JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Cite This: J. Chem. Inf. Model. 2019, 59, 181−189

Article

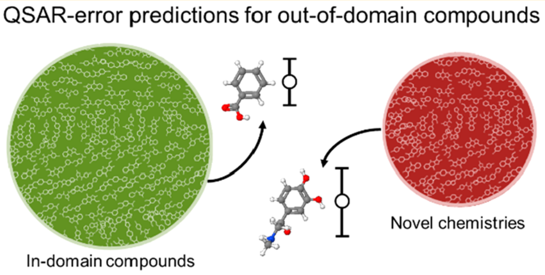
pubs.acs.org/jcim

Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds

Ruifeng Liu* and Anders Wallqvist*

Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland 21702, United States

Supporting Information



ABSTRACT: Domain applicability (DA) is a concept introduced to gauge the reliability of quantitative structure−activity relationship (QSAR) predictions. A leading DA metric is *ensemble variance*, which is defined as the variance of predictions by an ensemble of QSAR models. However, this metric fails to identify large prediction errors in melting point (MP) data, despite the availability of large training data sets. In this study, we examined the performance of this metric on MP data and found that, for most molecules, ensemble variance increased as their structural similarity to the training molecules decreased. However, the metric decreased for "out-of-domain" molecules, i.e., molecules with little to no structural similarity to the training compounds. This explains why ensemble variance fails to identify large prediction errors. In contrast, a new *molecular similarity*-based DA metric that considers the contributions of all training molecules in gauging the reliability of a prediction successfully identified predictions of MP data for which the errors were large. To validate our results, we used four additional data sets of diverse molecular properties. We divided each data set into a training set and a test set at a ratio of approximately 2:1, ensuring a small fraction of the test compounds are out of the training domain. We then trained random forest (RF) models on the training data and made RF predictions for the test set molecules. Results from these data sets confirm that the new DA metric significantly outperformed ensemble variance in identifying predictions for out-of-domain compounds. For within-domain compounds, the two metrics performed similarly, with ensemble variance marginally but consistently outperforming the new DA metric. The new DA metric, which does not rely on an ensemble of QSAR models, can be deployed with any machine-learning method, including deep neural networks.

INTRODUCTION

In the field of quantitative structure−activity relationship (QSAR) modeling of molecular activities, a subject of active research is the estimation of the reliability of QSAR predictions.[1−4] The concept of domain applicability (DA) was introduced in accord with the hypothesis that each QSAR model is applicable to molecules from a certain part of the chemical space. The reliability of a model prediction can then be judged from the relationship between the molecules under investigation and the domain: the prediction is considered reliable if a molecule is within the domain, but increasingly less so the further it is from the domain.

Many DA metrics have been defined to facilitate the quantitative estimation of prediction errors. The most intuitive is the distance-to-model metric, i.e., the distance between a molecule and the model training set.[5] This metric is commonly defined as the distance between a molecule and its closest neighbor in the training set, or the average distance between a molecule and its k closest neighbors in the training set.[6] To date, a leading DA metric is *ensemble variance*,[5,7] which is defined as the variance of predictions given by an ensemble of QSAR models for the same molecule. This metric was the focus of several detailed studies in which it outperformed the distance-to-model metric for both regression and classification problems.[5,8] Intriguingly, the results achieved by ensemble variance led the investigators to conclude that the prediction error associated with a molecule does not depend on the machine-learning method or input features but rather on the distance to the training molecules.[5] The distance-to-model metric and ensemble variance were also used to estimate the prediction errors of melting point (MP) models trained on large data sets. In this case, ensemble variance failed to identify large prediction errors, while the distance to the nearest training molecule performed best among the evaluated DA metrics.[9] However, the investigators noted that the distance-to-model metric was not sufficient for practical use.

The distance-to-model metrics evaluated in previous studies have two flaws: (1) only a limited number of nearest training molecules were considered to contribute to prediction accuracy and (2) the nearest training molecules were considered to contribute equally. In our view, all training molecules contribute, but not equally, to the prediction accuracy of a model. Specifically, we have suggested that the

Received: September 2, 2018
Published: November 7, 2018

DOI: 10.1021/acs.jcim.8b00597

contribution of a training molecule should be inversely correlated with the distance to the target molecule for which the prediction accuracy is assessed. On the basis of these considerations, we defined the sum of distance-weighted contributions (SDC) of all training molecules as a new DA metric and recently showed that it correlates well with prediction error.[10]

In this study, we examined the performance of SDC on MP data for which ensemble variance failed to identify large prediction errors. We found that SDC successfully identified predictions with large errors. Detailed analyses indicated that the ensemble variance for compounds with little to no structural similarity to the training samples were surprisingly lower than expected. This finding explains why ensemble variance cannot identify MP data for which prediction errors are large. To ensure that the findings were not restricted to MP data but were generally valid for any data set, we confirmed these findings in similar studies on four additional data sets of distinctively different molecular properties.

MATERIALS AND METHODS

Data Sets. We downloaded the MP data from Online Chemical Modeling Environment (OCHEM)—a web platform for data storage, model development, and publishing of chemical information.[11] The data were collected from several sources and curated by Tetko et al., who used them in their QSAR study. They are comprised of four data sets: the Bergstrom set of 277 drug-like compounds,[12] the Bradley set of 2886 compounds,[13] the OCHEM set of 21 883 compounds,[9] and the Enamine set of 22 404 compounds.[14] Although there were considerably more compounds in the original data sets, Tetko et al. removed mixtures, salts, and compounds that failed at least one descriptor calculation program. They also ensured that each compound belonged to only a single data set so that the data sets did not share any compounds. We refer the reader to the article by Tetko et al.[9] for further details on the data sets. We used the downloaded data sets without making any changes.

