JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 1561−1575

pubs.acs.org/jcim

General Approach to Estimate Error Bars for Quantitative Structure−Activity Relationship Predictions of Molecular Activity

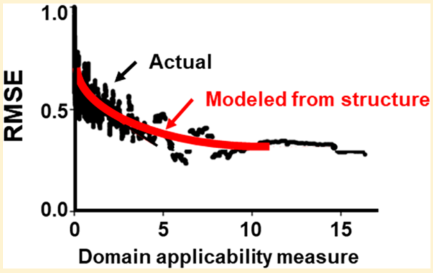
Ruifeng Liu,*[†] Kyle P. Glover,[‡] Michael G. Feasel,[§] and Anders Wallqvist*[†]

[†]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

[‡]Defense Threat Reduction Agency, Aberdeen Proving Ground, Maryland 21010, United States

[§]U.S. Army—Edgewood Chemical Biological Center, Operational Toxicology, Aberdeen Proving Ground, Maryland 21010, United States

Supporting Information



ABSTRACT: Key requirements for quantitative structure−activity relationship (QSAR) models to gain acceptance by regulatory authorities include a defined domain of applicability (DA) and appropriate measures of goodness-of-fit, robustness, and predictivity. Hence, many DA metrics have been developed over the past two decades. The most intuitive are perhaps distance-to-model metrics, which are most commonly defined in terms of the mean distance between a molecule and its *k* nearest training samples. Detailed evaluations have shown that the variance of predictions by an ensemble of QSAR models may serve as a DA metric and can outperform distance-to-model metrics. Intriguingly, the performance of ensemble variance metric has led researchers to conclude that the error of predicting a new molecule does not depend on the input descriptors or machine-learning methods but on its distance to the training molecules. This implies that the distance to training samples may serve as the basis for developing a high-performance DA metric. In this article, we introduce a new Tanimoto distance-based DA metric called the sum of distance-weighted contributions (SDC), which takes into account contributions from all molecules in a training set. Using four acute chemical toxicity data sets of varying sizes and four other molecular property data sets, we demonstrate that SDC correlates well with the prediction error for all data sets regardless of the machine-learning methods and molecular descriptors used to build the QSAR models. Using the acute toxicity data sets, we compared the distribution of prediction errors with respect to SDC, the mean distance to *k*-nearest training samples, and the variance of random forest predictions. The results showed that the correlation with the prediction error was highest for SDC. We also demonstrate that SDC allows for the development of robust root mean squared error (RMSE) models and makes it possible to not only give a QSAR prediction but also provide an individual RMSE estimate for each molecule. Because SDC does not depend on a specific machine-learning method, it represents a canonical measure that can be widely used to estimate individual molecule prediction errors for any machine-learning method.

INTRODUCTION

Quantitative structure−activity relationship (QSAR) modeling was initially introduced more than 50 years ago for predicting the chemical properties of congeneric compounds.[1] Its success in predicting congeneric chemical series encouraged applications of the technique to data sets consisting of compounds with increasingly diverse structures.[2] However, it became clear that QSAR model performance was not consistent across molecules, as it was typically better for compounds whose molecular structures were adequately represented by training samples.[3] In recent years, defining a model's domain of applicability (DA) has been an area of active research in QSAR modeling.[4−25] The goal is to provide not only QSAR predictions but also the degree of confidence in the predictions based on the relationship of the new molecules to the domain. This is important because most end users of QSAR models do not have direct knowledge of the structural information on the training molecules from which the models were derived. This is especially true for commercial software packages, whose training sets are usually proprietary and not disclosed to end users.

Although DAs can be defined in many ways, most can be grouped into two categories. One category defines a chemical space, where a QSAR prediction for a molecule is deemed reliable if the molecule is in the space but unreliable if it is not. The other category gives an estimate of the uncertainty of each QSAR prediction, with a smaller uncertainty indicating a more reliable prediction. We favor the latter approach because, regardless of how one defines chemical space, the prediction performance for molecules is more likely to deteriorate gradually as one moves from the center to the edges of the space. The difference in prediction performance between two structurally similar molecules that straddle a DA boundary is likely small. Thus, an uncertainty-based approach is advantageous because it requires neither a predefined boundary nor a leap of

Received: February 27, 2018
Published: June 27, 2018

DOI: 10.1021/acs.jcim.8b00114
J. Chem. Inf. Model. 2018, 58, 1561−1575

faith in the prediction reliability for molecules sitting across the boundary.

Many methods have been devised to gauge the reliability of QSAR predictions. The most intuitive is perhaps the distance-to-model metric, which is based on the distance (or similarity) between a new molecule and the training samples. However, counterintuitively, almost all variants of this type of metric have been defined as the average distance of a new molecule to the *k* nearest training samples. This definition essentially assumes that the presence or absence of other training samples has no effect on model performance.[10,23] Sun et al. reported that a significant fraction of reasonably good predictions might fall outside an applicability domain defined by the average distance to *k* nearest training samples.[26] In a comprehensive study evaluating the performance of many DA metrics, Tetko et al. found that the variance of predictions of an ensemble of QSAR models could serve as a DA metric and it outperformed distance-to-model metrics.[10,16] Intriguingly, they observed that the ensemble-variance metric always ranked the molecules in the same order, which led them to conclude that the error of predicting a new molecule did not depend on the descriptors or machine learning methods used, but on their similarity to the training set molecules.[10] Sheridan found that in addition to the variance of all random forest (RF) decision tree predictions, the predicted value itself was correlated with the error of predictions and could serve as a DA metric.[19] Interestingly, he and his colleagues also found that similarity to training samples is a good discriminator for prediction accuracy,[3] and in a comprehensive study using 15 drug discovery data sets, Sheridan found that the relative importance of distance-to-model and ensemble-variance metrics for estimating QSAR prediction errors depends on training set diversity,[23] which is not always transparent to end users of QSAR models.[20]

The observation that the ensemble-variance metric outperformed distance-to-model metric appears counterintuitive to the conclusion that the error of predicting a new molecule depends on distance to the training samples only. In our opinion, the distance-to-model metrics evaluated in previous studies have two flaws: (1) they assume only a limited number of nearest training samples contribute to prediction accuracy and (2) do not weight the contributions of the training samples. In this work, we introduce a new DA metric that considers the contributions of every training sample, each weighted by its distance to the molecule for which a QSAR prediction is made. We demonstrate that this metric correlates more strongly with prediction error than the distance-to-model or ensemble-variance metric and therefore has the potential for broad applications in modeling QSAR prediction errors.

METHODS AND MATERIALS

Sum of Distance-Weighted Contribution (SDC) Metric.

Our hypothesis is that all training molecules contribute to the prediction reliability of a given molecule, but they do not contribute equally. Instead, as dictated by the similar-structure similar-activity principle, structurally similar neighbors should contribute more than structurally dissimilar neighbors. A convenient metric of similarity between two molecules is the Tanimoto similarity,[27] which is defined as

$$\mathrm{TS} = \frac{C}{A + B - C} \qquad (1)$$

Here, *A* and *B* are the counts of unique molecular fingerprint features in molecules *a* and *b*, respectively, and *C* is the count of fingerprint features common to both molecules. According to eq 1, the value of TS ranges from 0, for two molecules not

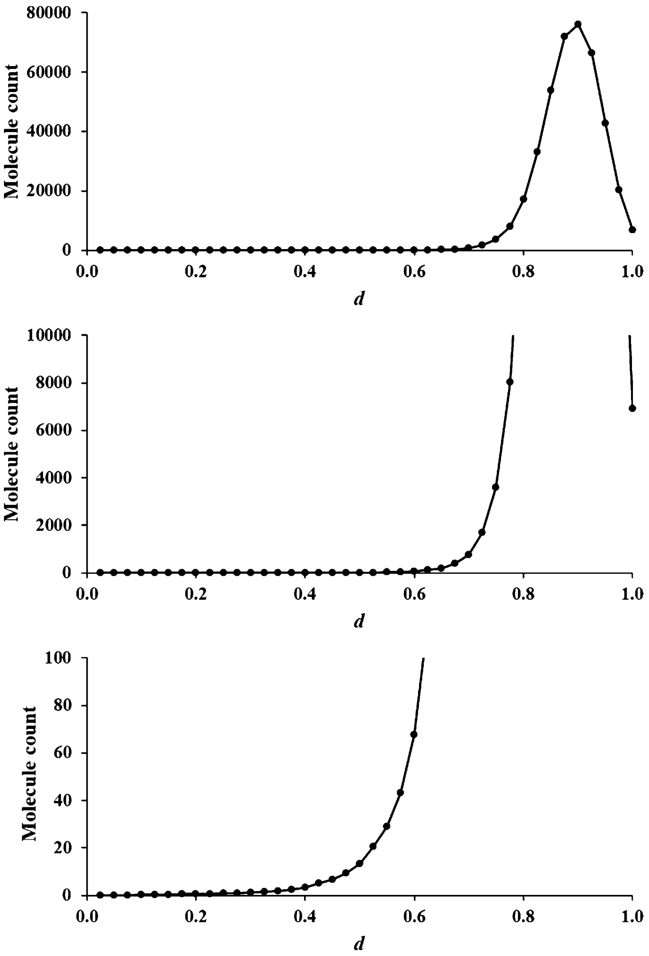


Figure 1. Distribution of the average number of molecules in the STITCH database that are within a specific Tanimoto distance (*d*) range to a given molecule. The three panels are of the same data but plotted on differently scaled *y*-axes. They show that on average, the number of neighbors close to a given molecule scales exponentially with the Tanimoto distance (*d*) between the molecule and the near neighbors.

