ACS OMEGA

Open Access

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

Article

Evaluation and Optimization Methods for Applicability Domain Methods and Their Hyperparameters, Considering the Prediction Performance of Machine Learning Models

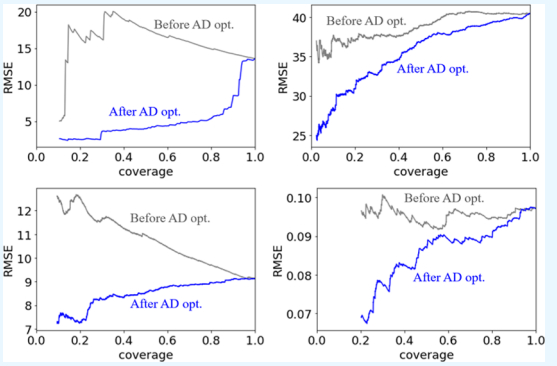
Hiromasa Kaneko*

Cite This: *ACS Omega* 2024, 9, 11453−11458

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: In molecular, material, and process design and control, the applicability domain (AD) of a mathematical model $y = f(x)$ between properties, activities, and features x is constructed. As there are multiple AD methods, each with its own set of hyperparameters, it is necessary to select an appropriate AD method and hyperparameters for each data set and mathematical model. However, there is no method for optimizing the AD model. This study proposes a method for evaluating and optimizing the AD model for each data set and a mathematical model. Using the predictions of double cross-validation with all samples, the relationship between coverage and root-mean-squared error (RMSE) was calculated for all combinations of AD methods and their hyperparameters, and the area under the coverage and RMSE curve (AUCR) was calculated. The AD model with the lowest AUCR value was selected as the optimal fit for the mathematical model. The proposed method was validated using eight data sets, including molecules, materials, and spectra, demonstrating that the proposed method could generate optimal AD models for all data sets. The Python code for the proposed method is available at https://github.com/hkaneko1985/dcekit.



1. INTRODUCTION

In molecular, material, and process design and control, a mathematical model $y = f(x)$ is constructed between objective variables y, including physical properties, activities, and product quality, and explanatory variables x, including molecular descriptors, experimental, synthesis, manufacturing, evaluation, process conditions, and variables. Using the constructed model, y values can be predicted from x values and x values can be designed with y as the target value.

Although it is critical to develop mathematical models with high predictive ability for data analysis and machine learning in molecular, material, and process design and control, the data domain in which the model can be applied is determined by the number of samples and their contents. When only a small number of samples exist, only a small data domain around the samples can be accurately predicted; however, as the number of samples increases, the data domain expands. This data domain is called the applicability domain (AD) of the model.[1] When the number of samples used to construct the model is small, there is a risk that the predicted y values are unreliable because they are outside the AD. However, because the AD is not prepared, the predicted y values are mistakenly accepted. Following the construction of model $y = f(x)$, it was necessary to develop an AD model. One of the organizations for economic cooperation and development principles for model validation requires defining the AD for machine learning models.[2]

Jaworska et al. examined and compared quantitative structure−activity relationship (QSAR) models in descriptor space via AD methods, including range, distance, geometry, and probability density distribution.[3] Sahigara et al. compared and visualized the results of AD methods for QSAR models, including range-based and geometric methods (for example, bounding box and convex full), distance-based methods, probability density distribution-based methods, and other methods (e.g., decision tree and stepwise approach), and concluded that it is preferable to evaluate the results from all possible methods before assessing a new data set.[4] Héberger used the sum of ranking differences to compare the statistics of QSAR models, statistical tests, and AD methods, including the Euclidean distance, Manhattan distance, Mahalanobis distance, five-nearest neighbor algorithm with the Euclidean distance, five-nearest neighbor algorithm with the Manhattan distance,

