TOXICOLOGICAL SCIENCES, 164(2), 2018, 512–526

doi: 10.1093/toxsci/kfy111
Advance Access Publication Date: May 2, 2018
Research Article

Assessing Deep and Shallow Learning Methods for Quantitative Prediction of Acute Chemical Toxicity

Ruifeng Liu,*,[1] Michael Madore,* Kyle P. Glover,†,‡ Michael G. Feasel,‡ and Anders Wallqvist*,[1]

*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; †Defense Threat Reduction Agency, Ft Belvoir, Virginia 22060; and ‡U.S. Army - Edgewood Chemical Biological Center, Operational Toxicology, Aberdeen Proving Ground, Maryland 21010

[1]To whom correspondence should be addressed. Fax: 1-301-619-1983; E-mail: rliu@bhsai.org and sven.a.wallqvist.civ@mail.mil.

ABSTRACT

Animal-based methods for assessing chemical toxicity are struggling to meet testing demands. *In silico* approaches, including machine-learning methods, are promising alternatives. Recently, deep neural networks (DNNs) were evaluated and reported to outperform other machine-learning methods for quantitative structure–activity relationship modeling of molecular properties. However, most of the reported performance evaluations relied on global performance metrics, such as the root mean squared error (RMSE) between the predicted and experimental values of all samples, without considering the impact of sample distribution across the activity spectrum. Here, we carried out an in-depth analysis of DNN performance for quantitative prediction of acute chemical toxicity using several datasets. We found that the overall performance of DNN models on datasets of up to 30 000 compounds was similar to that of random forest (RF) models, as measured by the RMSE and correlation coefficients between the predicted and experimental results. However, our detailed analyses demonstrated that global performance metrics are inappropriate for datasets with a highly uneven sample distribution, because they show a strong bias for the most populous compounds along the toxicity spectrum. For highly toxic compounds, DNN and RF models trained on all samples performed much worse than the global performance metrics indicated. Surprisingly, our variable nearest neighbor method, which utilizes only structurally similar compounds to make predictions, performed reasonably well, suggesting that information of close near neighbors in the training sets is a key determinant of acute toxicity predictions.

Key words: machine learning; deep neural networks; random forests; variable nearest neighbor method; acute toxicity; QSAR.

Computational approaches for predicting chemical toxicity are needed to keep pace with the increasing number of chemicals for which toxicity assessment is required (Taylor *et al.*, 2014; Vogelgesang, 2002). These nontesting approaches are cost-effective, can be scaled to address large data needs, and help to reduce, refine, and replace animal testing. However, to realize these benefits with any computational approach, its predictive accuracy must be transparent and its applicability domain clearly defined (Benfenati *et al.*, 2011).

Increasing data requirements are apparent in the industrial sector and for military preparedness. The initial data requirements to support the REACH legislation in the EU could not have been met without computational approaches, especially given the restrictions on animal testing (Burden *et al.*, 2015). Read-across has proven to be a useful approach, which utilizes structural analogs or chemical classes to estimate toxicity hazards (Ball *et al.*, 2014). Recent efforts have focused on optimizing these approaches, using a standard set of rules to make toxicity estimates (Ball *et al.*, 2016). Although read-across proved useful in supporting hazard identification for REACH, this type of approach is resource-intensive and not suitable for large-scale screening initiatives.

Table 1. Toxicity Categorization Criteria, Signal Words, and Labeling Symbols of the U.S. Environmental Protection Agency Office of Pesticide Programs, and Distribution of Samples Among These Categories in the Acute Toxicity Datasets

Category Signal Word	I Danger	II Warning	III and IV Caution
Symbol	Skull & Crossbones	No Symbol	No Symbol
Skin LD50 (mg/kg)	$\leq$200	200 $<$ and $\leq$ 2000	$>$2000
Oral LD50 (mg/kg)	$\leq$50	50 $<$ and $\leq$ 500	$>$500
Rabbit skin toxicity dataset	183 (10.4%)	555 (31.8%)	1014 (58.1%)
Rat oral toxicity dataset	1092 (10.5%)	2463 (23.8%)	6808 (56.7%)
Mouse oral toxicity dataset	981 (4.5%)	5550 (25.5%)	15 245 (70.0%)

Considering that the CAS registry—the gold standard for chemical substance information (https://www.cas.org/content/chemical-substances; last accessed May 13, 2018)—currently contains 133 million chemical substances with approximately 15, 000 new substances added each day (Chemical Abstract Service Annual Report 2015, https://acswebcontent.acs.org/annualreport/program_cas.html; last accessed May 13, 2018), the Department of Defense (DoD) faces a formidable task in maintaining readiness to chemical threats. The DoD recently commissioned a National Academy of Sciences (NAS) report to develop a strategy for addressing this data gap, using 21st century predictive toxicology tools (NAS, 2015). The report recommended a tiered testing strategy, using *in silico* profiling in early tiers to aid in prioritizing and filtering large chemical lists. Although quantitative structure-activity relationship (QSAR) models hold promise to meet this need, their predictive accuracy requires careful assessment and optimization.