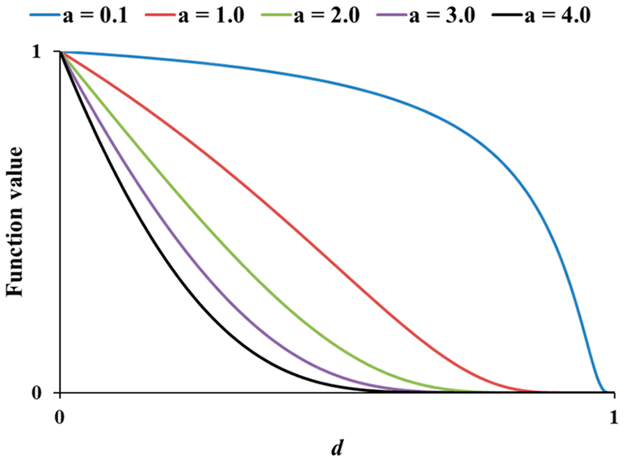


Figure 2. Dependence of $\exp\left(-\frac{ad}{1-d}\right)$ on *a*. The value of this function decays faster the higher the value of *a* (i.e., the larger the penalty for dissimilar molecular structures).

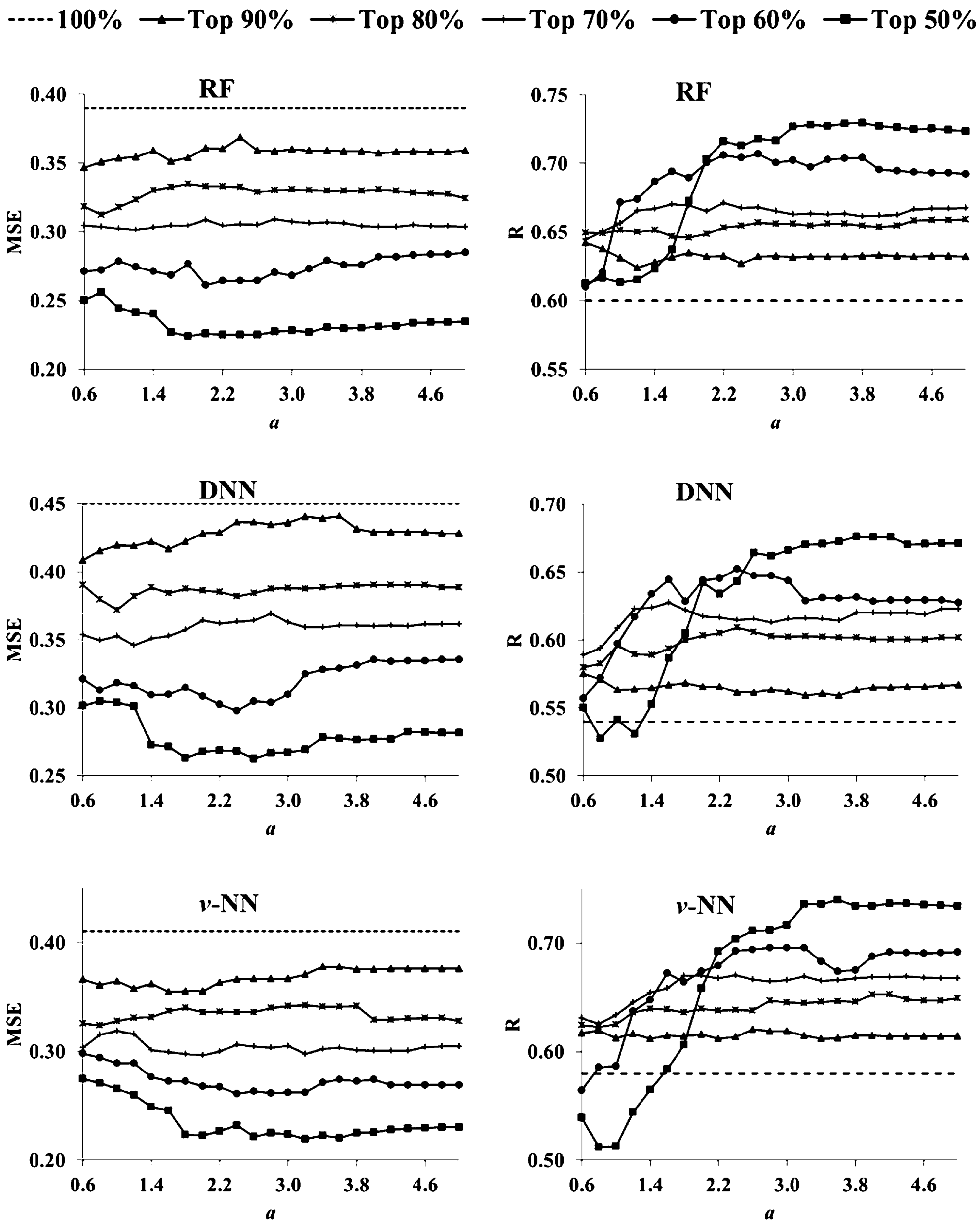


Figure 3. MSE and correlation coefficient (R) between predicted and experimental log(LD50) values of the compounds remaining in the rabbit skin toxicity data set after successive removal of 10% of the lowest-ranked compounds based on SDC calculated with different *a* values. The dashed lines represent the MSE or *R* values of the whole data set.

sharing any common structural features, to 1, for two molecules sharing all structural features. A corresponding metric of the distance between two molecules can be defined as

$$d = 1 - TS \tag{2}$$

In the remainder of the paper, we refer to *d* or distance as the Tanimoto distance.

To get a general idea of the distribution of structurally similar molecules in a given data set, we randomly selected 500 compounds from the chemical collection of the Search Tool for Interacting Chemicals (STITCH) database (version 5.0).[28] We also randomly selected another 100 000 chemicals from the same database and calculated distances between each of the 500 molecules and each of the 100 000 molecules, using extended connectivity fingerprints with a diameter of four chemical bonds (ECFP_4).[29] To generate a distribution of the number of structurally similar compounds with respect to the distance between the molecules, we divided the distance scale into 40 equal-sized bins and counted the number of molecules in each bin. We divided the counts by 500 to derive an average number of neighbors a molecule has in each bin. Figure 1 shows that whereas the number of molecules with high structural similarity is extremely small, the number of molecules increases exponentially with *d*. Note that a *d* of 0.9 or higher indicates virtually no structural similarity between two molecules.

The similar structure–similar activity principle posits that structurally similar compounds contribute more to the reliability of a prediction. However, Figure 1 shows that for a given molecule, the number of compounds highly similar to it is very small. Hence, the contribution of a training molecule to the prediction reliability of a given molecule should be weighted down exponentially by the distance between the two molecules. An estimate of the contribution of all training molecules to the prediction reliability for a given molecule can then be written as

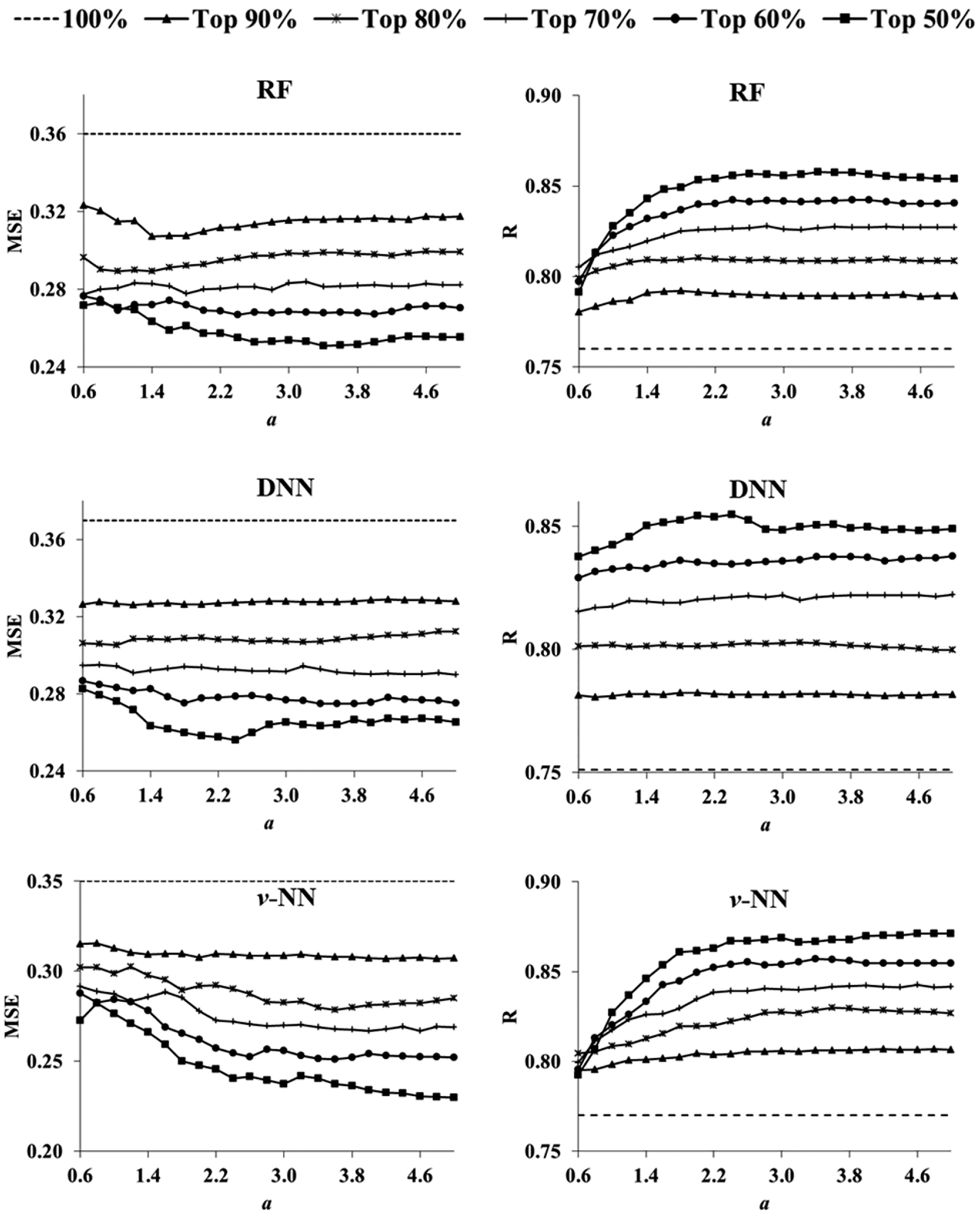


Figure 4. MSE and R between predicted and experimental log(LD50) values of the compounds remaining in the rat oral toxicity data set after successive removal of 10% of the lowest-ranked compounds based on SDC calculated with different a values. The dashed lines represent the MSE or R values of the whole data set.