To ensure that the findings of this study were not limited to MP data, we applied the same approach to four additional molecular property and bioactivity data sets. They included a molecular lipophilicity data set consisting of 10 178 molecules, an acute rat oral toxicity data set of 6734 molecules, a human leukemia cell line growth inhibition data set of 2000 molecules, and an aqueous solubility data set of 1144 molecules. The lipophilicity data set was an example data set in BIOVIA's Pipeline Pilot (http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/). The lipophilicity of each compound in the data set is given by the logarithm of the partition coefficient of the compound between the 1-octonal and water phases (logP). We downloaded the acute rat oral toxicity data set from the U.S. National Toxicology Program Web site (https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/acute-systemic-tox/models/index.html). This is the training data set for the Predictive Models for Acute Oral Systemic Toxicity Challenge, sponsored by the national Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM). Each compound in the data set has an experimentally determined LD50 value in milligrams per kilogram of body weight. We downloaded the leukemia cell growth inhibition data set from PubChem (https://pubchem.ncbi.nlm.nih.gov/) with assay ID 121. The assay determined 50% growth inhibition (GI50) values for 3223 chemical samples (of the 41 721 tested) that met active criteria via dose response measurements. Of these, we removed samples without molecular structure information, as well as replicate entries of the same compound by taking the average of their GI50 values as the GI50 for that compound. After these steps, we ended up with a data set consisting of 2000 structurally unique compounds with GI50 values. We downloaded the aqueous solubility data set from the *Journal of Chemical Information and Modeling* Web site (https://pubs.acs.org/doi/suppl/10.1021/ci034243x). This is the data set that Delaney used in his study of aqueous solubility.[15]

New DA Metric. Our DA metric is defined as

$$\mathrm{SDC} = \sum_{i=1}^{n} \mathrm{e}^{-3\mathrm{TD}_i/1-\mathrm{TD}_i} \tag{1}$$

where SDC—the sum of the distance-weighted contributions—gauges QSAR prediction accuracy, TD_i is the Tanimoto distance (TD) between a target molecule and the *i*th training molecule, and *n* is the total number of training molecules. The TD between two molecules is calculated by using the extended connectivity fingerprint with a diameter of four chemical bonds (ECFP_4).[16] The TD value between two molecules ranges between 0 and 1—the lower and upper limits corresponding to two molecules sharing all and no fingerprint features, respectively.

Machine-Learning Method. In this study, we chose to use random forest (RF) to build QSAR models based on the following considerations: (1) it is one of the most popular machine-learning methods, and (2) it is an ensemble method. For an RF model, which employs a large number of decision trees, the standard deviation of all tree predictions serves as a measure of ensemble variance.[17] Thus, RF models allow for an expedient comparison of SDC and ensemble-variance metrics. In this study, each RF model consisted of 500 decision trees. The input molecular descriptors were the counts of 2048 ECFP_2 fingerprint features (predefined molecular fragments) termed ECFC_2, which we showed to perform well in our previous QSAR studies of log*P* and MP.[18,19]

RESULTS AND DISCUSSION

Comparison of SDC and Ensemble Variance for MP Data Sets. Tetko et al. trained QSAR models with the OCHEM and Enamine data sets, using five machine-learning methods and multiple descriptor sets. To assess the performance of the developed QSAR models, they first carried out 5-fold cross validation of the training sets and then used the developed models to predict the MPs for the compounds in the other three data sets. They found that models developed with an associative neural network (ASNN) performed best with Estate index descriptors. They also trained many ASNN models using different descriptor sets and considered the average of all model predictions as the consensus prediction.[9] We compared the root mean squared errors (RMSEs) of their ASNN models with those derived from our RF/ECFC_2 models (Table 1). The overall performance of the ASNN models was comparable to that of the RF models. For both methods, model performance, as measured by the RMSE derived from 5-fold cross validation, was in line with the RMSE of the test sets for models trained with the OCHEM data set but not for models trained with the Enamine data set. The cross validation RMSEs of the models trained with the

DOI: 10.1021/acs.jcim.8b00597

Table 1. Root Mean Squared Errors of Melting Point Predictions for Different Data Sets[a]

method	training set	CV training	Bradley	Bergstrom	Enamine
ASNN best (Estate)[b]	OCEHM	41.6	36.6	36.0	43.1
RF/ ECFC2_2048[c]	OCHEM	41.7	36.8	36.8	42.6
ASNN best (Estate)[b]	Enamine	38.7	66.0	44.0	54.6
RF/ ECFC2_2048[c]	Enamine	38.8	76.6	42.0	57.0

[a] The predictions were made by random forest models trained on the OCHEM or Enamine data sets. [b] Results of Tetko et al.[9] using the associated neural network method and Estate index descriptors. [c] Results of the current study using a random forest model consisting of 500 decision trees and the counts of 2048 ECFP_2 fingerprint features as descriptors.

Enamine data set were significantly lower than the RMSEs of the models for the test sets, indicating that cross validation performance is not an accurate performance indicator for general applications.

To assess whether ensemble variance and SDC could effectively identify predictions associated with large errors, we calculated the values of SDC and standard deviation (STD) for all RF model predictions and then ranked the predictions according to each metric. We hypothesized that the higher the SDC and the lower the STD, the more reliable the prediction. If this hypothesis is correct, then removing the lowest-ranked predictions (those associated with the lowest SDC or highest STD values) should lead to lower RMSEs for the remaining predictions.