In recent years, deep neural networks (DNNs) have emerged to outperform all other machine-learning methods in image and speech recognition, and enabled unprecedented progress in artificial intelligence. This success has spurred applications of DNNs in many other fields, including QSAR modeling of molecular activities. In 2012, the pharmaceutical company Merck sponsored a Kaggle competition to examine the ability of modern machine-learning methods to solve QSAR problems in drug discovery. The DNN approach was the method in many of the winning entries in the competition. 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 datasets generated as part of Merck's drug discovery efforts (Ma *et al.*, 2015).

Most published studies comparing the performance of DNN to other machine-learning methods use global performance metrics, such as the correlation coefficient or the root mean squared error (RMSE) between the predicted and experimental values of all samples. These metrics provide a convenient global assessment of prediction performance; however, they may not be appropriate for datasets in which the samples are unevenly distributed across the activity spectrum. Unfortunately, most chemical toxicity and drug discovery datasets have a highly uneven distribution of samples. This is because the number of highly toxic chemicals is small compared with the vast number of existing chemicals. The situation is similar to chemical activity at a specific drug target, where the number of compounds highly active at the target is small compared with all of the chemicals tested. In both predictive toxicology and drug discovery, the main goal of QSAR predictions is to help identify highly active molecules. Global performance metrics, however, may be biased toward the most populous and usually marginally active compounds, owing to an uneven distribution of compounds. To examine the impact of such a distribution of samples on deep and conventional (shallow) machine-learning methods, we used 7 *in vivo* acute chemical toxicity datasets and 15 *in vitro* molecular activity datasets of different sizes to evaluate the prediction performance of DNN, RF, and our recently developed variable nearest neighbor (v-NN) methods (Liu *et al.*, 2012).

MATERIALS AND METHODS

Acute Chemical Toxicity Datasets

In this study, we used the acute chemical toxicity data in the Leadscope Toxicity Database (http://www.leadscope.com/toxicity_database/; last accessed May 13, 2018). These data were curated from the Registry of Toxic Effects on Chemical Substances. To evaluate the impact of training set size on performance, we initially used 4 datasets of varying sizes: rabbit skin, rat oral, mouse oral, and mouse intraperitoneal toxicity datasets with 2296, 15 752, 34 233, and 52 228 entries, respectively. However, some entries in the database were incompatible with QSAR studies and needed to be removed prior to modeling. These included entries without associated molecular structures as well as those of chemical mixtures and salt formulations. After we removed these entries, the numbers of entries for the corresponding datasets were 1745, 10 363, 21 776, and 29 476. Each entry contained a molecular structure and an experimentally determined LD50 value in mg/kg. In complementary calculations to broadly examine large *in vivo* toxicity datasets ($>$2000 compounds) as well as model parameter selections, we further included datasets of rat subcutaneous toxicity, mouse subcutaneous toxicity, and mouse intravenous toxicity containing 2191, 4115, and 11 716 compounds, respectively. We standardized the molecular structures of all datasets by protonating acids and de-protonating bases, and converted LD50 values from mg/kg to log(mmol/kg) before model development and performance evaluation.

To examine the distribution of toxic chemicals in the datasets, we used the criteria of the U.S. Environmental Protection Office of Pesticide Programs for Classification and Labeling of Chemical Hazards. Table 1 summarizes these criteria, as well as the distribution of samples among the different toxicity categories for the acute toxicity datasets. It shows that in these datasets, only about 10% or less of the compounds were highly toxic (category I), about 20%–30% had intermediate toxicity (category II), and 60% or more had marginal or no toxicity (categories III and IV). Currently, there is no official classification and labeling criterion for acute injection toxicities. However, the LD50 distributions of the 4 datasets presented in Figure 1 show that the distribution of samples across the mouse intraperitoneal toxicity spectrum is very similar to that of the oral and skin toxicity