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

where SDC is our DA metric defined as the sum of the distance-weighted contribution of all training molecules; d_i is the distance between the ith molecule in the model training set and the test molecule; and a is an adjustable parameter that modulates the distance penalty for the contribution of a training set molecule. Figure 2 shows $e^{-ad/(1-d)}$ as a function of d for different values of a. It shows that a larger a value gives a higher penalty to a distant training molecule.

Data Sets Used to Evaluate Performance of the SDC Metric. To evaluate the performance of SDC, we developed QSAR models for four acute chemical toxicity data sets of varying sizes. They are rabbit skin, rat oral, mouse oral, and mouse intraperitoneal toxicity data sets consisting of 1745, 10 363, 21 776, and 29 476 structurally unique compounds, respectively. Each compound in the data sets has an experimentally measured LD50 value in mg/kg. For this study, we converted the LD50 values to log(mmol/kg). We downloaded all of the toxicity data from the Leadscope Toxicity Database (http://www.leadscope.com/toxicity_database/), which were curated from the Registry of Toxic Effects on Chemical Substances.

To demonstrate the general applicability of the SDC metric to a broad range of molecular properties, we also built QSAR models and examined the prediction error distribution with respect to SDC for molecular lipophilicity (10 178 compounds with measured log P values), melting point (4444 compounds with measured melting point [MP] values), solubility (1,033 compounds with logarithm of molecular solubility [log S]), and a data set of 756 compounds with measured IC50 for inhibiting the activity of dihydrofolate reductase (DHFR). We used the log P data set in an example Pipeline Pilot protocol

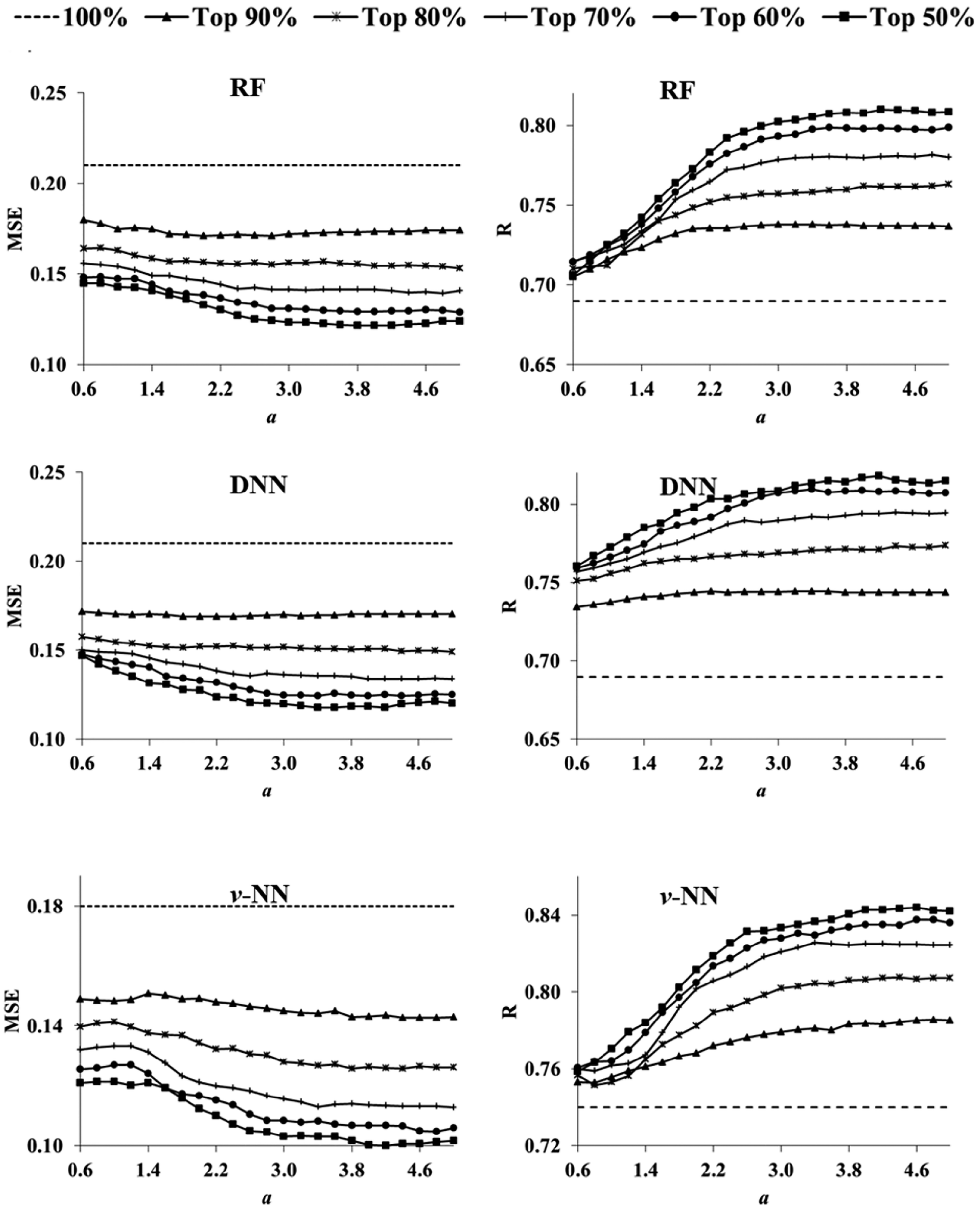


Figure 5. MSE and *R* between predicted and experimental log(LD50) values of the compounds remaining in the mouse oral toxicity data set after successive removal of 10% of the lowest-ranked compounds based on SDC calculated with different *a* values. The dashed lines represent the MSE or *R* values of the whole data set.

for building a log *P* prediction model. The MP data set was provided by Karthikeyan et al. in the Supporting Information of their paper investigating QSAR prediction of melting points.[30] The log *S* data set was provided by Huuskonen in the Supporting Information of his paper investigating QSAR prediction of aqueous solubilities,[31] and the DHFR data set was from the Supporting Information of the paper of Sutherland et al.[32]

Molecular Descriptors. In this study, we used three machine-learning methods to build our QSAR models for the acute toxicity data sets: deep neural network (DNN), RF, and variable nearest neighbor (*v*-NN) methods. We used ECFP_4 fingerprint features as the input descriptors for the four toxicity data sets. To accelerate Tanimoto distance calculations for the *v*-NN method, we folded the raw fingerprints into a fixed length of 2048 bits. Our calculations indicated that fingerprint folding to 2048 bits has a negligible impact because much longer or shorter fingerprints produce similar Tanimoto distances.

For the RF models, we used ECFP_4 fingerprints as input features without folding because the RF approach can handle a large number of input descriptors. To build each decision tree, RF models use only a subset of the input descriptors. These descriptors are selected to give optimal splits of the training samples.

Because of the large number of network weights to be determined in a DNN model, the number of input descriptors has a marked impact on computational cost. To contain this cost and ensure comparability with the *v*-NN approach, we used a total of 2048 ECFP_4 fingerprint features as input descriptors for all data sets in performing DNN calculations. These 2048 fingerprint features were not derived from folding the fingerprints but selected based on the frequencies of features in the data sets. For each data set, we selected the 2048 fingerprint features according to the following procedure:

(1) Identify all unique fingerprint features present in a data set;