Received: October 13, 2023
Revised: January 19, 2024
Accepted: February 12, 2024
Published: February 26, 2024

five-nearest neighbor algorithm with the Mahalanobis distance, bounding box, convex hull, and potential function; however, the hyperparameters of each AD method were not discussed.[5] Zhong et al. developed an AD method based on uncertainty-based active learning for QSAR models that employ Gaussian process regression.[6] Kaneko proposed a machine learning model for accuracy based on model prediction results and AD.[7] Berenger and Yamanishi proposed a trivial-to-interpret and fully automatic distance-based Boolean AD method for category QSAR for high-throughput screening data classification, improving classification performance, early classifier retrieval, and scaffold diversity among top-ranking active molecules.[8] Rakhimbekova et al. investigated AD methods of QSAR and quantitative structure−property relationship models for chemical reactions and proposed new AD methods for reactions.[9] Banerjee and Roy used AD to identify prediction outliers and machine learning models were used for predictions after removing them from a human ether-a-go-go-related gene toxicity data set.[10] Conformal prediction is a rigorous and mathematically proven framework for in silico modeling that guarantees error rates and consistent handling of the AD, which is intrinsically linked to the underlying machine learning model.[11]

As briefly outlined in the AD study, there are various AD methods and hyperparameters for each. It is preferable to optimize the AD method and its hyperparameters for each data set and machine learning method. However, because AD modeling is an unsupervised learning process, AD cannot be optimized on its own. Therefore, this study proposes a method for optimizing the AD method and its hyperparameters, considering the predictive ability of the model $y = f(x)$. The predicted y values were calculated using double cross-validation (DCV),[12] and the relationship between coverage and root-mean-squared error (RMSE) was determined for each combination of AD methods and hyperparameters. The AD method and associated hyperparameters are selected to yield a favorable RMSE curve.

The proposed method was validated against eight real data sets, including molecules, materials, and spectra. Following the selection of the best machine learning method for each data set based on the predictive ability of the model among various machine learning methods, the proposed method is used to optimize the AD method and its hyperparameters.

2. METHODS

2.1. Flow of Machine Learning and Prediction. Due to the necessity of a data set for the construction of machine learning models, a comprehensive data set was first collected. Subsequently, they were preprocessed. For example, when each sample included a chemical structure, descriptors were calculated for each structure. Smoothing and differentiation were performed for each sample that represented a spectrum. Preprocessing included outlier detection, feature transformation, and feature standardization.

Following that the model is generalized for the target data set and y. For example, in machine learning method selection, the data set is divided into training and test data, and each machine learning method is evaluated by predicting the test data using the model constructed with the training data, or DCV is performed by repeating the division of training and test data. Subsequently, the machine learning method that produced the model with the highest predictive ability was chosen. This model optimization also includes feature engineering, feature selection, and preprocessing method selection.

The AD was prepared using the data set for the optimized model. The model and AD are then used to predict y values for new candidates for x or to design x such that y attains a target value. Bayesian optimization[13] and direct inverse analysis of the model[14] should be used when searching for extrapolation domains of existing data or outside the AD.

2.2. Applicability Domain. One method for determining an AD is to measure the distance from the mean of the data set. The AD is defined as the difference from the mean within the AD. Although AD can be determined by considering all x variables simultaneously, the distance from the mean works only when the samples are concentrically distributed away from the mean. Therefore, AD was determined via the data density. The number of training samples surrounding a sample is considered to determine whether the sample is within the AD. When a significant number of samples were near the sample, the data density was high and the sample was within the AD.

The k-nearest neighbor algorithm (kNN) is a data density index that takes the average of the distances between the k samples that are closest to one another.[15] The distance between a given sample and all other samples in the training data was calculated; k samples were taken in decreasing order, and the average of k distances was calculated. The lower the average, the higher the data density. Although k is commonly set to 5 or 10 to account for neighboring samples, there is no optimal value for k.

For kNN, a method that can account for differences in local data density in the data distribution is the local outlier factor (LOF),[16] where the index of AD is calculated by considering not only the distance to the k-nearest samples but also the distance to the k samples closest to those k samples. Although the k value is generally set to 5 or 10 to account for neighboring samples, there is no optimal value for k.

The one-class support vector machine (OCSVM)[17] method uses a support vector machine to solve the data domain estimation problem, allowing it to detect outlier samples while considering all x variables.

