JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 117−126

pubs.acs.org/jcim

Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure−Activity Relationship Models Based on Deep Neural Networks?

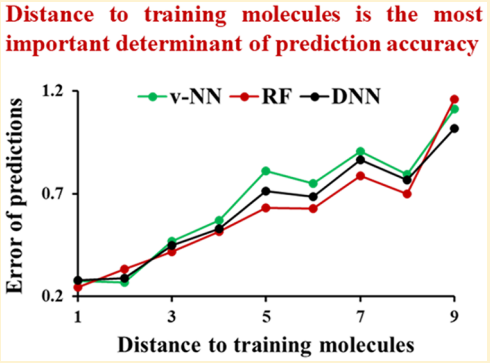
Ruifeng Liu,*,† Hao Wang,† Kyle P. Glover,§ Michael G. Feasel,§ and Anders Wallqvist*,†

†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

§U.S. Army−Edgewood Chemical Biological Center, Aberdeen Proving Ground, Maryland 21010, United States

Supporting Information

ABSTRACT: Deep neural networks (DNNs) are the major drivers of recent progress in artificial intelligence. They have emerged as the machine-learning method of choice in solving image and speech recognition problems, and their potential has raised the expectation of similar breakthroughs in other fields of study. In this work, we compared three machine-learning methods—DNN, random forest (a popular conventional method), and variable nearest neighbor (arguably the simplest method)—in their ability to predict the molecular activities of 21 *in vivo* and *in vitro* data sets. Surprisingly, the overall performance of the three methods was similar. For molecules with structurally close near neighbors in the training sets, all methods gave reliable predictions, whereas for molecules increasingly dissimilar to the training molecules, all three methods gave progressively poorer predictions. For molecules sharing little to no structural similarity with the training molecules, all three methods gave a nearly constant value—approximately the average activity of all training molecules—as their predictions. The results confirm conclusions deduced from analyzing molecular applicability domains for accurate predictions, i.e., the most important determinant of the accuracy of predicting a molecule is its similarity to the training samples. This highlights the fact that even in the age of deep learning, developing a truly high-quality model relies less on the choice of machine-learning approach and more on the availability of experimental efforts to generate sufficient training data of structurally diverse compounds. The results also indicate that the distance to training molecules offers a natural and intuitive basis for defining applicability domains to flag reliable and unreliable quantitative structure−activity relationship predictions.



INTRODUCTION

In recent years, deep neural networks (DNNs) have emerged as the machine-learning method of choice in image[1] and speech recognition,[2] and their versatility has led to unprecedented progress in artificial intelligence.[3] This success has spurred applications of DNNs in many other fields, including quantitative structure−activity relationship (QSAR) prediction of molecular activities.[4] In a Kaggle competition sponsored by Merck in 2012 to examine the ability of modern machine-learning methods to solve QSAR problems in pharmacology and drug discovery, DNNs were among the winning entries. Merck researchers followed this up in a detailed study that specifically compared the performance of DNN models to that of random forest (RF) models and showed that DNN models could routinely make better prospective predictions on a series of large, diverse QSAR data sets generated as part of Merck's drug discovery efforts.[5] Since then, many other studies have been published comparing DNNs and conventional machine-learning methods in terms of their ability to predict molecular activities.[6−11] The bulk of these studies showed better performance with DNNs than with other machine-learning methods, raising the hope that DNNs may help overcome key modeling challenges in drug discovery, one of which is to provide guidance for efficient exploration of new chemical spaces and the discovery and evaluation of structurally novel drugs. Interestingly, with Merck Challenge data, Winkler and Le showed that a single-hidden layer shallow neural network performed similarly as the DNNs.[6] Their results may look surprising, but they are consistent with the Universal Approximation Theorem, which states that a feedforward network with a single hidden layer is sufficient to represent any function.[12,13]

Most published studies evaluating the performance of DNNs and conventional machine-learning methods have relied on global performance metrics, such as the correlation coefficient (R^2) or the root mean squared error (RMSE) between the predicted and experimental results of all molecules, because of the large number of data sets they examined (a few tens to more than a thousand). Although such global metrics are

Received: June 3, 2018
Published: November 9, 2018

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

appealing because they conveniently provide a single number to interpret, they may miss important prediction details for individual molecules or groups of molecules in different activity ranges. In a recent study, we examined the performance of DNN, RF, and variable nearest neighbor (v-NN) methods with *in vivo* chemical toxicity and *in vitro* molecular activity data sets. Judged by R^2 and RMSE, DNN performance improved with increasing data set size and outperformed RF and v-NN models for large data sets, consistent with previously published studies. However, closer examination revealed that all machine-learning methods gave good predictions for molecules with marginal activity and markedly poorer predictions for highly active and highly inactive molecules.[14] Because one of the main objectives of predictive toxicology and drug discovery is to identify highly active molecules, these results suggest that the potential for machine learning to advance predictive toxicology and drug discovery might be substantially lower than the global performance metrics R^2 and RMSE indicate. They also suggest that, when evaluating the performance of machine-learning methods, one should not only rely on global metrics but also examine detailed prediction performance to better understand the strength and weakness of the machine-learning methods.

In the present study, we analyzed details of machine learning predictions of molecular activities with the aim of understanding if DNNs can truly learn new relationships and provide more reliable predictions than conventional machine-learning methods for molecules whose molecular structures are not very similar to training samples. This is one of the most challenging issues facing conventional machine learning methods, as analyses of applicability domains of conventional machine-learning models indicate that the most important determinant for error of predicting molecular properties is not a machine learning method but the similarity of the molecules to the training set molecules.[15,16]

MATERIALS AND METHODS

Data Sets. We derived seven *in vivo* acute chemical toxicity data sets from the Leadscope Toxicity Database (http://www.leadscope.com/toxicity_database/). After removing entries not suitable for QSAR modeling, we collected data sets of 1745, 2191, 4115, 10 363, 11 716, 21 776, and 29 476 compounds for rabbit skin, rat subcutaneous, mouse subcutaneous, rat oral, mouse intravenous, mouse oral, and mouse intraperitoneal toxicity, respectively. Each compound has an experimentally derived LD50 value in milligrams per kilogram body weight. We converted the LD50s into log(millimoles per kilogram) before modeling. Details of our data cleaning procedure can be found elsewhere.[14]

We also used 14 of the 15 *in vitro* data sets in the Merck Molecular Activity Challenge. The LogD data set was excluded in this study mainly because of computational cost due to its large size (50 000 compounds, the largest of the Merck Challenge data sets) and also because it is a relatively straightforward property for QSAR modeling, as indicated by the good performance of multilinear regression for LogP predictions,[17] where LogD is LogP under a specific pH condition.