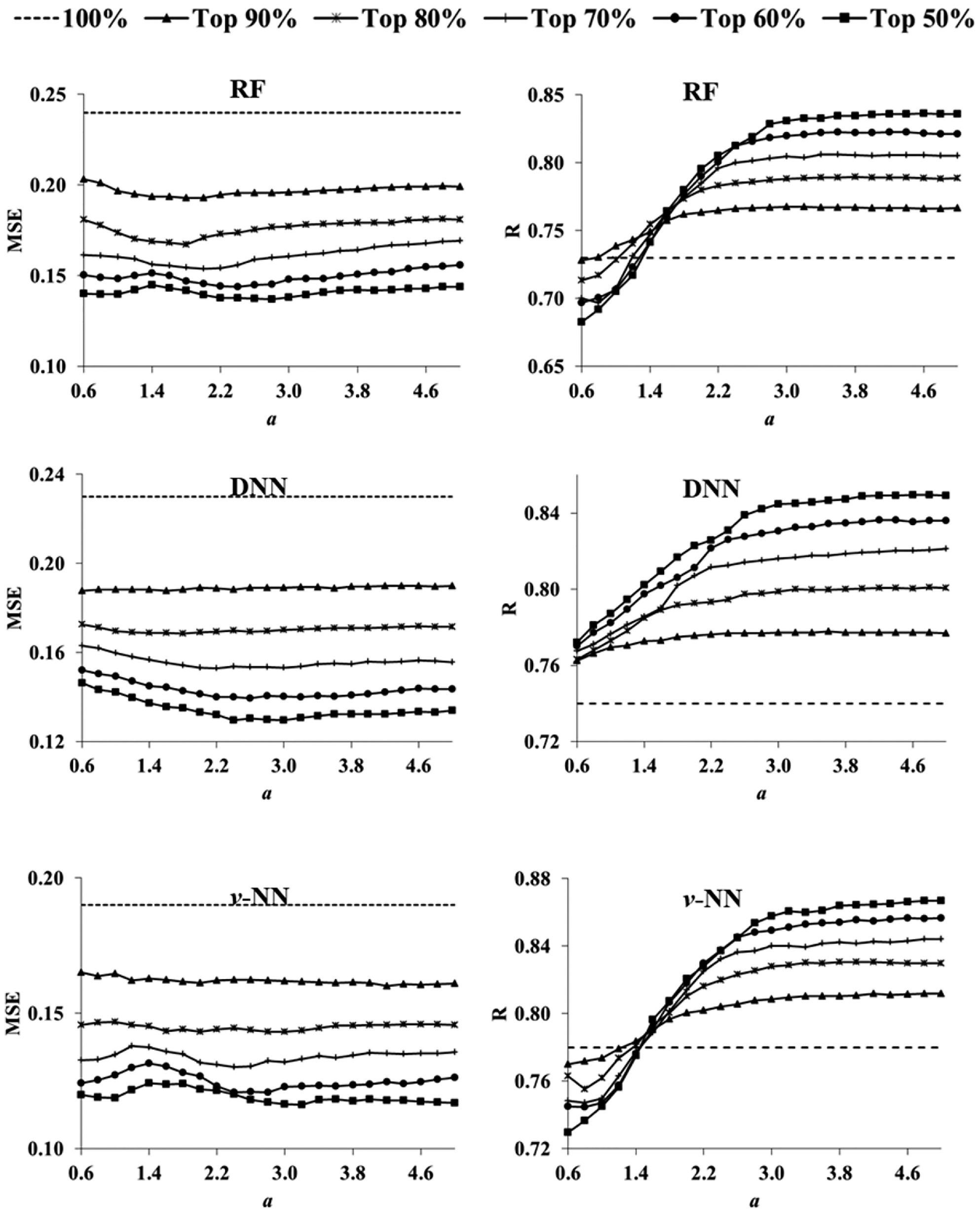


Figure 6. MSE and *R* between predicted and experimental log(LD50) values of the compounds remaining in the mouse intraperitoneal toxicity data set after successive removal of 10% of the lowest-ranked compounds based on SDC calculated with different *a* values. The dashed lines represent the MSE or *R* values of the whole data set.

(2) Calculate the frequency of each fingerprint feature appearing in all molecules in the data set;
(3) Select the fingerprint features appearing in 50% of the molecules and those closest to 50% of the molecules, until the total number of selected features reaches 2048. This selection process excludes the least important fingerprints because it deselects fingerprint features that appear in all or nearly none of the molecules.

Recent studies suggest that circular ECFP fingerprints are particularly suitable for deep learning of molecular properties because training DNNs to learn a representation of molecular structures directly from a graph representation led to learned features that were conceptually similar to circular fingerprints.[33]

To demonstrate the broad applicability of the SDC metric to QSAR models built on different machine-learning methods and molecular descriptors, we used a different set of molecular descriptors for each of the other four data sets. For the log *P* data set, we used the counts of 120 molecular fragments of the *A* log *P* model[34] as the input descriptors. For the MP data set, we used 202 conventional molecular descriptors provided by Karthikeyan et al.,[30] which included almost all molecular topological indices, counts of various atom types, all types of projected molecular surface areas, molecular volume, and quantum mechanical descriptors such as dipole moment, ionization potential, HOMO, and LUMO derived from semiempirical quantum mechanical calculations. We used estate_keys[35] as molecular descriptors for the log *S* data set and ECFP_4 fingerprint features as descriptors for the DHFR data set.

Details of Machine-Learning Approaches. *DNN.* To develop DNN prediction models, we used the open source Python library Keras (https://keras.io/) on top of Theano backend. We used mean squared error as the loss function for regression and probed the impact of multiple parameters,

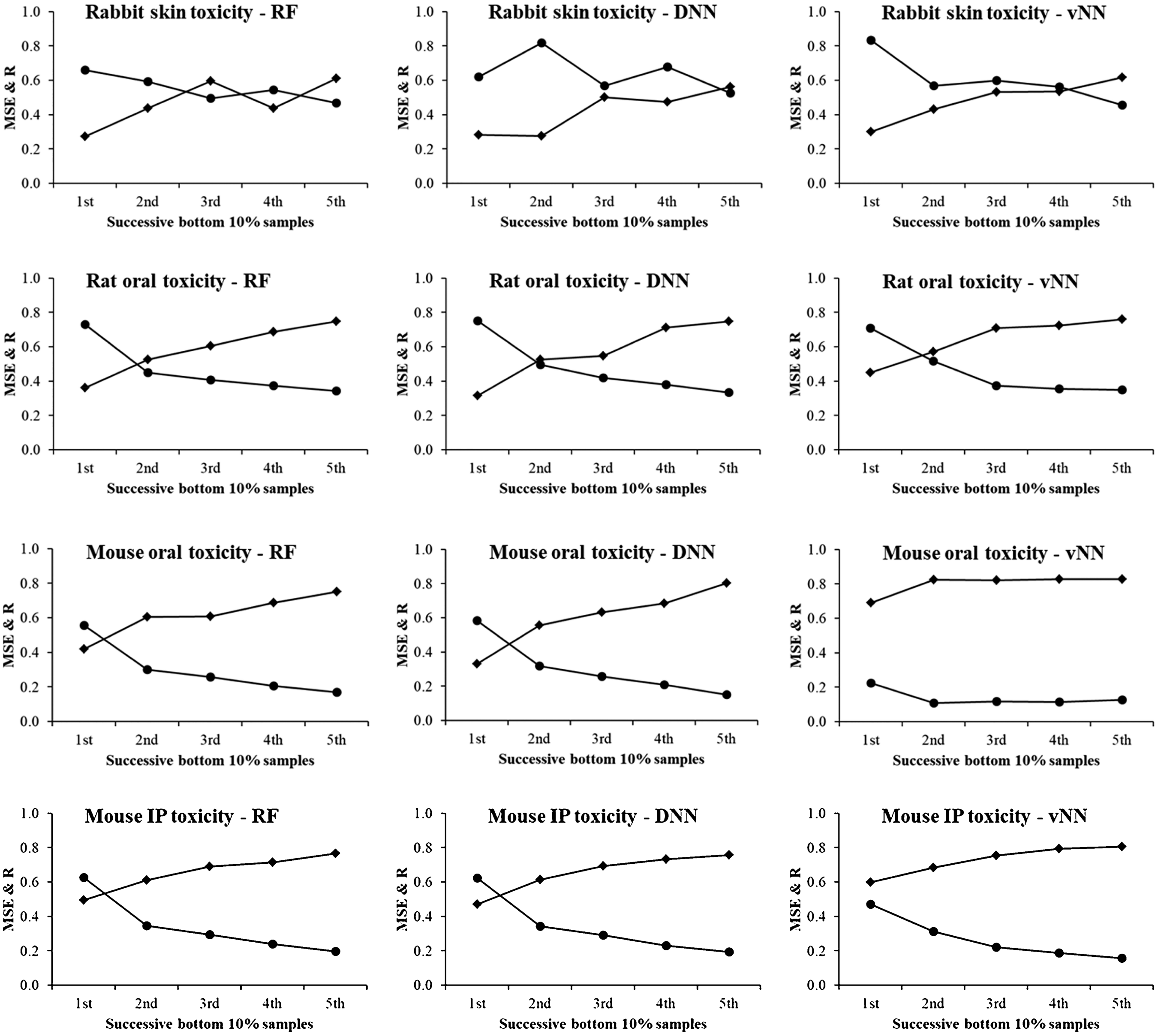


Figure 7. MSE and R between the predicted and experimental log(LD50) values of successive 10% of compounds ranked lowest in terms of SDC for the four acute toxicity data sets. The plots show that the MSE and R values are worst for the 10% of compounds with the lowest SDC-rankings but improve gradually for each successive 10% of lowest SDC-ranked compounds. The only exception is the pattern for rabbit skin toxicity, which is likely due to the small number of compounds in the data set.

dropout rates, optimizers, and initialization methods. Most default parameters in Keras performed reasonably well. Ultimately, we built all fully connected feed-forward multilayer neural networks using the ReLU activation function for the input and hidden layers, the Adam optimizer, a kernel initializer with a normal distribution, and a dropout rate of 30% on all input and hidden layers. For each data set, we performed a large number of 10-fold cross validation calculations to examine the performance of different network architectures, i.e., the number of hidden layers and the number of neurons in each hidden layer. We found that a fully connected feed-forward network consisting of 3 hidden layers with 300, 300, and 30 neurons in the three layers works reasonably well for the 4 toxicity data sets. Because there are 2048 structural features as inputs and a single neuron in the output layer generating the predicted log(LD50) for each molecule, we used 2048:300:300:30:1 to represent the neural net architecture. The total number of model parameters (weights of the connections between the neurons) was 713 430.