Figure 1A presents the RMSEs of predictions by the RF model trained with the OCHEM data set and the resulting RMSEs after successive removal of the predictions ranked lowest by SDC and STD. For clarity of exposition, we have omitted the results for the Bergstrom data set, which showed a less consistent trend because of the small number of compounds. Successive removal of the predictions ranked lowest by either metric resulted in smaller RMSEs. The RSMEs of the two metrics tracked each other closely, with STD marginally outperforming SDC in all cases.

These results were in stark contrast to the RMSEs of the predictions made by the RF model trained on the Enamine data set (Figure 1B). The 5-fold cross validation RMSEs (red and blue lines) closely resembled those of the model trained on the OCHEM data set (Figure 1A, red and blue lines). In contrast, the performance on the test sets markedly differed. For example, when the OCHEM data set served as the test set, the RMSE remained nearly constant despite successive removal of predictions ranked lowest by STD but decreased with removal of predictions ranked lowest by SDC (Figure 1B). That is, STD failed to identify predictions with large errors, whereas SDC successfully identified and removed these predictions. This contrast was more pronounced in the RMSEs of the Bradley test set: whereas successive removal of predictions ranked lowest by SDC considerably reduced the RMSEs of the remaining predictions, removal of predictions ranked lowest by STD gradually increased the RMSE of the remaining predictions. In other words, only SDC was successful in removing predictions with large errors.

Differences in the distribution of samples across the melting point range and coverage of chemical space by the data sets offer clues for understanding the performance disparity between the RF models trained with the OCHEM and Enamine data sets. Tetko et al. showed in Figure 1 of their article that the OCHEM data set had the broadest distribution of samples across the −100 to 400 °C MP range.[9] In contrast, in the Enamine data set, which had slightly more compounds, the number of compounds with a MP below freezing was zero. The two smaller data sets had markedly different sample distributions. Most compounds in the Bergstrom data set were drug-like, with MPs between 50 and 250 °C, closely tracking the MP distribution of the Enamine data set. In contrast, of all the data sets, the Bradley data set had the highest percentage of compounds with a MP below freezing.

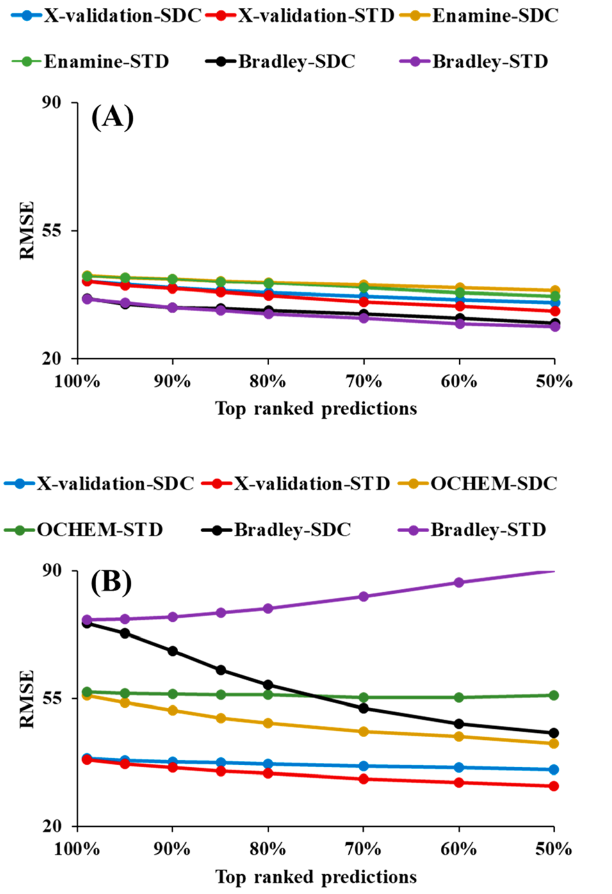


Figure 1. Root mean squared error (RMSE) of top SDC- and STD-ranked melting point predictions by random forest (RF) models trained on the OCHEM data set (A) and on the Enamine data set (B). The top panel shows that successive removal of the lowest SDC- and STD-ranked predictions by the RF model trained on the OCHEM data set consistently led to lower RMSEs. In contrast, the bottom panel shows that removal of the lowest STD-ranked predictions by RF models trained on the Enamine data set did not lead to lower RMSEs for the test sets, whereas removal of the lowest SDC-ranked predictions did.

To assess the overlap of chemical spaces between data sets, we calculated the TDs between the test set compounds and their closest neighbors in OCHEM and Enamine training sets and counted the number of compounds with TDs of 0.8 or longer—those with little to no structural similarity to (i.e., outside the domain of) the training set compounds. Table 2 shows the counts indicating that the OCHEM chemical space almost completely encloses the chemical spaces of the Enamine, Bradley, and Bergstrom data sets, given that the

Table 2. Number of Compounds of a Data Set with Little or No Structural Similarity (Tanimoto Distance ≥ 0.8) to Compounds in Another Data Set

	OCHEM	Enamine	Bradley	Bergstrom
to OCHEM data set		3	5	0
to Enamine data set	459		91	2

number of compounds with TDs 0.8 or longer to compounds of the OCHEM data set is nearly zero for each of these test sets. In contrast, a small but non-negligible percentage of the Bradley and OCHEM data sets is outside the domain of the Enamine data set. The results of our previous study indicated that for compounds with little or no structural similarity to those of a training set, the predictions of various machine-learning models were nearly constant and uncorrelated with the experimental results.[20] To examine whether this was also the case for MP predictions, we plotted MPs predicted by the RF model trained on the Enamine data against the experimental results for compounds in the Bradley data set with the shortest TDs 0.8 or longer to the training set (Figure 2). For these compounds, the predictions were nearly constant across MPs ranging from −200 to +200 °C. That the predictions for these "out-of-domain" compounds are nearly constant suggests that the ensemble variance for these compounds is low. This is corroborated by our observation that successive removal of predictions ranked lowest by STD failed to remove predictions with large errors (Figure 1B).