Molecular Descriptors. For the *in vivo* toxicity data sets, we used extended connectivity fingerprints with a diameter of four chemical bonds (ECFP_4)[18] as input molecular features. The ECFP_4 fingerprints were directly calculated from molecular structures. For the *in vitro* molecular activity data sets, the Merck Challenge only provided molecular activities and atom-pair descriptor values, and the molecular structures were not disclosed. Therefore, we used the provided atom-pair descriptor values as input features.

Machine-Learning Methods. *Deep Neural Networks.* For the *in vivo* toxicity data sets, we used a fully connected feed-forward network architecture of dimensions 2048:300:300:30:1, where the first number represents 2048 ECFP_4 fingerprint features as inputs for all data sets, followed by 300, 300, and 30 neurons in the first, second, and third hidden layers, respectively, and a single neuron in the output layer. We built seven single-task DNNs, each for an individual toxicity end point. For the DNN calculations, we used the open source Python library Keras (https://keras.io/) on top of the Theano[19] backend, 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. We reported our hyperparameter selection and DNN performance in a recent paper.[14] For each data set, we selected the 2048 ECFP_4 fingerprint features as input for the DNNs according to the following procedure:

(1) Identify all unique fingerprint features present in the whole data set.

(2) Calculate the frequency of each fingerprint feature appearing in the 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.

For the *in vitro* data sets, we first preprocessed Merck data sets using Merck-provided Python code downloaded from GitHub, and then implemented Merck DNN models, again using Merck-provided Python code downloaded from GitHub (https://github.com/RuwanT/merck). The Merck DNN models consisted of a variable number of input features, ranging from 2796 to 6559 depending on the data set: 4000, 2000, 1000, and 1000 hidden neurons in the first, second, third, and fourth hidden layers, respectively; and a single output for each model. Because we could not calculate any molecular descriptors given that Merck did not disclose molecular structure information, we used the Merck provided atom-pair descriptors as input features for the DNN calculations for the *in vitro* data sets.

Random Forests. We used the Pipeline Pilot implementation of the random forest (RF) algorithm called Forest of Random Trees (http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/) to develop the RF models. The RF model for each data set consisted of 500 decision trees. For the 7 *in vivo* toxicity data sets, we used ECFP_4 fingerprint features as molecular descriptors. For the 14 Merck *in vitro* molecular activity data sets, we used the Merck-provided atom-pair descriptors as input features. For both the *in vivo* and *in vitro* data sets, the maximum tree depth was 50, and a third of all molecular descriptors were tested as split criteria within each tree.

Variable Nearest Neighbor. The v-NN method is based on the principle that similar structures have similar activity. Its prediction is a distance-weighted average of all qualified nearest neighbors in the training set

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

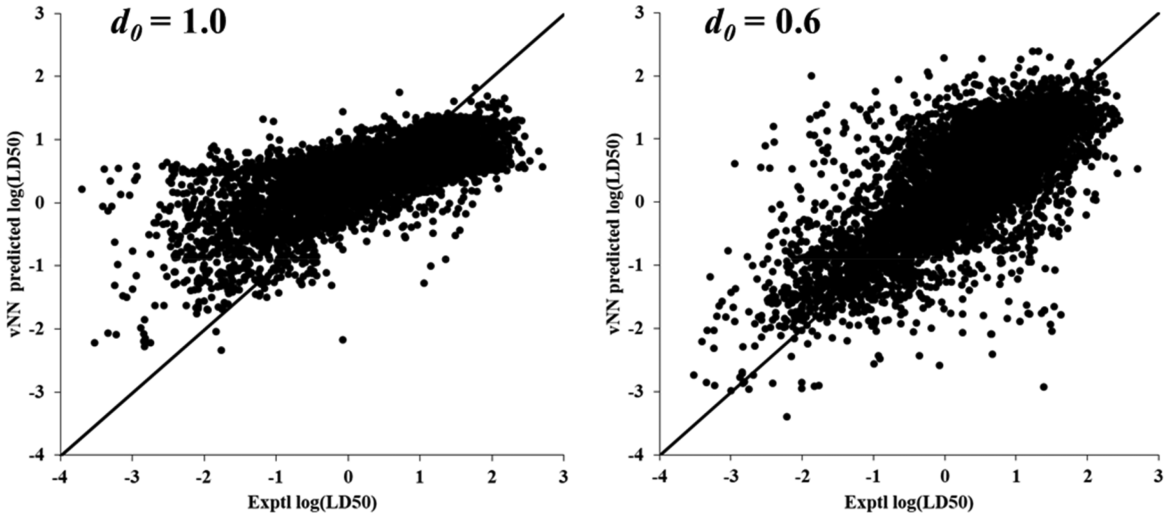


Figure 1. Impact of Tanimoto distance threshold (d_0) on the ability of the variable nearest neighbor (v-NN) method to predict the experimental log(LD50) values of a rat oral toxicity data set. In this and all subsequent figures plotting the predicted value against the experimental value, a data point on the diagonal line indicates that these values are identical. The predictions were made via 10-fold cross validation.

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

In this equation, y_i is the activity of the ith nearest neighbor in the training set, d_i is the distance between the ith nearest neighbor and the molecule for which v-NN is making a prediction, h is a smoothing factor that modulates the distance penalty, and v 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. We consider all training set near neighbors that meet the condition $d_i \leq d_0$ qualified training compounds. d_0 and h are the only model parameters to be determined from the training data.

In previous studies, we found that v-NN performance depends strongly on d_0.[20] If d_0 is small, then information on only structurally very similar compounds is used in making predictions and the results are more likely to be reliable. However, with a small d_0, the number of training set molecules meeting the Tanimoto distance threshold is small, and therefore, the number of molecules for which the method can make predictions is low. When we used the Tanimoto distance calculated from ECFP_4 molecular fingerprints, we found that the combination of $d_0 = 0.6$ and $h = 0.3$ worked well for predicting molecular activities, with both reasonable reliability and acceptable coverage (percentage of molecules for which v-NN predictions could be made).[14] For example, Figure 1 shows the results of 10-fold cross validation for the rat oral toxicity data set, calculated using ECFP_4 fingerprints with d_0 values of 0.6 and 1.0. With $d_0 = 1.0$, the RMSE of prediction was higher than that obtained with $d_0 = 0.6$, although it allowed predictions for 100% of the compounds, compared to 86% of the compounds with $d_0 = 0.6$.