RF. To develop RF models, we used the Pipeline Pilot implementation Forest of Random Trees (http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/). The RF model for each data set consisted of 500 decision trees. The maximum tree depth was 50, and a third of all molecular descriptors were tested as split criteria within each tree. These and other parameters were set to the default parameters of the RF module in Pipeline Pilot. The default parameters performed reasonably well in most test scenarios. Therefore, we used them to develop RF models for all of the data sets studied here.

v-NN. This method is based on the principle that similar structures have similar activities. It gives a prediction y for a query compound as a distance-weighted average of all nearest neighbors in the training set,

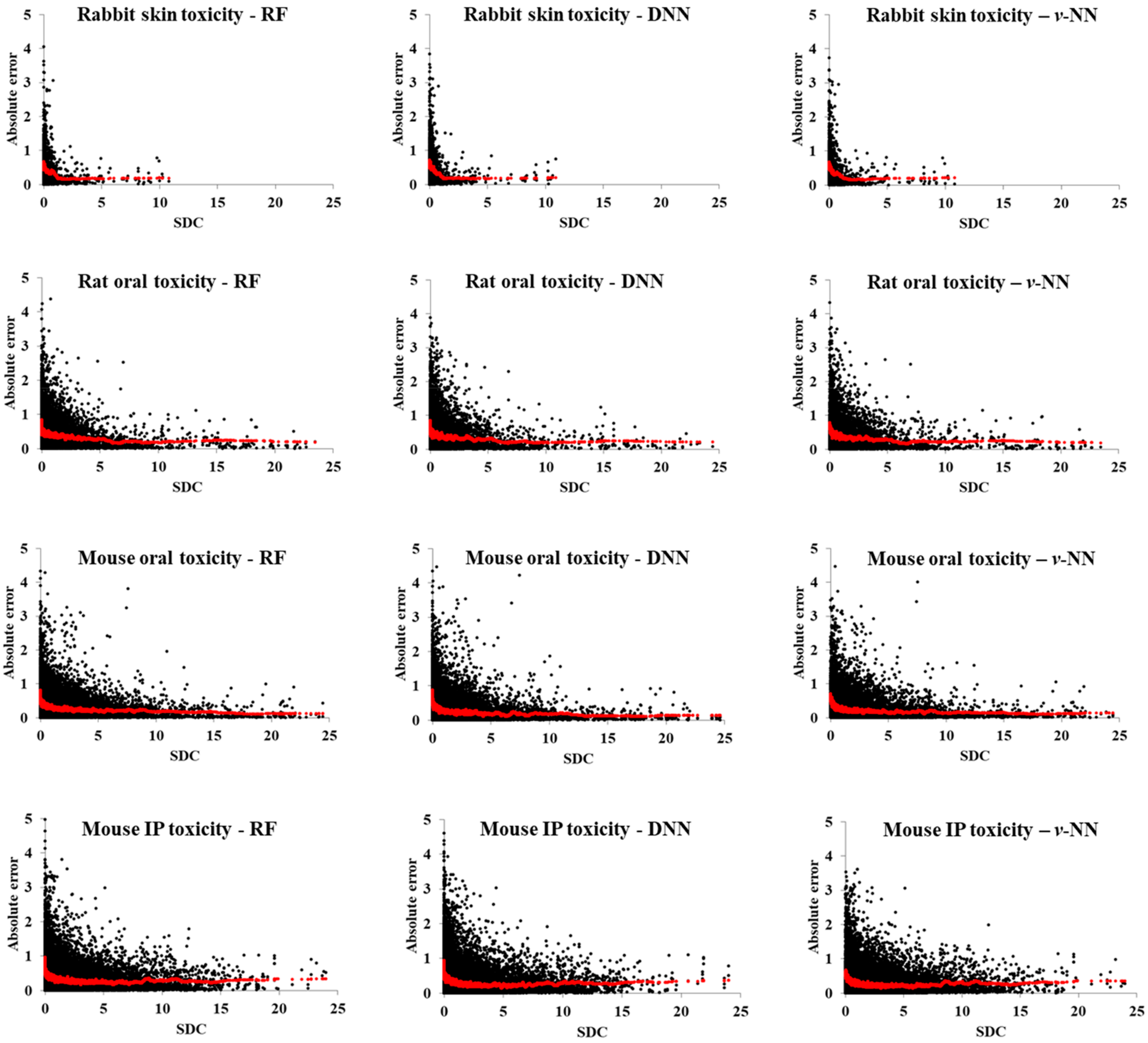


Figure 8. Absolute prediction error plotted against SDC for the four acute toxicity data sets. Each black dot represents a molecule with its absolute prediction error on the *y*-axis and its SDC value on the *x*-axis. The moving averages of the prediction errors calculated with a 100-compound window are displayed in red. The plots show a clear reduction of prediction error with increasing SDC value, irrespective of the data set and machine-learning method used. SDC was calculated using eq 3 with *a* set to 3.

$$y = \frac{\sum_{i=1}^{v} y_i e^{-\left(\frac{d_i}{h}\right)^2}}{\sum_{i=1}^{v} e^{-\left(\frac{d_i}{h}\right)^2}} \qquad (4)$$

In this equation, y_i is the toxicity of the *i*th nearest neighbor in the training set, d_i is the distance between the *i*th nearest neighbor and the molecule for which a *ν*-NN model is making a prediction, *h* is a smoothing factor that modulates the distance penalty, and *ν* is the count of all nearest neighbors in the training set that satisfy the condition $d_i \leq d_0$, where d_0 is a distance threshold that ensures the validity of the similar structure−similar activity principle when the distance between two molecules satisfies the condition. In the *ν*-NN approach, d_0 and *h* are the only model parameters to be determined from training data. To predict the property of a compound, a *ν*-NN model searches through currently available data to identify all qualified nearest neighbors and then uses eq 4 to make a prediction. For a given compound, a *ν*-NN model does not give a prediction if there are no qualified nearest neighbors. In our recent study of the same data sets, we found that a *ν*-NN model with a Tanimoto distance threshold of 0.60 and a smoothing factor of 0.30 performed reasonably, although predictions for a small percentage of compounds (5−25%, depending on data set size) could not be provided as they did not have qualified near neighbors in the training samples. In this study, we used d_0 = 0.60 and *h* = 0.30 for all *ν*-NN calculations.

To build QSAR models for the four nonacute toxicity data sets, we used different machine-learning methods. To develop log *P* and MP prediction models, we used the partial least-squares (PLS) method using the first 10 latent variables only. To develop a log *S* prediction model, we used support vector machines (SVM) with a radial kernel. To develop a DHFR

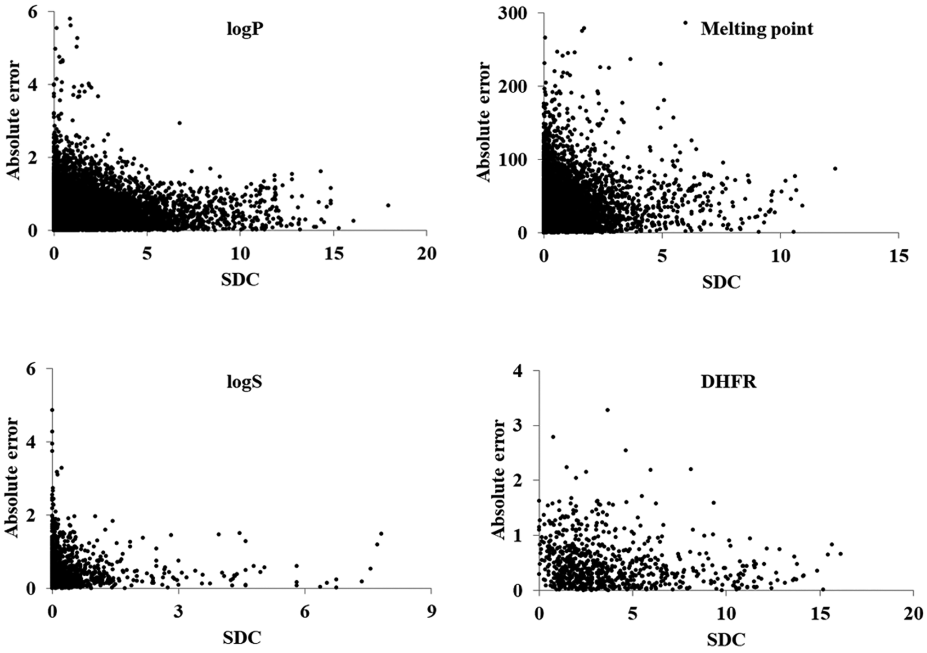


Figure 9. Absolute prediction error plotted against SDC for four additional data sets. The plots show a clear reduction of prediction error with increasing SDC value, irrespective of the data set and machine-learning method used. SDC was calculated using eq 3 with *a* set to 3.

inhibition model, we used RF with 500 decision trees. Because our objective here was to assess the correlation between the SDC and the prediction error, instead of developing the best prediction models for these properties, we did not optimize the hyperparameters of these methods for these data sets. Rather, we used the default parameters of the *R* package.

RESULTS AND DISCUSSION

Dependence of SDC Performance on *a*. For SDC to serve as a performance metric of QSAR predictions, one should be able to use it to identify unreliable or highly uncertain predictions. To assess SDC on this task, we performed 10-fold cross validation calculations for the four acute toxicity data sets, using the three machine-learning methods. With the DNN and RF methods, the models gave predictions for all compounds in each data set. With the *v*-NN method, however, the models gave predictions only for compounds having qualified neighbors within a Tanimoto distance of 0.6. The percentages of compounds for which *v*-NN models gave predictions depended on data set size, and were 75, 86, 94, and 95% for the rabbit skin, rat oral, mouse oral, and mouse intraperitoneal toxicity data sets, respectively. For each data set, we calculated the mean squared error (MSE) and correlation coefficient (*R*) between the predicted and experimental log(LD50) values.

To determine the optimal value of *a* in Equation 3, we calculated the SDC values for each compound by systematically varying *a* from 0.6 to 5.0 in increments of 0.2. We ranked the compounds in each data set based on their SDC values, threw out 10% of the compounds with the lowest SDC values, and recalculated MSE and *R* for the remaining compounds, iteratively repeating this process. Figures 3−6 show the resulting MSE and *R* values plotted against *a*. The MSE plots in al four figures show that SDC values were inversely related to prediction errors. As SDC increased, the prediction error decreased. After removal of the lowest 10% of SDC-valued samples, the MSE of the remaining compounds markedly decreased. Successive removal of 10% of the samples with the lowest SDC values resulted in a continuous decrease in MSE of the remaining samples. Similarly, successive removal of the lowest-ranked 10% of samples led to a steady increase in *R* for the remaining samples when *a* was ~3 or higher. At smaller *a* values, *R* values showed marked variability, especially for small data sets. Thus, based on the results of the four data sets, we set *a* to a value of 3 in eq 3 for the remainder of the study.

