning methods are used, the ADs are too wide and the classification results are incorrect even though the ADs indicate that the results are 100% reliable. In this case, we confirmed that the proposed AD based on both ensemble learning and data density works well. The numbers of molecules that are outside the AD are large potentially in our case studies. The point is that the AD based on ensemble learning in classification analysis is too wide and only the reliable molecules are selected to be inside the AD by combining ensemble learning and data density. Indices of data density and their thresholds are subjects of future investigation.

Data density can be combined with ensemble learning when the ADs are set in regression analysis. However, they were not combined, which is supported in ref 22, because the ADs become small when the ADs based on data density are considered and only the ADs based on ensemble learning work well.

In this study, we did not consider a combination of sample bagging and variable jagging, which has a possibility of improving the accuracy of the ADs. In future work, the bias of predicted values must be considered in ADs used in a regression analysis. A method exists whereby AD hyperparameters such as ν in a one-class SVM in a regression analysis are determined automatically.[19] A similar method would be desirable for use in classification analysis.

AUTHOR INFORMATION

Corresponding Author

*E-mail: funatsu@chemsys.t.u-tokyo.ac.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge the financial support of the Japan Society for the Promotion of Science (JSPS) through a Grant-in-Aid for Young Scientists (B) (Grant 24760629).

ABBREVIATIONS

QSAR, quantitative structure−activity relationship; QSPR, quantitative structure−property relationship; *k*-NN, *k*-nearest neighbor algorithm; RF, random forest; SVM, support vector machine; PLS, partial least-squares; SVR, support vector regression; AD, applicability domain; DM, distance to model; IMPR, index of monitoring prediction reliability; PR, prediction reliability; STD, standard deviation; RMSE, root-mean-square error; $pIGC_{50}$, logarithm of 50% growth inhibitory concentration.

REFERENCES

(1) Gasteiger, J.; Engel, T. *Chemoinformatics—A Textbook*; Wiley-VCH: Weinheim, Germany, 2003.

(2) Ajmani, S.; Jadhav, K.; Kulkarni, S. A. Three-dimensional QSAR Using the *k*-Nearest Neighbor Method and Its Interpretation. *J. Chem. Inf. Model.* **2006**, *46*, 24−31.

(3) 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.

(4) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1999.

(5) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(6) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.

(7) Shen, Q.; Jiang, J. H.; Tao, J. C.; Shen, G. L.; Yu, R. Q. Modified Ant Colony Optimization Algorithm for Variable Selection in QSAR Modeling: QSAR Studies of Cyclooxygenase Inhibitors. *J. Chem. Inf. Model.* **2005**, *45*, 1024−1029.

(8) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against *Tetrahymena pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.

(9) Lepp, Z.; Huang, C. F.; Okada, T. Finding Key Members in Compound Libraries by Analyzing Networks of Molecules Assembled by Structural Similarity. *J. Chem. Inf. Model.* **2009**, *49*, 2429−2443.

(10) Igne, B.; Hurburgh, C. R. Local Chemometrics for Samples and Variables: Optimizing Calibration and Standardization Processes. *J. Chemom.* **2010**, *24*, 75−86.

(11) Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−799.

(12) Novotarskyi, S.; Sushko, I.; Korner, R.; Pandey, A. K.; Tetko, I. V. A Comparison of Different QSAR Approaches to Modeling CYP450 1A2 Inhibition. *J. Chem. Inf. Model.* **2009**, *49*, 2429−2443.

(13) Horvath, D.; Marcou, G.; Varnek, A. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.* **2009**, *49*, 1762−1776.

(14) Kaneko, H.; Arakawa, M.; Funatsu, K. Applicability Domains and Accuracy of Prediction of Soft Sensor Models. *AIChE J.* **2011**, *57*, 1506−1513.

(15) Kaneko, H.; Arakawa, M.; Funatsu, K. Novel Soft Sensor Method for Detecting Completion of Transition in Industrial Polymer Processes. *Comput. Chem. Eng.* **2011**, *35*, 1135−1142.

(16) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, *45*, 839−849.

(17) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Lo, J. Z.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L. L.; Liu, H. X.; Yao, X. J.; Oberg, T.; Hormozdiari, F.; Dao, P. H.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094−2111.

(18) Baskin, I. I.; Kireeva, N.; Varnek, A. The One-Class Classification Approach to Data Description and to Models Applicability Domain. *Mol. Inf.* **2010**, *29*, 581−587.

(19) Kaneko, H.; Funatsu, K. Estimation of Predictive Accuracy of Soft Sensor Models Based on Data Density. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(20) Kaneko, H.; Funatsu, K. A Soft Sensor Method Based on Values Predicted from Multiple Intervals of Time Difference for Improvement and Estimation of Prediction Accuracy. *Chemom. Intell. Lab. Syst.* **2011**, *109*, 197−206.

(21) Sheridan, R. P. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J. Chem. Inf. Model.* **2012**, *52*, 814−823.

(22) Sheridan, R. P. Using Random Forest To Model the Domain Applicability of Another Random Forest Model. *J. Chem. Inf. Model.* **2013**, *53*, 2837−2850.

(23) Kaneko, H.; Funatsu, K. Adaptive Soft Sensor Based on Online Support Vector Regression and Bayesian Ensemble Learning for Various States in Chemical Plants. *Chemom. Intell. Lab. Syst.* **2014**, *137*, 57−66.

(24) 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.

(25) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643−651.

(26) Sun, H. A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−757.

(27) Wegner, J. K.; Fröhlich, H.; Zell, A. Feature Selection for Descriptor Based Classification Models. 1. Theory and GA-SEC Algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 921−930.

(28) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.

(29) Clark, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **2005**, *45*, 30−38.

(30) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

(31) Kaneko, H.; Funatsu, K. Development of a New Regression Analysis Method Using Independent Component Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 534−541.

(32) http://www.talete.mi.it/products/dragon_description.htm (accessed March 31, 2014).

(33) Wold, S. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

(34) http://www.cadaster.eu/node/65 (accessed July 17, 2014).

(35) Owen, J. R.; Nabney, I. T.; Medina-Franco, J. L.; López-Vallejo, F. Visualization of Molecular Fingerprints. *J. Chem. Inf. Model.* **2011**, *51*, 1552−1563.

(36) http://www.chemaxon.com/ (accessed July 17, 2014).