Figure 1 shows that with $d_0 = 0.6$, the data points are more symmetrically distributed around the diagonal identity line than with $d_0 = 1.0$, with the latter condition leading to underestimation of toxicity for highly toxic compounds and overestimation of toxicity for nontoxic compounds. Because v-NN predictions with $d_0 = 1.0$ represent a weighted average toxicity of all training samples, whereas predictions with $d_0 = 0.6$ represent a weighted average toxicity of only training samples that met this Tanimoto distance threshold, these results indicate that within the v-NN approach, predictions based on information from all compounds are no better than predictions based on information from qualified compounds only. Thus, information from unqualified compounds may make the predictions worse. On the basis of this observation, we decided to adopt a layered v-NN prediction approach. That is, for a given test compound, we segregate the chemical space, with the test compound at the center, into ten partitions. The first partition is a sphere with a radius of $d_0 = 0.1$. The other partitions represent shells of spaces, defined by Tanimoto distances between 0.1 and 0.2, 0.2 and 0.3, ... up to 0.9 and 1.0 (Figure 2). We give a v-NN prediction for a test compound by using information on training set compounds in the closest partition only. For example, if the test compound has neighbors in the training set in layer 1 ($d_i \leq 0.1$), then only information for these compounds is used to make a prediction

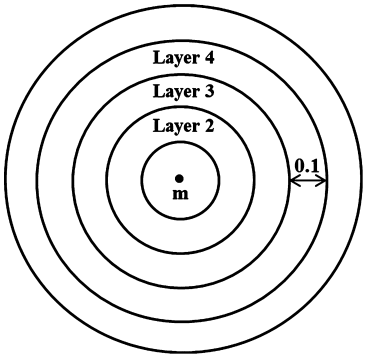


Figure 2. Scheme illustrating the layered v-NN approach. To make a prediction for molecule **m**, the entire chemical space is segregated into 10 spherical layers of equal depth (a Tanimoto distance of 0.1) with **m** at the center. The training molecules are then distributed among the layers by their Tanimoto distance to **m**, and only those in the layer closest to **m** are used in eq 1 for v-NN predictions. For comparison, we made RF and DNN predictions of **m** using models trained with all training samples. We then grouped the predictions into layers by the distance between **m** and the training samples.

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

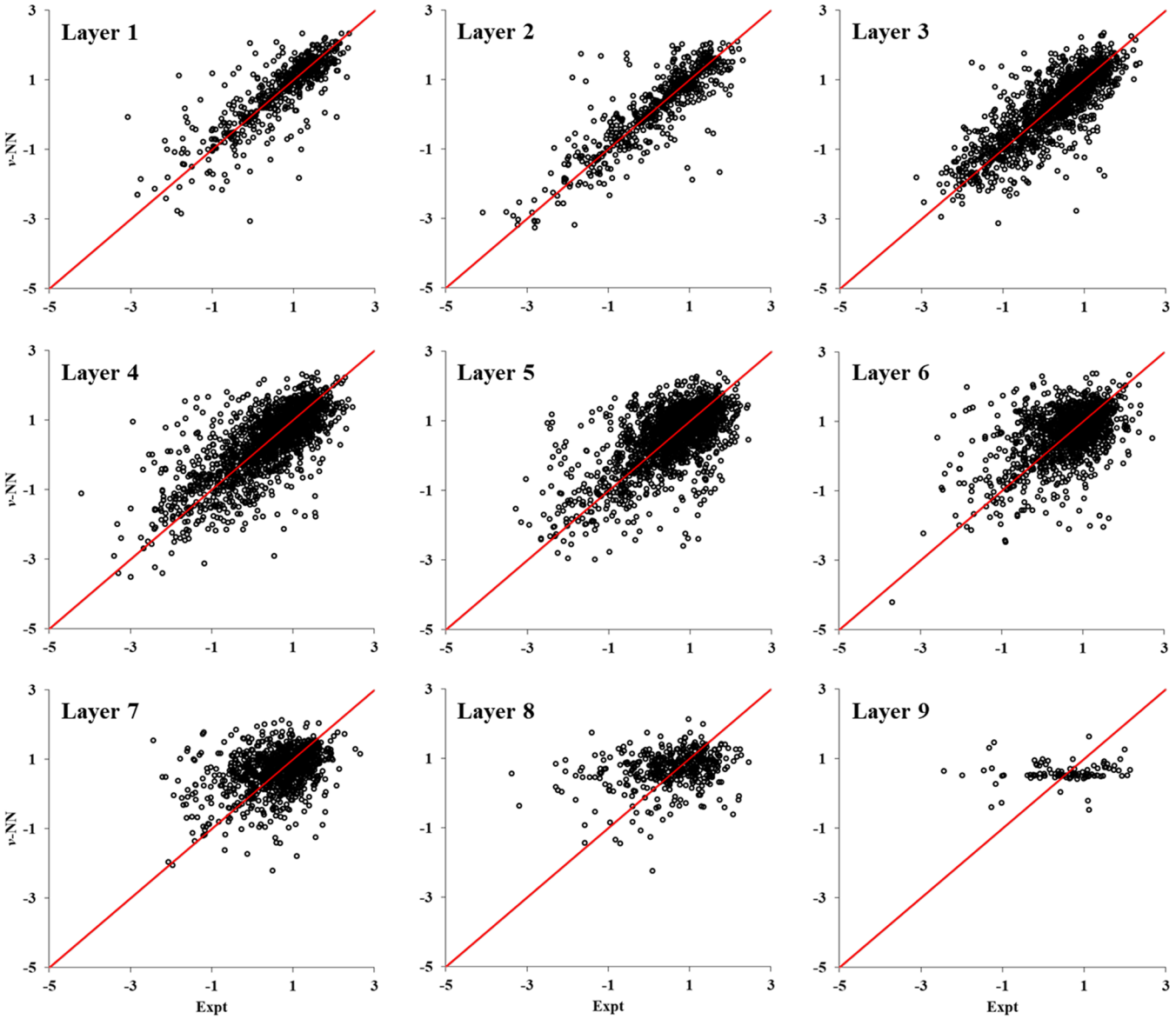


Figure 3. Plots of predicted versus experimental log(LD50) values of the rat oral toxicity data set. The predictions were made by variable nearest neighbor (v-NN) model via 10-fold cross validation. The predicted results were grouped, based on the shortest Tanimoto distance to the training molecules, into layers described in Figure 2.

regardless of whether or not training samples exist in the remaining layers. If the test compound has no neighbors in layer 1, then only information on training samples in layer 2 is used to make predictions. We refer to such v-NN predictions as layered predictions. For the acute toxicity data sets, we define the layers based on Tanimoto distances calculated using ECFP_4 fingerprints; for Merck *in vitro* molecular activity data sets, the layers were defined by Tanimoto distances calculated using the atom-pair fingerprints derived from Merck-provided descriptor values, by stripping the counts of atom-pairs in a molecule and retaining only information on their presence or absence.