The fundamental formula for OCSVM is expressed as follows:

$$f(\mathbf{x}) = \phi(\mathbf{x})\mathbf{w} - b = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}^{(i)}, \mathbf{x}) - b \tag{1}$$

where $\mathbf{w}$, ϕ, $\mathbf{x}^{(i)}$, b, and K represent a weight vector, a nonlinear function, the x variables of the ith sample, a constant, and a kernel function, respectively. In this study, the following Gaussian kernel is used

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp(-\gamma \| \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \|^2) \tag{2}$$

In eq 1, α_i is computed by minimizing

$$\frac{1}{2} \| \mathbf{w} \|^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - b \tag{3}$$

subject to

$$\phi(\mathbf{x}^{(i)})\mathbf{w} \geq b - \xi_i$$
$$\xi_i \geq 0 \tag{4}$$

$\nu \in (0, 1)$ is interpreted as the fraction of outliers in the training data, that is, data for which $f(\mathbf{x}^{(i)}) < 0$. Γ can be determined to

https://doi.org/10.1021/acsomega.3c08036
ACS Omega 2024, 9, 11453−11458

maximize the variance in the Gram matrix of the Gaussian kernel for 2^{-15}, 2^{-14}, ..., 2^{0}, and 2^{1}. However, ν cannot be optimized.

To compare the performances of the AD indices, the coverage[7] was calculated after sorting the data in descending order of the AD index value

$$\text{coverage}_i = i/M \tag{5}$$

where M denotes the number of data points. Thus, coverage_i represents the proportion of data up to the ith data point, which is estimated to have a lower prediction error. The value RMSE_i corresponding to coverage_i is calculated as follows using Ii data points sorted in descending order of the AD index value

$$\text{RMSE}_i = \sqrt{\frac{\sum_{j=1}^{i} (y_{\text{obs},j} - y_{\text{pred},j})^2}{i}} \tag{6}$$

where $y_{\text{obs},j}$ represents the jth measured value of y, and $y_{\text{pred},j}$ represents the jth predicted value of y. For AD indices, RMSE_i should be low when coverage_i is low, and RMSE_i increases when coverage_i is large. If the two machine learning models are the same, then their RMSE_i values at $\text{coverage}_i = 1$ are equal for different methods of setting the AD.

2.3. Proposed AD Evaluation and Optimization Method. This study proposes a method for determining the optimal AD method and its hyperparameters using the prediction results of a machine learning model and the performance of AD. The proposed method proceeds as follows:

(1) Perform DCV on all samples and calculate the predicted y value for each sample.
   For each AD method and hyperparameter candidate, the following 2, 3, 4, and 5 are calculated:
(2) Calculate the AD index for each sample.
(3) Sort the samples by AD index values.
(4) Calculate coverage and RMSE, adding samples one by one.
(5) Calculate the area under the coverage and RMSE curve (AUCR),[7] which is the area of the lower part of the area, using the horizontal axis as the coverage and the vertical axis as the RMSE. The AUCR is calculated as follows:

$$\text{AUCR} = \frac{1}{2} \sum_{i=m}^{M} (\text{RMSE}_i + \text{RMSE}_{i-1}) (\text{coverage}_i - \text{coverage}_{i-1}) \tag{7}$$

   where m is the number of data points used to calculate the initial RMSE, which must be sufficiently large to stabilize the RMSE value.
(6) Select the AD method and hyperparameter that minimize the AUCR.

The proposed method allows for the selection of an appropriate AD model while simultaneously considering the predictive ability of the machine learning model and the performance of the AD.

Python codes for the proposed method are available at https://github.com/hkaneko1985/dcekit.

3. RESULTS AND DISCUSSION

To verify the performance of the proposed AD evaluation and optimization method, molecular data sets of boiling point (BP),[18] solubility in water ($\log S$, S = solubility at 20−25 °C in moles per liter),[19] melting point (MP),[20] and environmental toxicity (Tox)[21] were used. The Tox data set was derived from an online challenge in which researchers were asked to estimate the toxicity of molecules against *Tetrahymena pyriformis*. It included entries corresponding to the logarithm of the 50% growth inhibitory concentration (pIGC_{50}). Material data sets included superconducting (Tc)[22,23] and thermoelectric conversion (ZT) materials.[24] Similarly, the tablet data sets Shootout2002[25] (API1) and Shootout2012[26] (API2) were used for spectral data sets. These data sets represent real-world data. For compound data sets, molecular descriptors, or x variables, were calculated using RDKit,[27] which provides basic descriptors such as the number of atoms for each atom type, molecular weight, and descriptors including information on fragments, topology, and physicochemical properties. In Tc and ZT, x represents the fraction of each metal element, while y represents the critical temperature. In ZT, y represents the efficiency of thermoelectric conversion[24] and is calculated as

$$ZT = \frac{S^2\sigma}{\kappa} T \tag{8}$$

here κ, S, and σ represent the thermal conductivity, the Seebeck coefficient, and the electrical conductivity, respectively, while T represents the average temperature of the material proportional to conversion efficiency.