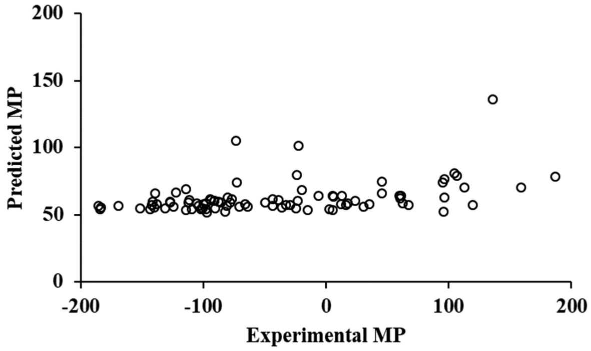


Figure 2. Predicted versus experimentally measured melting points of compounds in the Bradley data set, which are at least a Tanimoto distance of 0.8 away from compounds in the Enamine data set. The predictions, made by a random forest model trained on the Enamine data set, are nearly constant, in sharp contrast to the experimental values that span a range of 400 °C.

To confirm that out-of-domain compounds caused ensemble variance to fail in identifying large prediction errors, we removed 987 compounds (see Supporting Information for compound IDs) in the OCHEM training set to make 91 Bradley compounds out of the OCHEM domain. We retrained the RF model using the remaining molecules of the OCHEM data set and estimated its performance by 5-fold cross validation. We also used the new RF model to make predictions for the Bradley data set. The resulting RMSEs after successive removal of the predictions ranked lowest by SDC and STD (Figure 3) show that the RMSE curves of 5-fold cross validation of the reduced OCHEM training set (red and blue lines) were nearly identical to the corresponding RMSE curves in Figure 1A. However, the RMSE curves of the Bradley data set (Figure 3, black and purple lines) were markedly different from those obtained with the RF model trained on the full OCHEM data set (Bradley data set in Figure 1A); instead, they were remarkably similar to those obtained with the RF model trained on the Enamine data set (Figure 1B). These results confirm that the failure of ensemble variance to identify large prediction errors is due to out-of-domain compounds for which the ensemble variance was lower than expected.

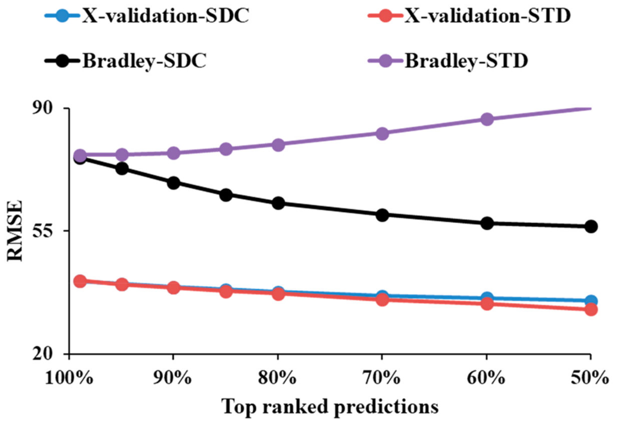


Figure 3. Root mean squared error (RMSE) of predictions of melting points of the Bradley data set. The predictions were made by a random forest model trained on the OCHEM data set with some compounds removed to render 91 compounds of the Bradley data set outside of the training domain (i.e., with the shortest Tanimoto distance of 0.8 or higher to the remaining compounds in the OCHEM data set). The RMSEs derived from 5-fold cross validation of the reduced OCHEM data set are also shown.

The results obtained with MP data suggest that for within-domain molecules, i.e., molecules with a TD up to 0.8 to a training molecule, the ensemble variance increases with TD, whereas for out-of-domain molecules, it decreases with TD. To test this conjecture, we plotted the STD of ensemble model predictions made for the Bradley compounds by the RF model trained on the Enamine data set, against the shortest TDs of these compounds to the training set molecules (Figure 4A). As expected, the plot revealed a tendency for STD to increase with TD and then decrease with increasing TD in the range of 0.8 to 1.0. We also plotted the average STDs of the predictions in each TD bin of size 0.05 across the entire range of TD values in Figure 4B, which shows the same tend we observed in Figure 4A.