RESULTS

Layered v-NN Predictions for *in Vivo* Toxicity Data Sets. To evaluate the performance of layered v-NN predictions, we performed 10-fold cross validation calculations for the *in vivo* toxicity data sets. Thus, we first split each data set randomly into 10 equal-sized groups and then used 9 of them as the training set to predict the toxicities of the compounds in the left-out group. This process was repeated nine times so that each and every group was left-out once and used as a test set. We used a smoothing factor of 0.3 and Tanimoto distances calculated with ECFP_4 fingerprints in performing the v-NN calculations. For all data sets, the number of compounds with layer 9 predictions, i.e., those without training set compounds within a Tanimoto distance of 0.8, was very small, and an even smaller number of compounds were present with layer 10 predictions. In analyzing the data, we combined the predictions for layers 9 and 10 as layer 9 predictions.

For the rat oral toxicity data set, we compared the predicted toxicities in different layers with the experimental results (Figure 3). Because the results for the other six data sets were similar, we have included them in the Supporting Information (Figure S1). In the Supporting Information (Table 1), we

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

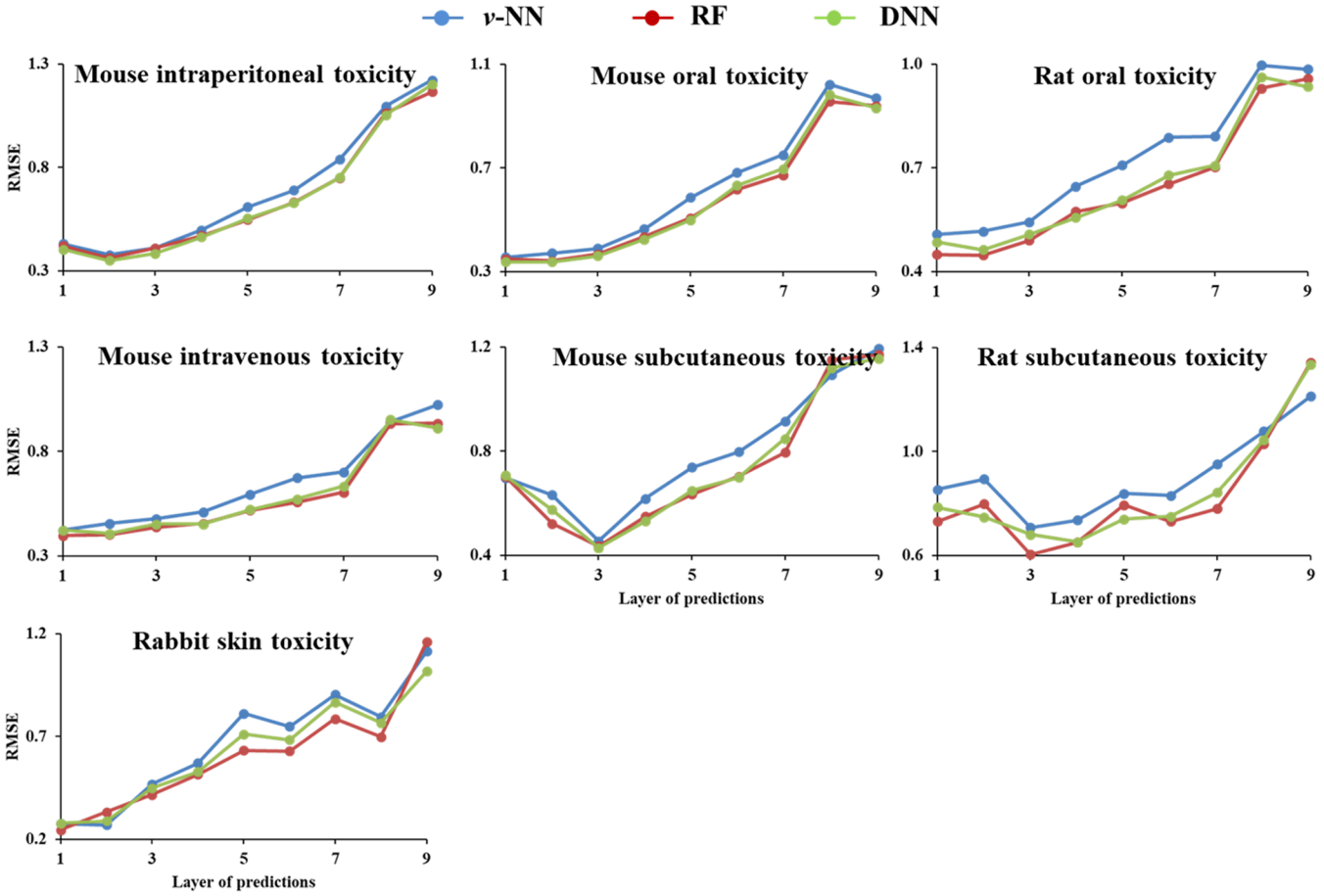


Figure 4. Root mean squared error (RMSE) between the predicted and experimental log(LD50) values of compounds in different layers of the seven *in vivo* acute toxicity data sets. The predictions were made via 10-fold cross validation, using the variable nearest neighbor (*v*-NN), random forest (RF), and deep neural network (DNN) models. The layers were described in Figure 2.

present the squared correlation coefficient, R^2, between the predicted and experimental log(LD50) values for each layer of the seven *in vivo* data sets. Both the table and figures show that the predictions were highly reliable when the test molecules had close neighbors in the training set but less reliable when they had no such neighbors. This became increasingly apparent for predictions of compounds in layers 7 to 9. The R^2 for these compounds was smaller than 0.1, indicating no correlation between the predicted and experimental results. A notable feature for compounds in layer 9 is that the predicted toxicity was roughly constant, regardless of molecular structure and experimentally measured log(LD50) value. This is not surprising for layered *v*-NN predictions, because, according to eq 1, a layer 9 or 10 *v*-NN prediction is simply the average toxicity of almost all training samples. Because RF and DNN are based on more intricate algorithms, it is interesting to assess how they perform under the same circumstances.