The following methods were used to construct machine learning models:

- Ordinary least-squares regression
- Partial least-squares regression
- Ridge regression
- LASSO regression
- Elastic net
- Support vector regression (linear and Gaussian kernels)
- Decision tree
- Random forests (RF)
- Gradient-boosting decision tree
- LightGBM
- XGBoost
- Gaussian process regression (GPR) (11 kernels)

Table 1. Machine Learning Method Selected with DCV for Each Data Set

data set	machine learning method	r^2 (DCV)	RMSE (DCV)
BP	GPR5	0.968	13.6
log S	GPR6	0.931	0.535
MP	GPR6	0.588	40.5
Tox	GPR10	0.850	0.414
Tc	RF	0.929	9.13
ZT	RF	0.755	0.0972
API1	GPR5	0.959	0.929
API2	RF	0.957	0.260

Table 2. AD Method and its Hyperparameters are Selected with the Proposed Method for Each Data Set

data set	AD method	hyperparameter
BP	kNN	6
log S	kNN	30
MP	kNN	6
Tox	kNN	3
Tc	LOF	1
ZT	OCSVM	0.02
API1	OCSVM	0.43
API2	OCSVM	0.13

https://doi.org/10.1021/acsomega.3c08036
ACS Omega 2024, 9, 11453−11458

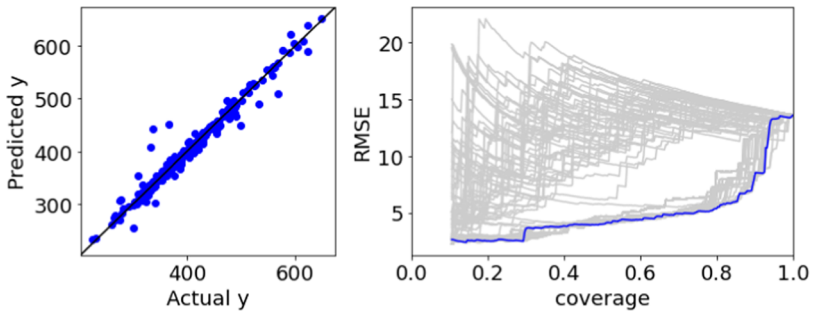


Figure 1. Plot of actual y vs. predicted y in DCV (left) and coverage and RMSE curve (right) for the BP data set.

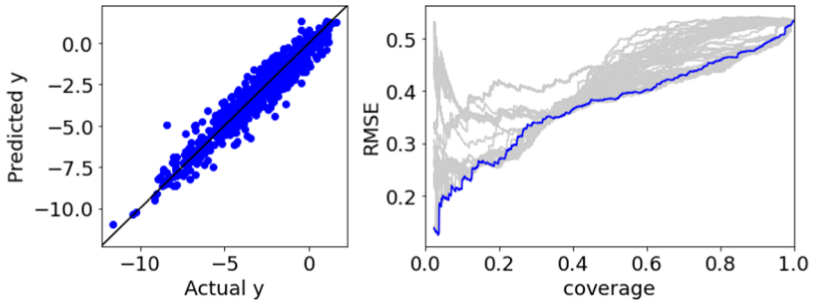


Figure 2. Plot of actual y vs. predicted y in DCV (left) and coverage and RMSE curve (right) for the log S data set.

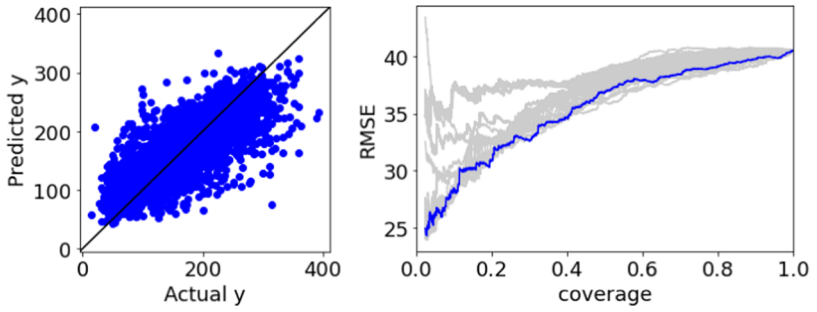


Figure 3. Plot of actual y vs. predicted y in DCV (left) and coverage and RMSE curve (right) for the MP data set.