This trend for the ensemble variance to first increase and then decrease as a function of TD may seem counterintuitive. To understand why, we can assume that the MP model is a function of descriptor sets **X** and **Y**, as shown in eq 2 below.

$$y = c + f(\mathbf{X}) + f'(\mathbf{Y}) \quad (2)$$

Here, c is a constant, **X** is a set of molecular descriptors, with x_i present in training set molecules, and **Y** is a set of molecular descriptors, with y_i present in out-of-domain test molecules. Because out-of-domain molecules share little to no structural similarity with the training molecules, we can reasonably assume that none of the descriptors y_i are present in **X** and none of the descriptors x_i are present in **Y**. Because **X** and **Y** do not overlap, a model developed from the training data alone will be reduced to

$$y = c + f(\mathbf{X}) \quad (3)$$

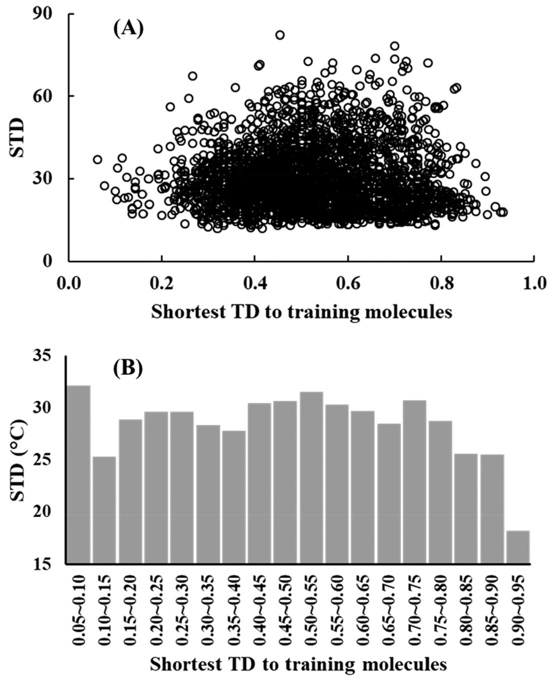


Figure 4. (A) Scatterplot of standard deviations (STDs) of melting point predictions for molecules in the Bradley data set made by 500 decision trees of the random forest model trained on the Enamine data set as a function of the shortest Tanimoto distance to the training molecules. (B) Mean STD of predictions in each TD bin of size 0.05. There are only two compounds in the first bin, which explains the usually high mean STD.

Given that none of x_i are present in out-of-domain molecules, the predicted MP for all out-of-domain molecules will be the same (given by the constant c above). This is consistent with the distribution of data points in Figure 2 and the unexpected decrease in ensemble variance for out-of-domain molecules.

Comparisons for Other Data Sets. To ensure the generality of the findings concerning predictions for out-of-domain compounds derived from the MP data, we compared SDC and STD for four additional data sets that cover a wide range of molecular properties. To model these properties, we first log-transformed the experimentally determined property values—a standard practice in the field. We then divided each data set into a training and a test set. To secure a fraction of the test samples outside of the training-set domain, for each data set we first examined the distribution of samples along the molecular property of interest. As in most molecular activity data sets, the distribution of samples was highly uneven (Figure 5). In each data set, most samples populated a limited number of activity bins and the percentage of compounds with extremely high or extremely low activity values was very low. An apparent exception is the distribution of samples for the data set on leukemia cell growth inhibition, which had no compounds with a log GI50 value greater than −6. This is an artifact, however, because such compounds were considered inactive and, therefore, were not subjected to dose−response measurements.

The highly uneven distributions shown in Figure 5 dictate that the value c in eq 3 is most likely very close to the activity of the highest populated bins of each data set. This is because the objective of training a model, irrespective of the machine-learning method used, is to minimize the error between the predicted and experimental values of all training samples. Statistically, we can achieve this objective most efficiently when the constant c is close to the activity of an overwhelmingly large number of training samples. Consequently, if the

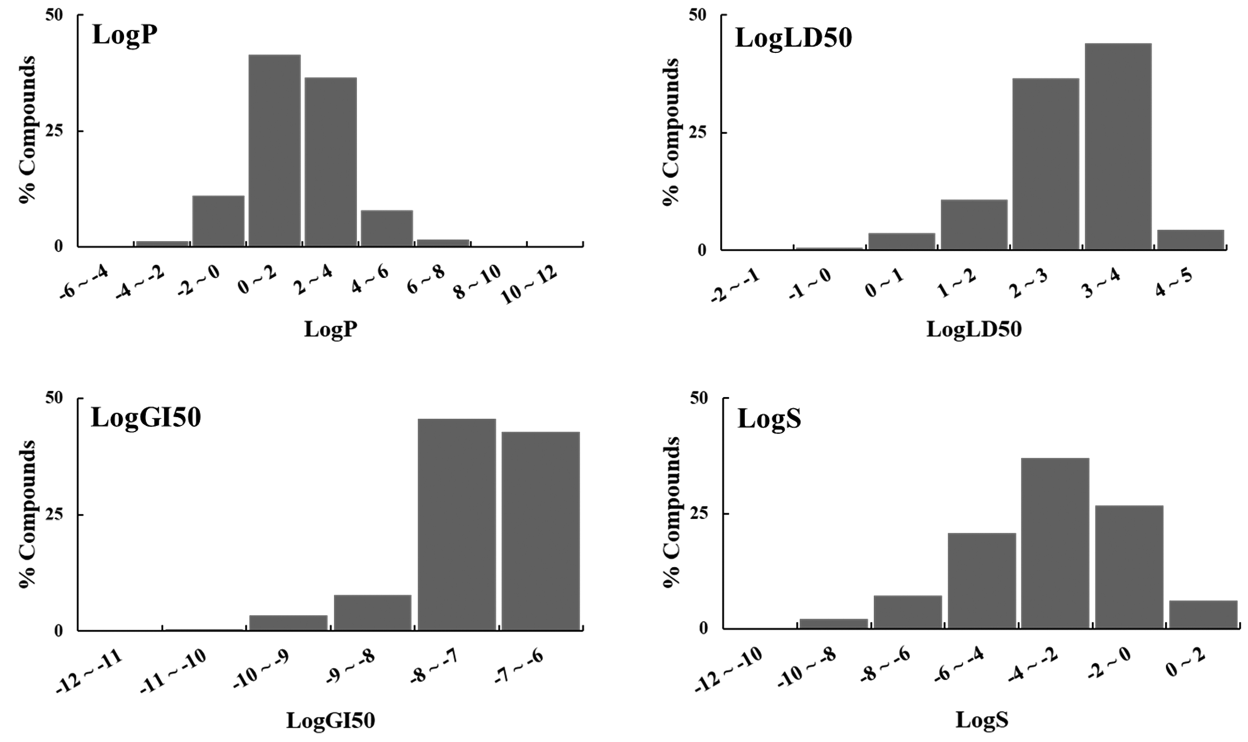


Figure 5. Distributions of samples for the four molecular property/activity data sets used in this study to test the generality of the findings derived from the melting point data.