RF and DNN Performance for the *in Vivo* Toxicity Data Sets. Unlike the layered *v*-NN approach, which uses information on only qualified neighbors in a training set to make predictions, the RF and DNN methods use information on all training samples to first build the models, which they then use to make predictions. To assess model performance for the RF and DNN methods in a manner similar to that for the layered *v*-NN approach, we first made RF and DNN predictions for all compounds in 10-fold cross validation, and then calculated Tanimoto distances between the test and training compounds. Subsequently, we segregated the compounds by their shortest Tanimoto distance to the training samples into groups similar to those of the layered *v*-NN approach. To our surprise, the distributions of RF- and DNN-predicted versus experimental log(LD50) values of the rat oral toxicity data set are remarkably similar to that shown in Figure 3, even though both RF and DNN are much more sophisticated machine-learning methods. Because they are so similar, we presented them in Figures S2 and S3 of the Supporting Information. Similar results were also observed for the other *in vivo* toxicity data sets and are presented in Figures S2 and S3. The R^2 between the predicted and experimental log(LD50) values for all seven *in vivo* data sets are presented in Table S1. Figure 4 plots the RMSE of compounds within each layer of predictions for all seven *in vivo* data sets.

Although the RF and DNN models are more complex than the layered *v*-NN models, the plots in Figures 3 and S1−S3 (Supporting Information) show similar results for all models, especially for predictions of the lower layers for molecules with structurally close near neighbors in the training sets. For the six larger data sets (i.e., those excluding the rabbit skin toxicity data set, which contains too few molecules), all three methods show highly reliable predictions for compounds within a Tanimoto distance of 0.4 to any training sample (layers 1−3). For compounds without any training samples within a Tanimoto distance of 0.4, but with those within a Tanimoto distance of 0.7 (layers 4−6), all models gave inferior but

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

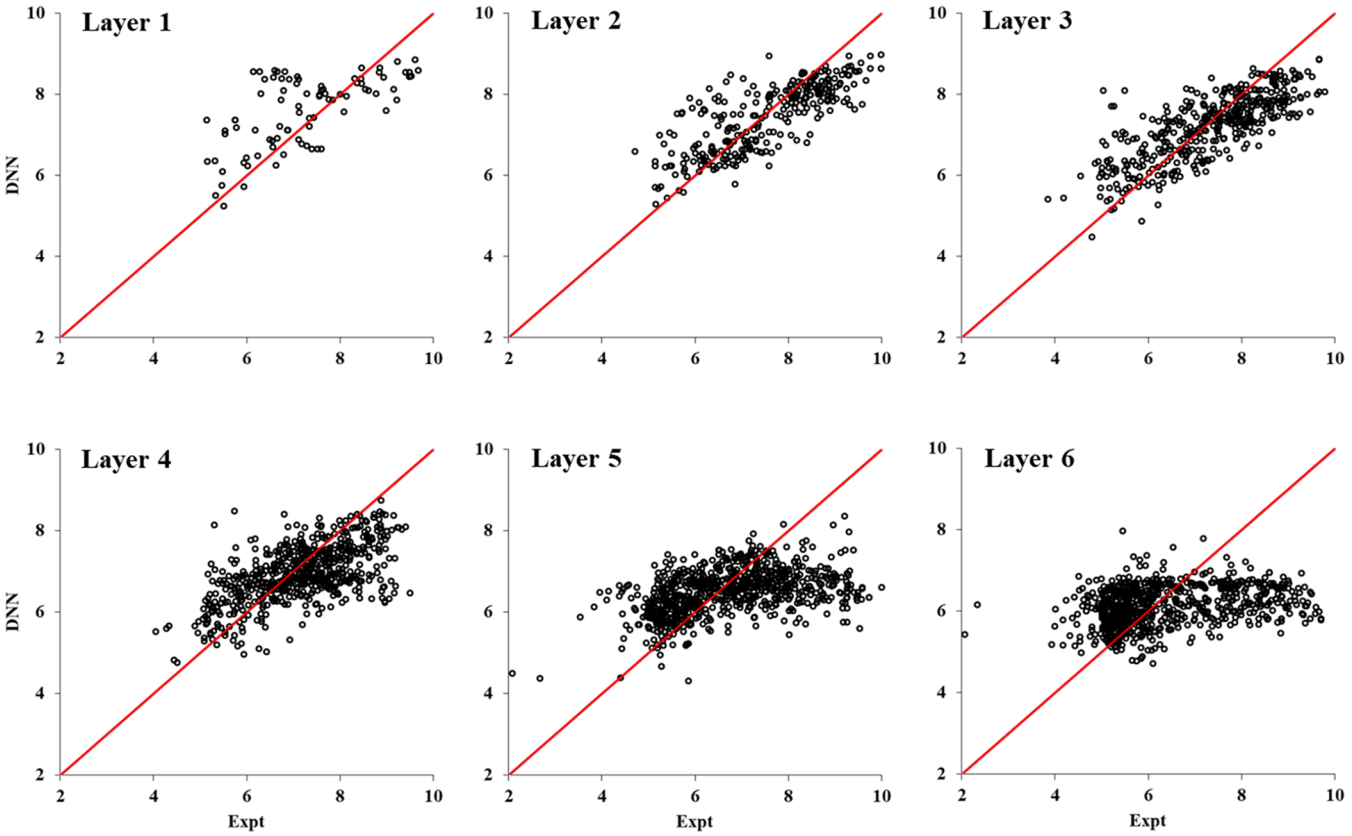


Figure 5. Plots of predicted versus experimental log(LD50) values of the CB1 data set. The predictions were made by a deep neural network (DNN) model using time-split training and test data provided by Merck Challenge (https://github.com/RuwanT/merck). The predicted results were grouped into layers by the Tanimoto distance to the training molecules, as described in Figure 2.

acceptable predictions that were clearly correlated with the experimental results (Table S1 and Figure 4). However, for test compounds with a Tanimoto distance of at least 0.7 from any training sample, none of the models gave acceptable predictions (i.e., there was simply no correlation between the predicted and experimental results). For the rabbit skin toxicity data set (the smallest acute toxicity data set), all machine-learning methods showed poorer performance, as judged by the RMSE (Figure 4) and R^2 (Table S1 in the Supporting Information) of the different prediction layers. Because DNNs employ a large number of model parameters, they require a large data set to develop a good model. These results show that all machine-learning methods performed better the greater the amount of training data.