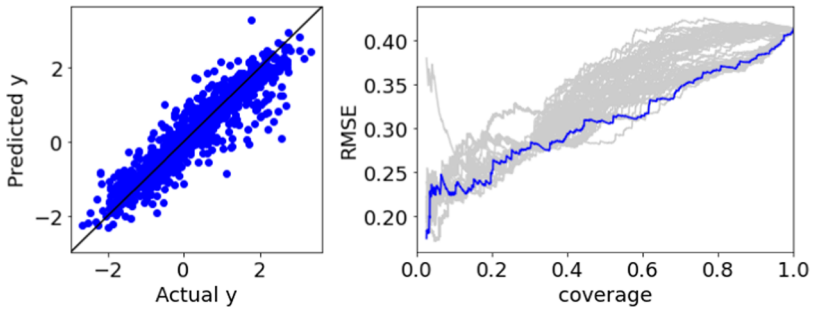


Figure 4. Plot of actual y vs. predicted y in DCV (left) and coverage and RMSE curve (right) for the Tox data set.

- Deep neural networks

The hyperparameters in each method are determined using 5-fold cross-validation.

For each data set, DCV was performed with 10 outer divisions, and the method with the highest r^2 was selected. Table 1 lists the selected method and the r^2 and RMSE for each data set. The table shows the method used to discuss AD optimization for each data set.

The AD methods used in this study are kNN, LOF, and OCSVM. The candidates for each hyperparameter were: k for kNN was 1, 2, 3, ..., 29, 30; k for LOF was 1, 2, 3, ..., 29, 30; and ν in OCSVM was 0.01, 0.02, 0.03, ..., 0.49, 0.50. The proposed method is used to calculate the AUCR for all AD method combinations and hyperparameter values.

Table 2 lists the AD method and hyperparameters that minimize the AUCR for each data set. The plots of measured y vs. predicted DCV and the coverage and RMSE curves for Table 2 are depicted in Figures 1−8. The blue line in each coverage and RMSE curve represents the result of optimizing the AD method and its hyperparameters using the proposed method, and the gray lines represent the other results. It was confirmed

https://doi.org/10.1021/acsomega.3c08036
ACS Omega 2024, 9, 11453−11458

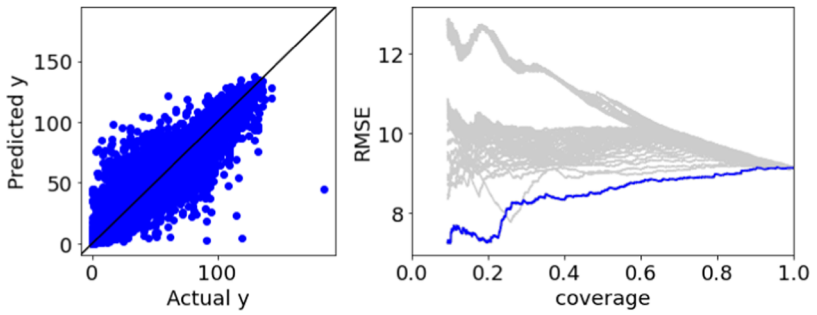


Figure 5. Plot of actual *y* vs. predicted *y* in DCV (left) and coverage and RMSE curve (right) for the Tc data set.

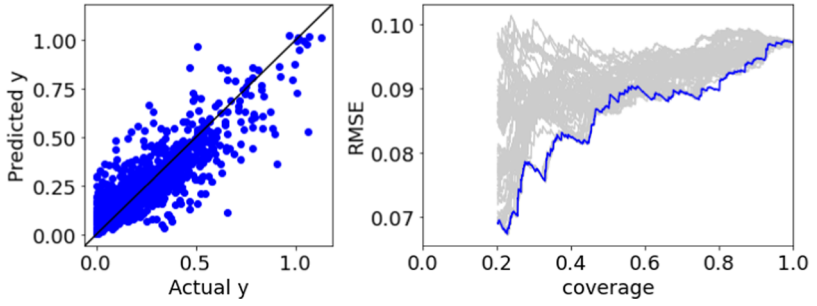


Figure 6. Plot of actual *y* vs. predicted *y* in DCV (left) and coverage and RMSE curve (right) for the ZT data set.