DOI: 10.1021/acs.jcim.8b00597

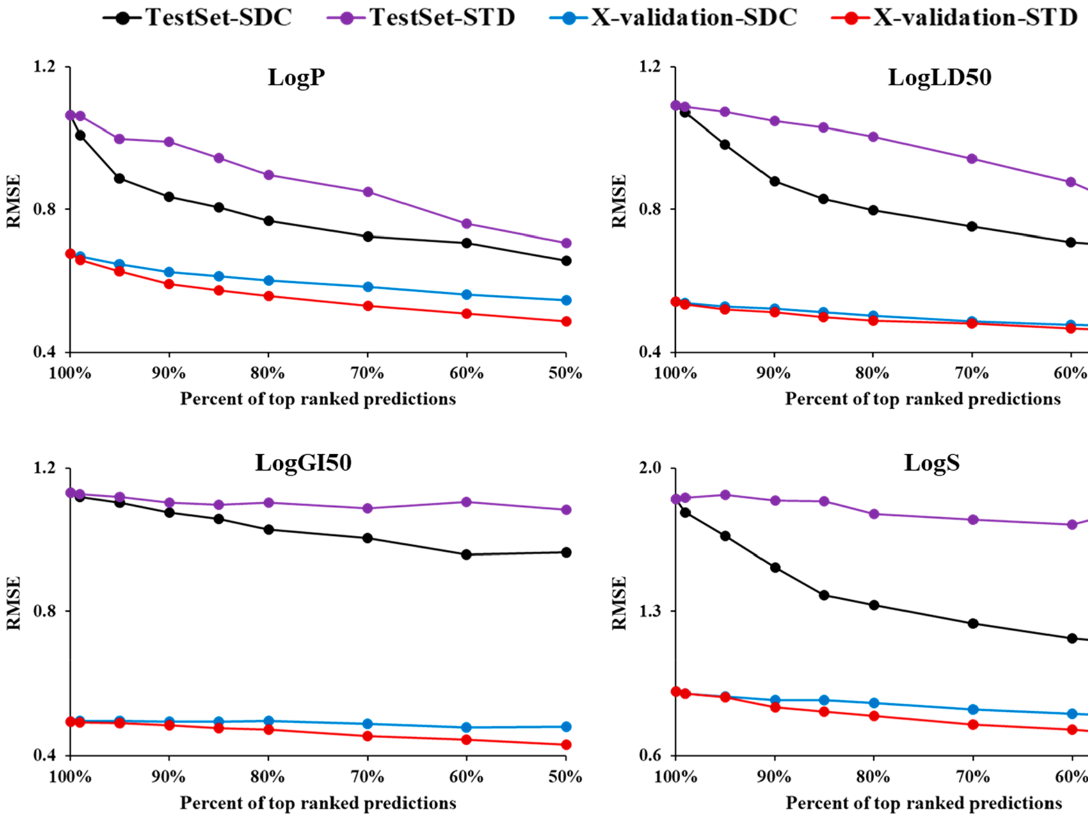


Figure 6. Root mean squared errors (RMSEs) of random forest predictions for test sets of the four molecular property data sets. The test sets contain molecules outside of the training set domains. Also shown are RMSEs of 5-fold cross validation of the training sets. The plots show that SDC was more efficient than STD at identifying predictions with large errors in the test sets. In contrast, SDC and STD performed similarly for 5-fold cross validation, with the latter performing slightly better for all data sets.

activities of out-of-domain compounds are in or around the highest populated bins, we could not achieve our objective because the prediction errors for these compounds will be small—not because the predictions are more reliable but because the experimental values happen to be close to *c*.

For MP data, the performance disparity between SDC and STD was most obvious when we used the Bradley data set as a test set for models trained on the Enamine data. The distributions of samples presented in Figure 1 of the article by Tetko et al. show that the MPs for a considerable portion of the Bradley data set compounds are below freezing, while none of the compounds in the Enamine data set have subzero MPs. When the compounds with MPs below freezing are outside of the Enamine data domain, models trained on the Enamine data set gave large prediction errors but with low ensemble variance. We believe this is why the disparity in performance between SDC and STD is so apparent.

On the basis of the considerations above, to divide each of the additional data sets into a training set and a test set with a portion of the compounds outside of the training domain, we first selected a small number of compounds with extreme (highest or lowest) activity levels. We then performed a structural similarity search using these compounds as queries to identify all compounds in the data set that fell within a TD of 0.8 to the query compounds. We then combined these compounds with the query compounds to serve as the test set. The remaining compounds served as the training set. In this way, we ensured that the selected compounds with extreme activity levels were outside of the training set domain.