Interestingly, the plots in Figures 3 and S1−S3 reveal a common trend regardless of the machine-learning method: the distribution of data points rotated clockwise from roughly 45° for data points in layer 1 to approximately 0° for those in layer 9. Thus, for compounds with close near neighbors in the training sets, the data points distribute symmetrically around the diagonal identity lines. Going from the lowest to highest layer, the models increasingly underestimated the toxicity of highly toxic compounds and overestimated that of the least toxic compounds, resulting in a nearly horizontal distribution of data points in layer 9. Because v-NN predictions for compounds in layer 9 are simply the average toxicity of all training samples, the horizontal distributions of layer-9 data points suggest that the RF and DNN predictions for these compounds were also close to the average toxicity of all training samples. Thus, when a compound is too far away from the training compounds, its predicted activity is close to the average activity of the training molecules regardless of the machine-learning method.

Results for *in Vitro* Molecular Activity Data Sets. For the *in vitro* data sets, each data set was provided in the form of a training set and a test set consisting of 75% and 25% of the compounds, respectively. The compounds were split into a training set and a test set based on the dates they were evaluated. This approach is intended to capture a set of training compounds representing chemistries that were synthesized and evaluated before the newer test set compounds were synthesized and made available for evaluation. This time-split test provides a more realistic estimate of model performance for new compounds, because by design, chemical and pharmaceutical research constantly explores new chemical spaces that differ from the space of the training set.[21] They are ideal test sets for assessing how good a machine-learning method can learn from a known region of chemical space and make reliable predictions for compounds of a previously unexplored region. Conventional machine-learning methods do not perform well in this respect: they give poor predictions for molecules whose structures are not well-represented by the training set. To remedy this issue, various applicability domains have been defined for flagging compounds where reliable predictions cannot be made.[22,23] Although deep learning is currently considered the most powerful machine-learning method,[9] an interesting question is whether it represents an incremental improvement over traditional machine-learning methods or a fundamental change (i.e., it learns something new and infers relationships

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

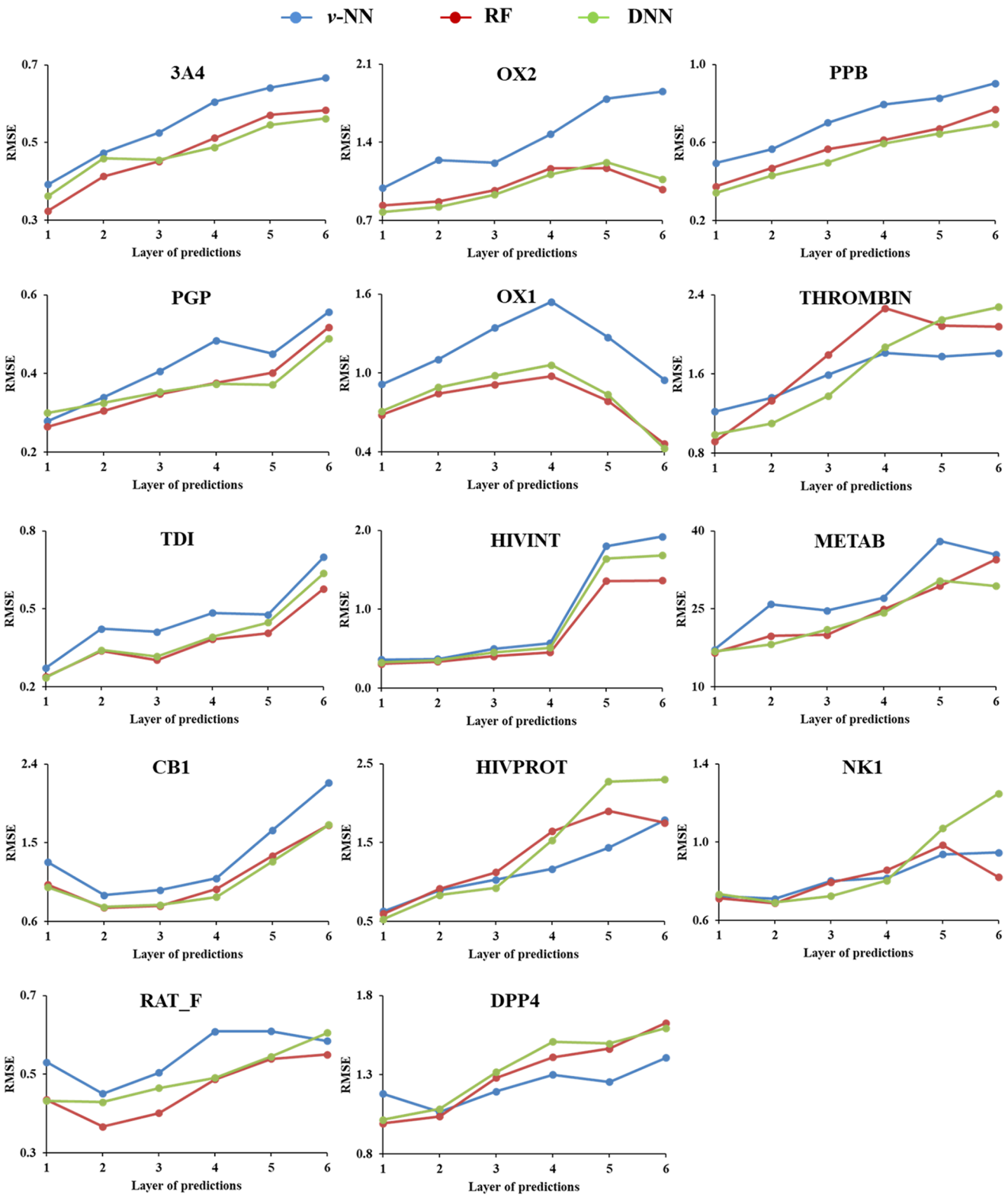


Figure 6. Root mean squared error (RMSE) between the predicted and experimental log(LD50) values of compounds in different layers of the *in vitro* molecular activity data sets. The predictions were made with the time-split training and test data provided by the Merck Challenge, using the variable nearest neighbor (*v*-NN), random forest (RF), and deep neural network (DNN) models. The layers were as described in Figure 2. Assay abbreviations: 3A4, Cyp 3A4 inhibition (pIC50); OX2, Orexin 2 receptor inhibition (pKi); PPB, human plasma protein binding [log(bound/unbound)]; PGD, transport by p-glycoprotein [log(BSA/AB)]; OX1, Orexin 1 receptor inhibition (pKi); THROMBIN, human thrombin inhibition (pIC50); TDI, time-dependent Cyp3A4 inhibition [log(IC50 without NADPH/IC50 with NADPH)]; HIVINT, inhibition of HIV integrase in a cell based assay (pIC50); METAB, percent remaining after 30 min microsomal incubation; CB1, binding to cannabinoid receptor 1 (pIC50); HIVPROT, inhibition of HIV protease (pIC50); NK1, inhibition of neurokinin 1 receptor binding (pIC50); RAT_F, log(rat bioavailability) at 2 mg/kg; DPP4, inhibition of dipeptidyl peptidase 4 (pIC50).