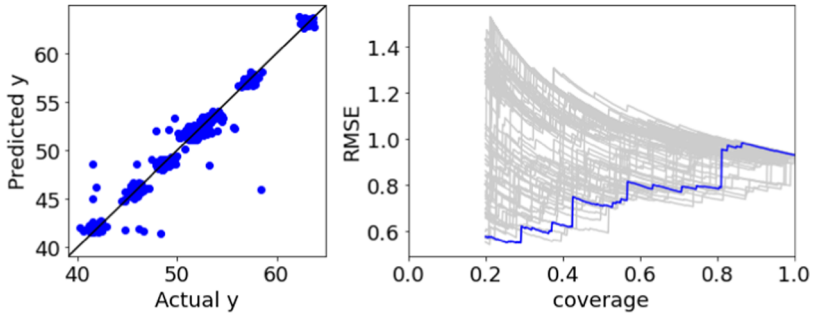


Figure 7. Plot of actual *y* vs. predicted *y* in DCV (left) and coverage and RMSE curve (right) for the API1 data set.

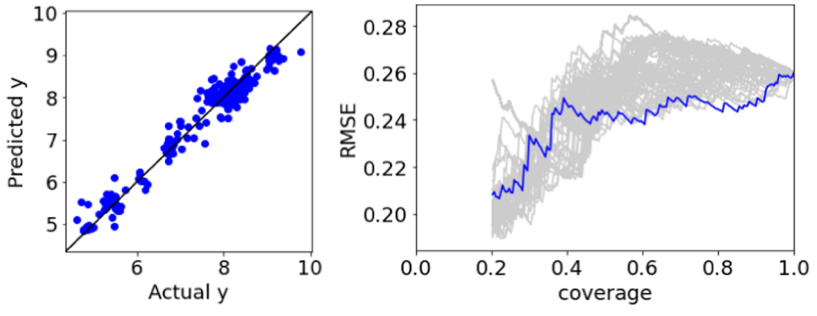


Figure 8. Plot of actual *y* vs. predicted *y* in DCV (left) and coverage and RMSE curve (right) for the API2 data set.

that the appropriate AD methods and hyperparameters for which the coverage and RMSE curves were located were to the lower right of the gray line for all data sets. The reliability of predictions for new samples is discussed using the AD model selected for each data set. In the gray API2 results, RMSE was small when coverage was low, but this was only because the RMSE was not stable when coverage was low, and the number of samples used to calculate RMSE was also low; thus, RMSE increased when coverage was high. The proposed method prevents the selection of unstable results. The proposed method effectively optimizes the AD method and its hyperparameters.

4. CONCLUSIONS

This study proposes a method for evaluating and optimizing the AD model and its hyperparameters to properly operationalize the AD model. The construction of an AD model exhibiting efficacy with respect to new samples is accomplished using a combination of the AD method and hyperparameters that reduce the AUCR. The proposed method was validated against

https://doi.org/10.1021/acsomega.3c08036
ACS Omega 2024, 9, 11453−11458

eight data sets, including molecules, materials, and spectra. The proposed method was demonstrated to be capable of optimizing the AD method and its hyperparameters on all data sets. The reliability of the predicted *y* values can be discussed by using the selected AD method and hyperparameters. Although this study used only three methods: kNN, LOF, and OCSVM, the proposed method can be used to compare other AD methods and their hyperparameters. Furthermore, when a new AD method is developed, its validity is evaluated against the proposed method. The proposed method is expected to facilitate the design of molecules, materials, and processes using mathematical models constructed via machine learning.

■ ASSOCIATED CONTENT

Data Availability Statement

The data supporting the findings of this study are available as refs 18−26.

■ 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.3c08036

Notes

The author declares no competing financial interest.