The improvement in MSE and *R* with successive removal of the lowest SDC-ranked compounds indicated that the prediction errors for the removed compounds were large. To confirm this, we calculated the MSE and *R* values for the successively removed compounds in each data set. Figure 7 shows that compounds with the lowest SDC values had the highest MSE and the lowest *R* values. With the exception of the smallest rabbit skin toxicity data set, the MSE decreased and the *R* value increased for the second to fifth 10% sample sets that were successively removed, indicating that SDC was correlated with prediction reliability.

Distribution of Prediction Errors. Figure 8 shows the distribution of absolute deviations between the predicted and experimental log(LD50) values of molecules in the four acute toxicity data sets. In this figure, each black dot represents a molecule. The vertical axis is the absolute error of prediction for a molecule, and the horizontal axis is the SDC value of the molecule. The moving averages of the prediction errors calculated with a 100-compound window are displayed in red. For all plots, the prediction error decayed exponentially with increasing SDC. For rabbit skin toxicity—the smallest data set—the SDC values of all compounds were relatively low, and most data points were close to the vertical axis at SDC = 0. With increasing data set size, the data points were increasingly right-shifted, consistent with the general observation that QSAR models developed from larger data sets are more reliable.

To ascertain that the correlation between prediction error and SDC is not limited to the acute toxicity data sets and is independent of machine-learning methods and molecular descriptors, we also performed 10-fold cross validation calculations for the four other data sets, using the machine-learning methods and molecular descriptors described in the Method and Materials Section. The resulting prediction error distributions (Figure 9) were similar to those in Figure 8, indicating that SDC

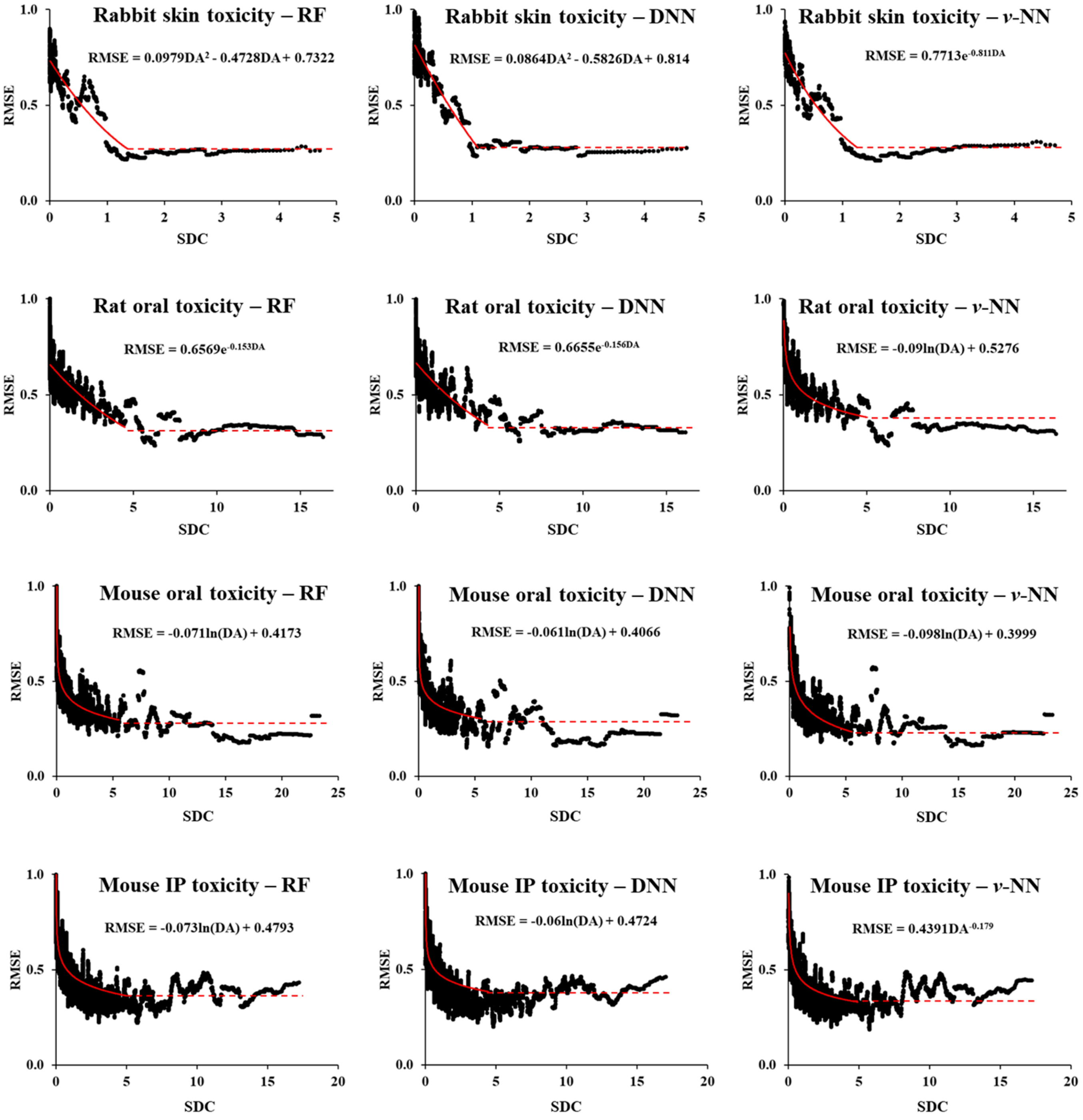


Figure 10. Moving average of the RMSE calculated with a 100-compound window plotted against the moving average of SDC for four acute toxicity data sets and three machine-learning methods. The equations shown in the plots were derived from the Trend Analysis function in Microsoft Excel for the data points in the exponential decay range (solid red line). All moving RMSE plots level off in the high SDC range, possibly because the inherent variability in the experimental data. SDC was calculated using eq 3 with a set to 3.

can serve as an intuitive parameter that is generally applicable for modeling QSAR prediction errors.

SDC-Based Estimate of Prediction Uncertainties. The plots in Figures 8 and 9 show that regardless of the machine-learning method used, model predictions for different compounds are not equally reliable. Although the developers of a model and users with detailed information on how the model was built may have some ideas about the type of molecules for which the model may give reliable predictions, most users may have neither access to the information nor the training or knowledge to estimate the reliability of model predictions. Thus, it is important to provide not only a model prediction but also an estimate of its uncertainty. This recommended practice is implemented in some platforms, e.g., the Online Chemical Modeling Environment (OCHEM).[36] The strong correlation between prediction errors and SDC values offers a straightforward and practical approach to estimate prediction uncertainty. To this end, we developed prediction models for the root mean squared error (RMSE) of acute toxicity predictions by the following procedure.