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

heretofore unobserved in the training data). Therefore, we used the Merck-provided time-split training and test sets to evaluate performance on the *in vitro* data sets.

We calculated Tanimoto distances between a test molecule and the training molecules, using the presence or absence of Merck-defined atom-pair features (ignoring the counts of such features present) in a molecule (see Materials and Methods). The information content of the atom-pair features is much lower than that in ECFP_4 fingerprints. As a result, we could reasonably expect the *v*-NN results to be inferior. This was corroborated by the observation that only a small fraction of molecules had a Tanimoto distance of 0.6 or more to the training molecules in the *in vitro* data sets. For RF and DNN predictions, we used the atom-pair descriptors (including counts of the atom-pairs) as provided in the Merck Challenge data sets. We then grouped the predicted results by the Tanimoto distance to the training samples into different prediction layers. Because most data sets contained only a small fraction of compounds in layers 7 and 8, and no compounds in the outmost layers, we grouped the compounds in layers 7 and 8 into a terminal layer 6.

As an example, a scatterplot of DNN-predicted activity as a function of experimentally measured activity for the molecules in the CB1 data set is presented in Figure 5. The plots of the other data sets show the same general trend and are presented in Figures S4−S6 (Supporting Information). Although we used markedly different input descriptors and DNN architectures to model the *in vitro* and *in vivo* data sets, the general trend of the predicted activity plotted against the measured activity of the molecules in the *in vitro* data sets was similar to that in the *in vivo* data sets. That is, when a test molecule has close neighbors in the training set, the predicted values are generally more reliable, as indicated by the greater number of data points distributed along the diagonal identity line in each plot. However, for molecules in increasingly higher layers (structurally further away from the training sets), the distribution of the data points rotated clockwise from around 45° for data points of the first layer to approximately 0° for data points of the last layer. Thus, we found in the *in vitro* data sets the same trend as that which we had observed in the *in vivo* data sets, i.e., regardless of the machine-learning method used, the prediction for a molecule became increasingly unreliable as the distances of the molecule from the training compounds increased.

Figure 6 shows the RMSE values between the predicted and experimental activities of the compounds in different layers of the *in vitro* data sets (the R^2 data are presented in Table S2 of the Supporting Information). For most data sets, the RF and DNN results are similar, whereas the RMSE values for *v*-NN are notably higher than the corresponding values for RF or DNN. This is evident in how the scatter of data points in Figure S4 (Supporting Information) is broader than that in Figures S5 and S6. This differs from the plots in Figure 4, in which the RMSE of all three methods are similar for all data sets. We believe this is most likely due to information loss of the atom-pair fingerprint used in the *v*-NN calculations. Although we used atom-pair counts for RF and DNN calculations, we could not do so for the *v*-NN calculations owing to the way Tanimoto distance was defined.

DISCUSSION

Numerous published studies have shown that DNNs can outperform other techniques for many machine-learning tasks, including QSAR modeling of molecular activities. DNNs may thus provide the potential means to overcome many computational and modeling challenges in drug discovery and healthcare based on advanced data analysis. Here, we examined machine-learning predictions for QSAR modeling of molecular activity in detail, using 7 *in vivo* acute toxicity and 14 *in vitro* molecular activity data sets. For these data sets, we showed that current DNN implementations may represent an incremental improvement over the other machine-learning methods examined, although their performance is largely in line with the latter methods. Like the other machine-learning methods, for a molecule with close near neighbors in the training set, DNNs are able to accurately predict its activity. However, for a molecule representing a hitherto unexamined chemical series—and therefore having no near neighbors in the training set—DNNs assign predictions close to the average of all training molecule activities, much like the other machine-learning methods. Thus, current implementations of DNN for QSAR modeling of molecular activities lack the ability to learn beyond the training set, have limited potential for guiding the exploration of new chemical space or the discovery of structurally novel drugs, and still need an applicability domain for estimating reliability of predictions. A leading applicability domain metric is ensemble variance metric, which is defined as the standard deviation of predictions given by an ensemble of prediction models.[15] However, this metric requires the development of an ensemble of prediction models, which is not easily applicable to DNNs due to the amount work required to create these models. On the other hand, our results show that for molecules within 0.3 Tanimoto distance to a training molecule, all machine-learning predictions are reasonably reliable. Therefore, experimental measurements for these molecules can be safely replaced by machine-learning predictions, so that precious resources can be redirected to where they can have the highest impact.[24,25] We like to point out that for estimating prediction errors for individual molecules, studies have shown that the closest similarity to training set compounds does not perform as well as ensemble variance.[15] However, we believe this is due to inappropriate use of similarity to training compounds, as only similarity to the nearest neighbor in the training set was used and similarities to the other training molecules were ignored. In a recent study, we defined a new similarity-based DA metric as a sum of distance-weighted contributions of all training molecules, which performs as well as if not better than the ensemble variance metric.[26]

A likely caveat of our study is that we did not examine the performance of multitask DNNs, i.e., those using a single neural network to learn and predict multiple end points. Several recently published studies have found that multitask DNNs can perform slightly better than their single-task counterparts on classification problems.[9,27,28] For regression problems, Ma et al. compared multitask and single-task DNNs on the Merck Challenge data sets.[5] They found that multitask DNNs performed slightly better for some data sets, whereas they performed similar or slightly worse than the corresponding single-task DNNs for others. When averaged across all Merck Challenge data sets, multitask DNNs performed slightly better, with smaller data sets benefiting more at the expense of worse performance on the largest data sets. The seemingly better performance of multitask DNNs generated some excitement, as one of the plausible explanations is that the relationships (weights) linking the input features to nodes in

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126

the neural network are related due to an inherent biological response similarity (i.e., what can be "learned" from one assay can be transferred to another assay end point). Interestingly, Xu et al. performed detailed data analyses on the Merck Challenge data sets and found that the apparently better performance of multitask DNNs was in large part due to assay relatedness (i.e., test molecules for one assay were in the training set of a correlated assay).[29] Thus, current implementations suggest that the transferability of a learned task is limited for molecular activities.