The training and test compounds selected for the four data sets by the approach described above are given in the Supporting Information. The ratio of the number of training compounds to the number of test compounds is roughly 2 to 1. For each data set, we trained an RF model with the training data and made predictions for the test compounds. We also carried out a 5-fold cross validation using only the training data. We calculated SDC and STD for each compound. These calculations followed the same approach we used for the MP data (see Materials and Methods). After sorting the predictions based on SDC and STD values separately, we successively removed a portion of predictions with the lowest SDC and the highest STD values and calculated the RMSE for the remaining predictions. Figure 6 shows the resulting RMSEs as a function of the percentage of remaining predictions for the four data sets. The graphs show that (1) for 5-fold cross validation, SDC and STD performed similarly, with STD performing slightly better for all data sets; (2) for test sets with out-of-domain compounds, SDC was superior to STD, given that removal of the lowest SDC-ranked predictions led to a steeper reduction in the RMSE of the remaining predictions; (3) for all four data sets, the RMSEs of cross validation were lower than the RMSEs of the test sets with out-of-domain compounds, indicating that the model performance estimate derived from cross validation is an overly optimistic estimate for predicting future compounds, given that chemical research tends to explore new chemical spaces, generating previously unseen (and therefore out-of-domain) chemical structures.

The results of the additional data sets confirmed that our finding of SDC being more efficient than STD in identifying out-of-domain compounds is not restricted to MP data. Because SDC does not rely on building an ensemble of QSAR models, it can be easily deployed with all machine-learning methods, including the most popular of them all—deep neural networks.

We demonstrated the benefits of SDC over STD by making a portion of the compounds with extreme activities outside of the training-set domain. Our calculations showed that in cross validation with compounds randomly separated into training and test sets, the benefits of SDC were inconsistent. We believe that two main factors obscure the benefits of SDC over STD in cross validation. First, in a cross-validation study, compounds are randomly separated into training and test sets. In chemical research, especially in drug discovery, compounds tend to be synthesized in chemical series (an active compound leads to the synthesis of many structurally similar compounds). For this reason, the likelihood of a test compound having close near neighbors in a training set is high in cross validation, and hence, fewer compounds will be outside of the training-set domain. The large prediction errors of a small number of out-of-domain compounds will be obscured by the smaller errors of a large number of within-domain test compounds when calculating the root mean squared error.

Second, as Figure 5 shows, the sample distributions of all molecular activity data sets are highly uneven, with most compounds distributed around the most probable activity of each data set. This study also showed that for all out-of-domain compounds—those with little to no structural similarity to the training molecules—the model predictions are nearly constant and close to the most probable activity of the training set. Thus, for a significant fraction of out-of-domain compounds, the prediction errors are expected to be small. However, this is not because the predictions are reliable; rather, it is because a model tends to give the most probable activity of a training set as its predicted activity of an out-of-domain compound, and the most probable activity of a training set is also the most probable activity of any compound. This is why we chose to move compounds with extreme activities out of the training-set domain to demonstrate the benefits of the SDC metric.

Equation 1 indicates that the DA of a model depends on the training set. The larger the training set, the more likely it is to contain more structurally diverse compounds. Consequently, the DA is larger. Because the model trained on the Enamine data set (all MPs > 0 °C) gave poor predictions for compounds with MPs of less than 0 °C, we sought to assess the impact of training set size on prediction reliability for compounds with MPs of less than 0 °C, given that the MPs of all training compounds are above 0 °C. To this end, we first made predictions, using the RF model trained on the Enamine data set, for compounds in the Bradley data set with MPs of less than 0 °C. We plotted the prediction errors for these compounds against SDC (Figure 7A). The SDC values of most of the 707 compounds were nearly zero, i.e., outside of the DA. A very small proportion of the SDC values were nonzero, with the highest being 4.51.

We then removed compounds of the OCHEM data set for which the MPs were 0 °C or less (1532 compounds) and combined the remaining 20 351 compounds with the Enamine data set to serve as a new expanded training set with all MPs greater than 0 °C. The expanded training set was nearly double the size of the Enamine data set. We trained an RF model on this expanded data set and made predictions for the same 707 Bradley compounds. We plotted the prediction errors against the SDC values calculated from the expanded training set (Figure 7B), which again shows that most compounds are still outside of the DA with SDC values near zero. However, a comparison of Figure 7A and B revealed that considerably more compounds are within the DA of the expanded training set, consistent with expectations based on eq 1. More importantly, the prediction errors for many of the compounds pulled into the DA by the expanded training set were reduced relative to the prediction errors by the model trained on the smaller Enamine training set. Thus, although the expanded training set did not contain any compounds with MPs of less than 0 °C, it still expanded the DA for compounds with MPs of less than 0 °C and reduced prediction errors for some compounds.