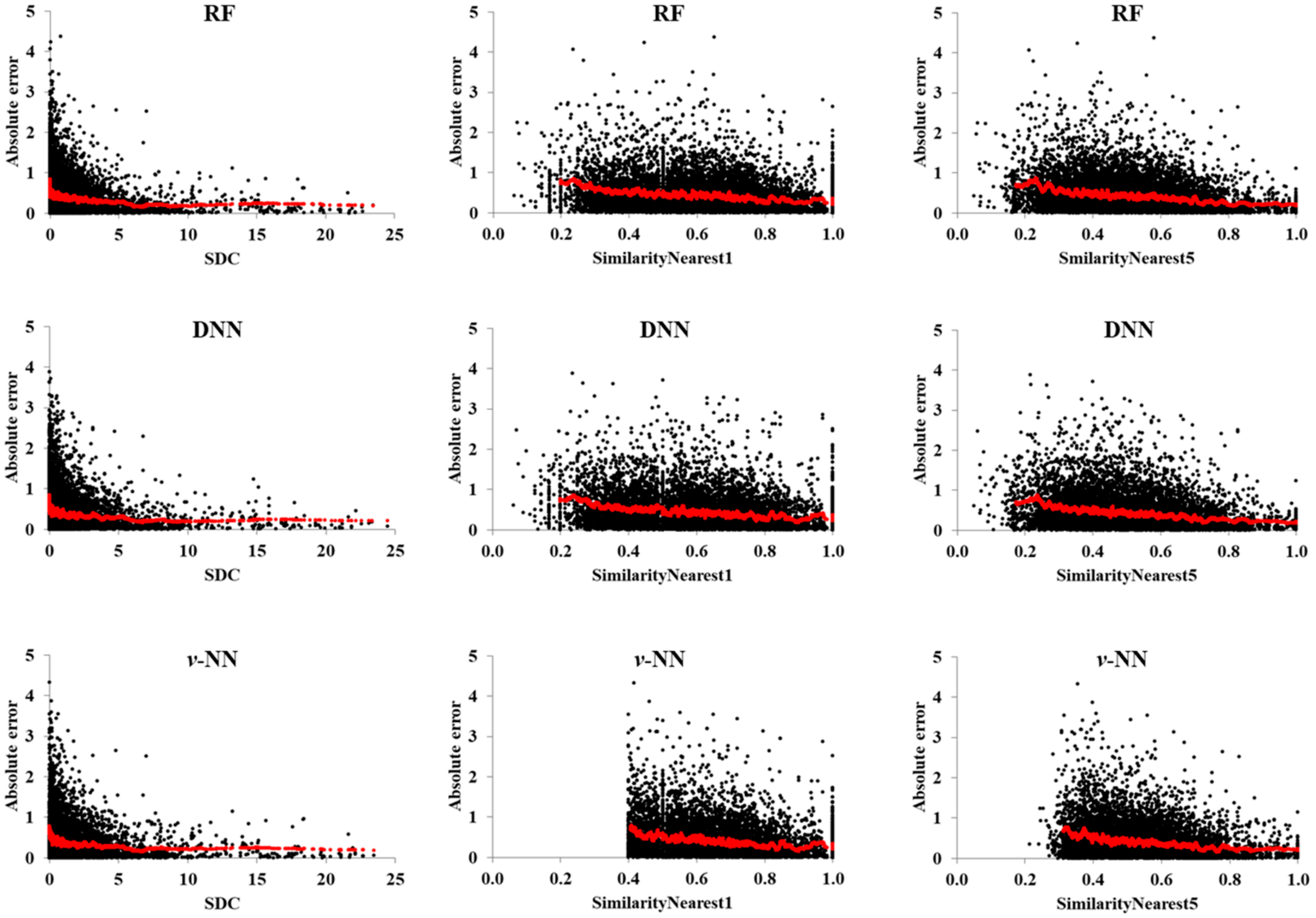


Figure 11. Prediction errors plotted against SDC, SimilarityNearest1, and SimilarityNearest5 for predictions of the rat oral toxicity data set made by RF, DNN, and ν-NN methods. The moving averages of prediction errors calculated with a 100-compound window are displayed in red. The plots show that SDC is most strongly correlated with the prediction error. Similar results obtained for the other data sets are detailed in the Supporting Information. SDC was calculated using eq 3 with a set to 3.

(1) For each data set, we sorted the compounds by their SDC values in increasing order. We then calculated moving averages of the SDC values with a 100-compound window. This was done by first taking the average SDC of the 100 compounds with the lowest SDCs and then replacing the first compound with the 101st and recalculating the average SDC. This was repeated until the compound with the highest SDC was included in calculating the average SDC. Similarly, we calculated moving averages of the RMSE between the predicted and experimental log(LD50) values, using the same 100-compound window.

(2) We plotted the resulting moving RMSE against the moving SDC for each of the four data sets in Figure 10, yielding a total of 12 RMSE graphs with the three machine-learning methods.

All RMSE plots showed a similar trend: the smaller the SDC, the greater the RMSE. In addition, the RMSE values decreased exponentially with increasing SDC. However, all RMSE values leveled off at higher SDC values, most likely because of the inherent uncertainty of the experimental measurements.

To derive a mathematical expression for the RMSE as a function of SDC, we fitted the declining portion of the moving RMSE data of each plot in Figure 10. After exploratory fitting of the data with an exponential, logarithmic, polynomial, or power function, we plotted the best fits (Figure 10, solid red curves, with equations shown in the upper right). Figure 10 also depicts the uncertainty limits for the machine-learning methods (dashed red lines), which are likely dictated by the uncertainty of the experimental data. The process of fitting the RMSE curve offers a practical means to derive an uncertainty estimate for every model prediction.

Comparison with Distance-to-Model Metrics for Error Modeling. Some other Tanimoto similarity-based metrics have been evaluated for modeling prediction errors. For example, SimilarityNearest1[23] denotes the Tanimoto similarity to the closest training set molecule; the assumption for this metric is that a prediction for a molecule is more likely to be reliable if the molecule is structurally similar to at least one molecule in the training set. Another metric, SimilarityNearest5, is the mean similarity to five nearest neighbors in the training set. They correspond to the κ and γ DA-indices of Harmeling.[37] To compare the association between their values and the magnitudes of prediction errors, we plotted the absolute prediction errors against SimilarityNearest1, SimilarityNearest5, and SDC for the rat oral toxicity data set in Figure 11. The results for the other three data sets were highly similar (Figures S1−S3, Supporting Information). The plots show that the correlation with the prediction error was marginal for SimilarityNearest1 and slightly stronger for SimilarityNearest5, especially at the tail end (toward higher mean similarity). Because the ν-NN

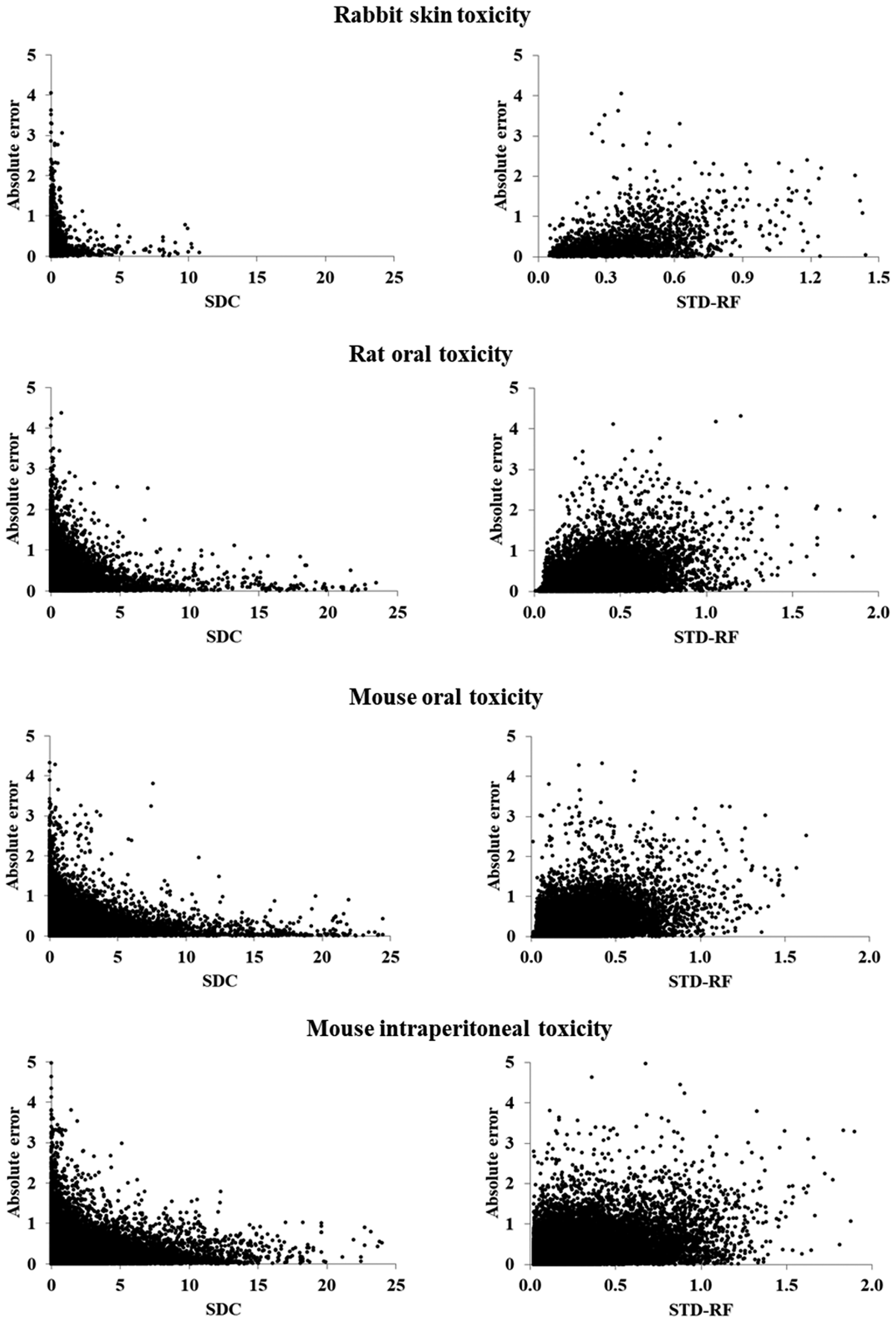


Figure 12. Distribution of RF prediction errors with respect to SDC compared to that with respect to the standard deviation of all RF tree predictions (STD-RF) for four acute toxicity data sets. The plots show a significantly stronger correlation between the prediction error and SDC.

predictions used a Tanimoto distance threshold of 0.6, the models gave no predictions for molecules with SimilarityNearest1 less than 0.4.

Note that the moving average was highest at SDC values close to 0 but showed a sharp drop with an initial increase in SDC from 0 and then remained largely unchanged across the higher SDC range. This indicates that the SDC values of the bulk of the bad predictions are very close to zero. The moving averages with respect to SimilarityNearest1 and SimilarityNearest5 were higher than those with respect to SDC, indicating that some predictions with large errors are spread across the range of SimilarityNearest1 and SimilarityNearest5 values.