Our study noted that machine-learning methods ranging from the deepest (multiple hidden layers, >700 000 weight parameters) to the shallowest (v-NN with two parameters) have similar prediction accuracy when considering molecules ranging from the closest to the farthest to the training sets. Thus, the most important determinant for the prediction accuracy of molecular activities is not machine-learning method, but the distance to training molecules. This corroborates and supports the observations of Sheridan et al.[30] and Tetko et al.,[15] inferred from conventional machine-learning methods, that the error of predicting a molecule does not depend on the descriptors or machine-learning method used, but rather on the similarity to the training set molecules.

These results indicate two paths forward for using machine-learning methods to improve predictions of molecular activities: (1) coupling focused data generation with strictly defined prediction errors and (2) developing machine-learning methods that truly learn transferable biological relationships based on sparse data.

With respect to the first path, the results of this study indicate that irrespective of current machine-learning methods, if assay data generation can be tailored for modeling (e.g., by examining as many diverse chemical scaffolds as possible rather than focusing on any selected chemical series), model accuracy and applicability will be optimized. Importantly, by coupling such tailored data generation with a strictly defined and validated applicability domain, it will be possible to provide a tool that can be used with confidence to prospectively gauge the broadest possible set of chemicals. It will also provide guidance on which chemicals are not part of the model's applicability domain and suggestion on where additional assay experiments are needed. In this sense, the ability to distinguish a reliable from an unreliable prediction is paramount, as the chemical space is vast, sparsely and unevenly populated by existing chemicals used to parametrize models.

The fundamental basis for QSAR predictions is that similar molecules have similar properties. Thus, the ability to learn this similarity provides a model with the means to predict molecular properties. This enormously successful principle has been exploited using different techniques, ranging from regression analysis to machine learning. If we are to take advantage of current developments in artificial intelligence and go beyond the similarity principle, machine-learning methods will need to truly learn not only similarities, but biological and chemical principles as well.

The situation in the drug development field, in which data are sparse and major efforts are required to generate data points for chemical and biological assays, is opposite to that of other fields, such as image and speech recognition, where data are in ample supply. Although most chemical fingerprints should provide sufficient chemical characterization, we lack sufficient chemical and biological data to enable standard machine-learning methods to learn the underlying processes governing assay outcomes.

Typical assay data range from physiochemical properties, single enzyme assay data, cell-based screening data, to animal *in vivo* outcomes, representing a wealth of chemical and biological processes. The diversity of processes reflected by these data suggests that to apply artificial intelligence in learning how to make predictions for these end points, the DNN constructs need to be trained on mechanisms rather than on outcomes. These mechanisms or DNN models could then be integrated and interrogated depending on the modeled assay end point, similar to an adverse outcome pathway framework in predictive toxicology.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00348.

Correlation coefficients (R^2) between predicted and measured molecular activities (Tables S1 and S2) and scatterplots of predicted versus measured molecular activities (Figures S1−S6) (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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Dr. Tatsuya Oyama in editing the manuscript. 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 CBCall14-CBS-05-2-0007. 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) Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural computation* **2017**, *29*, 2352−2449.

(2) Deng, L.; Li, X. Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing* **2013**, *21*, 1060−1089.

(3) Brown, N.; Sandholm, T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* **2018**, *359*, 418−424.

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

(5) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.

(6) Winkler, D. A.; Le, T. C. Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. *Mol. Inf.* **2017**, *36*, 1781141.

(7) Koutsoukas, A.; Monaghan, K. J.; Li, X.; Huan, J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminf.* **2017**, *9*, 42.
(8) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068−2076.
(9) Lenselink, E. B.; Ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminf.* **2017**, *9*, 45.
(10) Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today* **2017**, *22*, 1680−1685.
(11) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.
(12) Cybenko, G. Approximation by superposition of sigmoidal functions. *Math. Control Signals Syst* **1989**, *2*, 303−314.
(13) Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **1991**, *4*, 251−257.
(14) Liu, R.; Madore, M.; Glover, K. P.; Feasel, M. G.; Wallqvist, A. Assessing deep and shallow learning methods for quantitative prediction of acute chemical toxicity. *Toxicol. Sci.* **2018**, *164*, 512−526.
(15) 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.
(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) 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.
(18) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(19) Al-Rfou, R.; Alain, G.; Almahairi, A.; Angermueller, C.; Bahdanau, D.; Ballas, N.; Bastien, F.; Bayer, J.; Belikov, A.; Belopolsky, A.; Bengio, Y.; Bergeron, A.; Bergstra, J.; Bisson, V.; Snyder, J. B.; Bouchard, N. Theano: A Python framework for fast computation of mathematical expressions. *arXiv.org* **2016**, 1605.02688.
(20) Liu, R.; Tawa, G.; Wallqvist, A. Locally weighted learning methods for predicting dose-dependent toxicity with application to the human maximum recommended daily dose. *Chem. Res. Toxicol.* **2012**, *25*, 2216−2226.
(21) Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.
(22) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315−1326.
(23) Schwaighofer, A.; Schroeter, T.; Mika, S.; Blanchard, G. How wrong can we get? A review of machine learning approaches and error bars. *Comb. Chem. High Throughput Screening* **2009**, *12*, 453−468.
(24) Tetko, I. V.; Poda, G. I.; Ostermann, C.; Mannhold, R. Accurate in silico logP predictions: One can't embrace the unembraceable. *QSAR Comb. Sci.* **2009**, *28*, 845−849.
(25) Tetko, I. V.; Poda, G. I.; Ostermann, C.; Mannhold, R. Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chem. Biodiversity* **2009**, *6*, 1837−1844.
(26) 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.
(27) Kearnes, S.; Goldman, B.; Pande, V. Modeling industrial ADMET data with multitask networks. *arXiv.org* **2017**, 1606.08793.
(28) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions. *arXiv.org* **2014**, 1406.1231.
(29) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490−2504.
(30) 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. *Journal of chemical information and computer sciences* **2004**, *44*, 1912−1928.

DOI: 10.1021/acs.jcim.8b00348
J. Chem. Inf. Model. 2019, 59, 117−126