■ REFERENCES

(1) Kaneko, H. Discussion on regression methods based on ensemble learning and applicability domains of linear submodels. *J. Chem. Inf. Model.* **2018**, *58*, 480−489.
(2) OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models* OECD Publishing: Paris, France; 2007.
(3) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *Altern. Lab. Anim.* **2005**, *33* (5), 445−459.
(4) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791−4810.
(5) Héberger, K. Selection of optimal validation methods for quantitative structure−activity relationships and applicability domain. *SAR QSAR Environ. Res.* **2023**, *34* (5), 415−434.
(6) Zhong, S.; Lambeth, D.-R.; Igou, T.-K.; Chen, Y. Enlarging applicability domain of quantitative structure−activity relationship models through uncertainty-based active learning. *ACS ES&T Eng.* **2022**, *2* (7), 1211−1220.
(7) Kaneko, H. A new measure of regression model accuracy that considers applicability domains. *Chemom. Intell. Lab. Syst.* **2017**, *171*, 1−8.
(8) Berenger, F.; Yamanishi, Y. A distance-based boolean applicability domain for classification of high throughput screening data. *J. Chem. Inf. Model.* **2019**, *59* (1), 463−476.
(9) Rakhimbekova, A.; Madzhidov, T.-I.; Nugmanov, R.-I.; Gimadiev, T.-R.; Baskin, I.-I.; Varnek, A. Comprehensive analysis of applicability domains of qspr models for chemical reactions. *Int. J. Mol. Sci.* **2020**, *21*, 5542.
(10) Banerjee, A.; Roy, K. Machine-learning-based similarity meets traditional QSAR: "q-RASAR" for the enhancement of the external predictivity and detection of prediction confidence outliers in an hERG toxicity dataset. *Chemom. Intell. Lab. Syst.* **2023**, *237*, No. 104829.
(11) Alvarsson, J.; McShane, S.-A.; Norinder, U.; Spjuth, O. Predicting with confidence: using conformal prediction in drug discovery. *J. Pharm. Sci.* **2021**, *110*, 42−49.
(12) Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated double cross validation. *J. Chemom.* **2009**, *23*, 160−171.
(13) Ando, T.; Shimizu, N.; Yamamoto, N.; Matsuzawa, N.; Maeshima, H.; Kaneko, H. Design of molecules with low hole and electron reorganization energy using DFT calculations and Bayesian optimization. *J. Phys. Chem. A* **2022**, *126* (36), 6336−6347.
(14) Kaneko, H. Adaptive design of experiments based on Gaussian mixture regression. *Chemom. Intell. Lab. Syst.* **2021**, *208*, No. 104226.
(15) Korolev, V.; Nevolin, I.; Protsenko, P. A universal similarity-based approach for predictive uncertainty quantification in materials science. *Sci. Rep.* **2022**, *12*, No. 14931.
(16) Alghushairy, O.; Alsini, R.; Soule, T.; Ma, X. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. *Big Data Cogn. Comput.* **2021**, *5*, 1.
(17) Al Shorman, A.; Faris, H.; Aljarah, I. Unsupervised intelligent system based on one class support vector machine and Grey Wolf optimization for IoT botnet detection. *J. Ambient Intell. Hum. Comput.* **2020**, *11*, 2809−2825.
(18) 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.
(19) 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.
(20) Karthikeyan, M.; Glen, R. C.; Bender, A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* **2005**, *45*, 581−590.
(21) http://www.cadaster.eu/node/65.html (accessed August 19, 2023).
(22) Hamidieh, K. A data-driven statistical model for predicting the critical temperature of a superconductor. *Comput. Mater. Sci.* **2018**, *154*, 346−354.
(23) https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data# (accessed August 19, 2023).
(24) Katsura, Y.; Kumagai, M.; Kodani, T.; Kaneshige, M.; Ando, Y.; Gunji, S.; Imai, Y.; Ouchi, H.; Tobita, K.; Kimura, K.; Tsuda, K. Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials. *Sci. Technol. Adv. Mater.* **2019**, *20*, 511−520.
(25) Wehrens, R. *Chemometrics with R − Multivariate Data Analysis in the Natural Sciences and Life Sciences*; Springer, 2011.
(26) 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, DOI: 10.1366/0003702021955358.
(27) http://www.rdkit.org/ (accessed August 19, 2023).