Comparison with Ensemble-Variance Metric for Error Modeling. To compare the performance of the SDC metric with that of the ensemble-variance metric, we calculated the standard deviation of all 500 decision tree predictions of each RF prediction for the four acute toxicity data sets in 10-fold cross validation. We then compared the plots of absolute prediction error against the standard deviations of the RF predictions to those against the SDC in Figure 12. The side-by-side

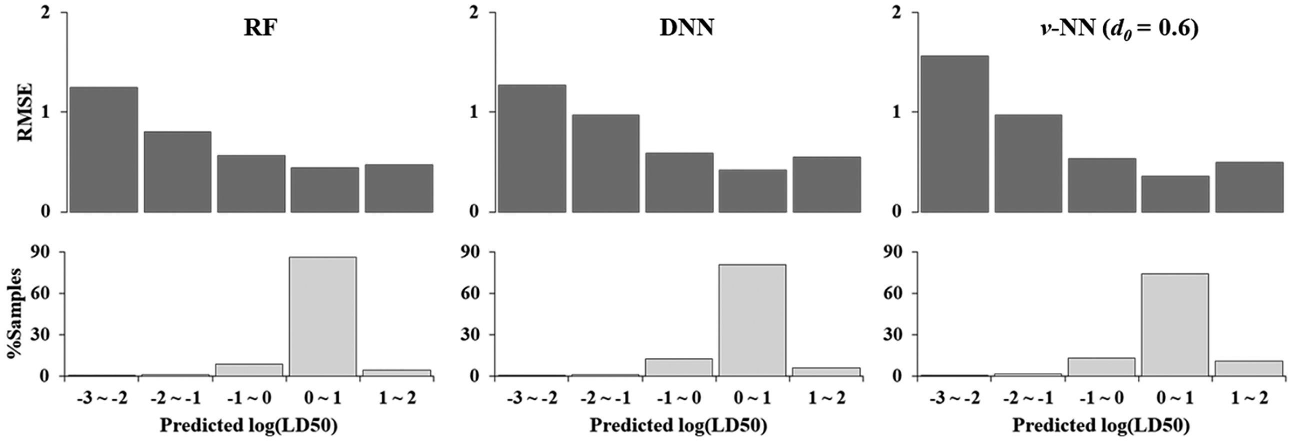


Figure 13. Root mean squared error (RMSE) of predictions for compounds grouped into each predicted log(LD50) unit interval of the mouse oral toxicity data set (upper graphs) and percentage of compounds in each predicted log(LD50) unit interval (lower graphs). The predictions were given by three machine-learning methods via 10-fold cross validation. They show a clear inverse relationship between RMSE and percentage of compounds in the log(LD50) unit intervals.

comparisons reveal that the correlation with prediction error was higher for the SDC metric. The distribution of the RF prediction error with respect to the standard deviation of the RF predictions was similar to that of the prediction error with respect to STD-CONS DM (Figure 4 of ref 10) and STD-ASNN (Figure 5 of ref 10) metrics of Tetko et al.

Sheridan proposed that regardless of the QSAR method, the predicted value itself was correlated with prediction error and therefore could be used as a DA metric.[19,23] To assess its performance, we binned the predicted toxicities of the compounds in the mouse oral toxicity data set with a bin-size of 1 log-unit and then calculated the percentage of compounds and the RMSE of each bin. Figure 13 shows that regardless of the machine-learning method used, the RMSE was the smallest for marginally toxic compounds (predicted log(LD50) in the 0−1 range), and increased with decreasing predicted log(LD50) values, becoming the highest for the most toxic compounds. A closer examination showed that the RMSE was actually inversely related to the percentage of compounds in the bins and tracked the sample distribution with respect to experimental log(LD50) ranges.

The correlation between the RMSE and molecular activity stems from the fact that almost all molecular activity data sets have highly uneven sample distributions, in which most compounds have marginal activities and only a very small fraction are highly active. Because the objective function of most machine-learning methods for regression problems is the overall or average prediction error, model parameters derived from minimizing the objective function are biased toward marginally active compounds and against highly active and inactive compounds. Thus, if the aim of QSAR modeling is to identify marginally active compounds, the predicted activity value itself can serve as a suitable DA metric. However, if the aim is to identify highly active compounds, it may not be a useful metric because it simply associates predicted high potency with a large prediction error.

■ SUMMARY

To develop and validate a QSAR model, it used to be common practice to begin by segregating a data set into a training set and a test set. The training set was used to develop a model, and model performance was assessed by calculating the RMSE between the predicted and experimental results of test set compounds. The RMSE was assumed to be representative of model performance for all other compounds. The concept of domain applicability was introduced when it was realized that regardless of the size of a test set, the RMSE of the test compounds is unlikely to be representative of model performance in real-world applications. Structurally novel compounds are routinely synthesized in chemical and pharmaceutical research. These are the molecules for which we have the greatest need of making high-quality QSAR predictions. However, the results of this and other studies clearly indicate that these molecules pose the greatest challenge for making accurate and consistent QSAR predictions because they are the least likely to have structurally similar compounds in the training sets. If predictions were accompanied by measures of their uncertainty, blind trust could be prevented, helping end-users make more informed decisions. This is why a defined DA and appropriate measures of goodness-of-fit, robustness, and predictivity are key requirements for QSAR models to gain acceptance by regulatory authorities.[38]

In this study, we introduced a new DA metric as the sum of the distance-weighted contributions of training molecules to prediction accuracy. We assessed the performance of the metric using eight molecular data sets of varying sizes and a number of commonly used machine-learning methods, ranging from arguably the simplest to the most sophisticated in terms of model parameter counts. The results indicated that the new metric captures the reliability of model predictions for individual molecules, independent of the machine-learning method and molecular descriptors used to build the prediction models. For all methods and data sets, we demonstrated that a SDC value close to zero is associated with the highest prediction errors and that increased SDC is associated with reduced prediction errors. We also demonstrated that we could use SDC to develop robust RMSE prediction models. The RMSE models display an exponential decrease in prediction variability with increasing SDC. Prediction uncertainty eventually levels off with increasing SDC, an outcome likely dictated by the variability of the training data. A comparison of the new metric with DA metrics previously evaluated for modeling prediction errors indicated that the former has a much stronger correlation than the latter with prediction uncertainty. We therefore propose the new DA metric as a robust measure for evaluating QSAR prediction reliability.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00114.

> Scatter plots of absolute prediction errors with respect to SDC, SimilarityNearest1, and SimilarityNearest5 domain applicability metrics for rabbit skin toxicity (Figure S1), mouse oral toxicity (Figure S2), and mouse intraperitoneal toxicity (Figure S3) data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: rliu@bhsai.org (R.L.)

*E-mail: sven.a.wallqvist.civ@mail.mil (A.W.).

ORCID

Ruifeng Liu: 0000-0001-7582-9217

Michael G. Feasel: 0000-0001-7029-2764

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) Hansch, C.; Maloney, P.; Fujita, T.; Muir, R. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178−180.

(2) Hansch, C.; Hoekman, D.; Leo, A.; Zhang, L.; Li, P. The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicol. Lett.* **1995**, *79*, 45−53.

(3) 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.

(4) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR models with error estimation: vapor pressure and log *P*. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046−1051.

(5) 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.

(6) Guha, R.; Jurs, P. C. Determining the validity of a QSAR model−a classification approach. *J. Chem. Inf. Model.* **2005**, *45*, 65−73.

(7) Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* **2006**, *11*, 700−707.

(8) Schroeter, T. S.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K. R. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 651−664.

(9) Guha, R.; Van Drie, J. H. Structure−activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.

(10) 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.

(11) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315−1326.

(12) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J. Chem. Inf. Model.* **2009**, *49*, 1762−1776.

(13) Kuhne, R.; Ebert, R. U.; Schuurmann, G. Chemical domain of QSAR models from atom-centered fragments. *J. Chem. Inf. Model.* **2009**, *49*, 2660−2669.

(14) Clark, R. D. DPRESS: Localizing estimates of predictive uncertainty. *J. Cheminf.* **2009**, *1*, 11.

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

(16) 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.

(17) Ellison, C. M.; Sherhod, R.; Cronin, M. T.; Enoch, S. J.; Madden, J. C.; Judson, P. N. Assessment of methods to define the applicability domain of structural alert models. *J. Chem. Inf. Model.* **2011**, *51*, 975−985.

(18) Soto, A. J.; Vazquez, G. E.; Strickert, M.; Ponzoni, I. Target-Driven Subspace Mapping Methods and Their Applicability Domain Estimation. *Mol. Inf.* **2011**, *30*, 779−789.

(19) Sheridan, R. P. Three useful dimensions for domain applicability in QSAR models using random forest. *J. Chem. Inf. Model.* **2012**, *52*, 814−823.

(20) Briesemeister, S.; Rahnenfuhrer, J.; Kohlbacher, O. No longer confidential: estimating the confidence of individual regression predictions. *PLoS One* **2012**, *7*, e48723.

(21) Keefer, C. E.; Kauffman, G. W.; Gupta, R. R. Interpretable, probability-based confidence metric for continuous quantitative structure-activity relationship models. *J. Chem. Inf. Model.* **2013**, *53*, 368−383.

(22) Wood, D. J.; Carlsson, L.; Eklund, M.; Norinder, U.; Stalring, J. QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 203−219.

(23) Sheridan, R. P. The Relative Importance of Domain Applicability Metrics for Estimating Prediction Errors in QSAR Varies with Training Set Diversity. *J. Chem. Inf. Model.* **2015**, *55*, 1098−1107.

(24) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability domain: towards a more formal definition. *SAR QSAR Environ. Res.* **2016**, *27*, 893−909.

(25) Kaneko, H.; Funatsu, K. Applicability Domains and Consistent Structure Generation. *Mol. Inf.* **2017**, *36*, 1−2.

(26) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Tice, R. R.; Huang, R. Prediction of Cytochrome P450 Profiles of Environmental Chemicals with QSAR Models Built from Drug-like Molecules. *Mol. Inf.* **2012**, *31*, 783−792.

(27) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - A Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(28) Szklarczyk, D.; Santos, A.; von Mering, C.; Jensen, L. J.; Bork, P.; Kuhn, M. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **2016**, *44*, D380−384.

(29) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(30) 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.

(31) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(32) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-fitting with a genetic algorithm: a method for developing classification structure-activity relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.

(33) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.

(34) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.

(35) Hall, L. H.; Kier, L. B. The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.

(36) 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.

(37) Harmeling, S.; Dornhege, G.; Tax, D.; Meinecke, F.; Muller, K. From outliers to prototypes: ordering data. *Neurocomputing* **2006**, *69*, 1608−1618.

(38) Benfenati, E.; Diaza, R. G.; Cassano, A.; Pardoe, S.; Gini, G.; Mays, C.; Knauf, R.; Benighaus, L. The acceptance of in silico models for REACH: Requirements, barriers, and perspectives. *Chem. Cent. J.* **2011**, *5*, 58.