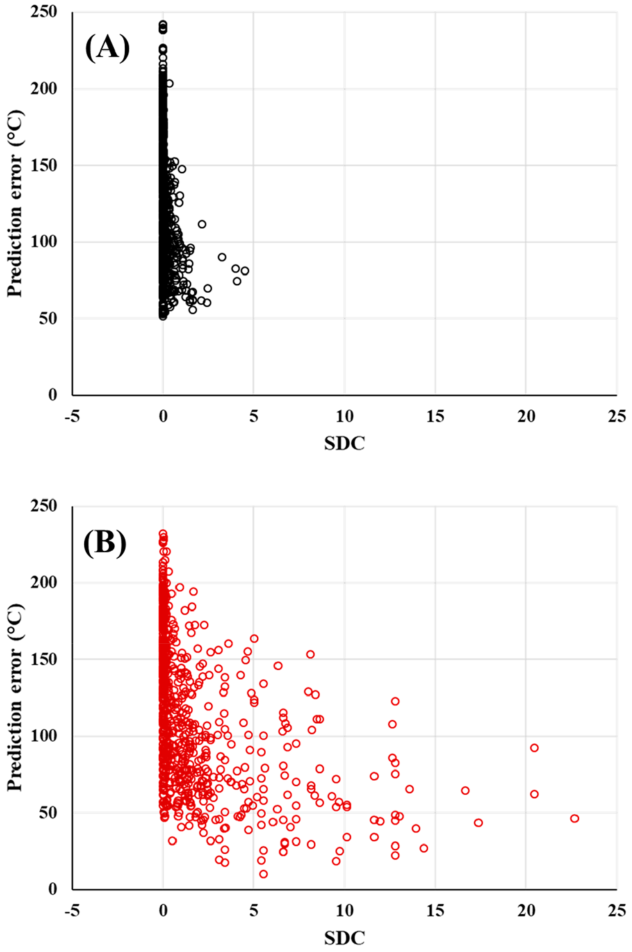


Figure 7. (A) Error in MP predicted by the RF model trained on the Enamine data set (22 404 compounds) for Bradley test set compounds with a MP of less than 0 °C, plotted as a function of SDC, and (B) error in MP predicted by a RF model trained on Enamine and OCHEM data sets (consisting of 42 755 compounds with a MP of greater than 0 °C) for the same test set compounds as those in A, plotted as a function of SDC.

■ SUMMARY

In this study, we demonstrated that SDC, a metric we recently developed for assessing the reliability of QSAR model predictions, identified large prediction errors in melting point data sets for which ensemble variance failed. Our analysis indicated that the failure to identify predictions for out-of-domain compounds is responsible for the failure of ensemble variance for melting point data. Interestingly, ensemble variance performed marginally better than SDC for within-domain compounds. To ensure the generality of the findings, we used four additional molecular property/activity data sets. For each data set, we divided the compounds into a training set and a test set and ensured that compounds in the test set included some outside of the training domain and with extreme activity levels. Calculations on these data sets confirmed that while SDC performed similarly to ensemble variance for within-domain compounds, it considerably outperformed ensemble variance in identifying predictions of out-of-domain compounds. Considering that SDC does not rely on an ensemble of QSAR models, it is easier to deploy with any machine-learning method and ideal for deep learning, which is perhaps the most powerful and popular method today.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00597.

> RECORDIDs of compounds in OCHEM data set removed to make 91 Bradley compounds outside of the resulting OCHEM data set and the training and test sets of the lipophilicity, rat oral toxicity, leukemia cell growth inhibition, and solubility data sets (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: rliu@bhsai.org.

*E-mail: sven.a.wallqvist.civ@mail.mil.

ORCID

Ruifeng Liu: 0000-0001-7582-9217

Funding

The research was supported by the U.S. Army Medical Research and Materiel Command (Ft. Detrick, MD) as part of the U.S. Army's Network Science Initiative, and by the Defense Threat Reduction Agency (Grant No. CBCall14-CBS-05-2-0007).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Dr. Tatsuya Oyama in editing the manuscript. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES

(1) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability Domain for QSAR Models: Where Theory Meets Reality. *Int. J. QSAR* **2016**, *1*, 45−63.

(2) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure−Activity Relationships. *ATLA* **2005**, *33*, 1−19.

(3) 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.

(4) Toplak, M.; Mocnik, R.; Polajnar, M.; Bosnic, Z.; Carlsson, L.; Hasselgren, C.; Demsar, J.; Boyer, S.; Zupan, B.; Stalring, J. Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. *J. Chem. Inf. Model.* **2014**, *54*, 431−441.

(5) 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.

(6) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.

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

(8) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Muller, K. R.; Xi, L.; Liu, H.; Yao, X.; Oberg, T.; Hormozdiari, F.; Dao, P.; 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.

(9) Tetko, I. V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A. E.; Charochkina, L.; Asiri, A. M. How Accurately Can We Predict the Melting Points of Drug-like Compounds? *J. Chem. Inf. Model.* **2014**, *54*, 3320−3329.

(10) Liu, R.; Glover, K. P.; Feasel, M. G.; Wallqvist, A. General Approach to Estimate Error Bars for Quantitative Structure-Activity Relationship Predictions of Molecular Activity. *J. Chem. Inf. Model.* **2018**, *58*, 1561−1575.

(11) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, II; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533−554.

(12) Bergstrom, C. A.; Norinder, U.; Luthman, K.; Artursson, P. Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177−1185.

(13) Bradley, J. C.; Lang, A. S. I. D.; Williams, A. J.; Curtin, E. ONS Open Melting Point Collection. *Nature Precedings* **2011**, DOI: 10.1038/npre.2011.6229.1 (accessed July 7, 2018).

(14) Enamine Ltd. http://www.enamine.net (accessed July 7, 2018).

(15) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

(16) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

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

(18) Liu, R.; Zhou, D. Using Molecular Fingerprint as Descriptors in the QSPR Study of Lipophilicity. *J. Chem. Inf. Model.* **2008**, *48*, 542−549.

(19) Zhou, D.; Alelyunas, Y.; Liu, R. Scores of Extended Connectivity Fingerprint as Descriptors in QSPR Study of Melting Point and Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *48*, 981−987.

(20) Liu, R.; Wang, H.; Wallqvist, A. Dissecting Machine-Learning Prediction of Molecular Activity: The Limits of Learning. *J. Chem. Inf. Model.* **2018**, Submitted.