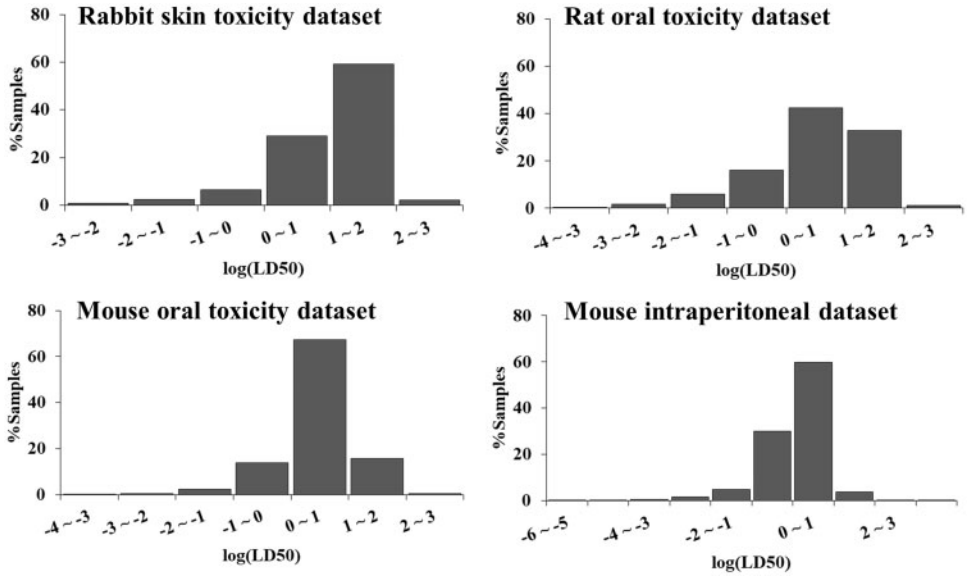

Figure 1. Distribution of samples for the 4 acute chemical toxicity datasets analyzed in this study.

datasets. The LD50 values of most compounds in the 4 datasets are in the 0–2 log(mmol/kg) range.

In Vitro *Molecular Activity Datasets*

To compare and examine parameter choices as well as the behavior of prediction models, we examined 15 *in vitro* molecular activity datasets from the Merck Challenge (Ma *et al.*, 2015). We downloaded the datasets (activity values and molecular descriptors), the provided Python codes for data preprocessing, as well as the published DNN models employing the recommended DNN architecture and hyperparameters. The data and Python codes for these DNN models are available from GitHub (https://github.com/RuwanT/merck; last accessed May 13, 2018); we used these to create the corresponding Merck DNN models for the 15 datasets.

Details of Machine-Learning Methods Used in the Study

Deep neural networks. To develop DNN prediction models, we used the open source Python library Keras (https://keras.io/; last accessed May 13, 2018) on top of Theano (Al-Rfou *et al.*, 2016) backend. We used the mean squared error (MSE) as the loss function for regression, and probed the impact of multiple parameters, dropout rates, optimizers, and initialization methods. Most default parameters in Keras performed satisfactorily. Ultimately, we built all fully connected feed-forward multi-layer neural networks with 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 dataset, we examined the performance of different network architectures, ie, the number of hidden layers and the number of neurons in each hidden layer. In the end, we selected a single architecture to build the DNN models for all of the datasets.

Random forests. To develop RF models, we used the Pipeline Pilot implementation called Forest of Random Trees (http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/; last accessed May 13, 2018). The RF model for each dataset 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 (not mentioned here) are all default parameters of the RF module in Pipeline Pilot. The default parameters worked reasonably well in most test scenarios. Therefore, we used them to develop RF models for all of the datasets studied here.

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

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

In this equation, y_i is the toxicity of the ith nearest neighbor in the training set, d_i is the distance between the ith nearest neighbor and the molecule for which υ-NN is making a prediction, h is a smoothing factor that modulates the distance penalty, and υ 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 2 molecules satisfies the condition. d_0 and h are the only model parameters to be determined from the training data. To predict the toxicity of a compound, υ-NN searches through currently available data to identify all qualified nearest neighbors, and then uses Equation (1) to make a prediction. For a given compound, υ-NN does not give a prediction if there are no qualified nearest neighbors.

Molecular descriptors. As pointed out by Shao *et al.* (2013), input molecular descriptors have a considerable impact on predictive performance. To ensure a fair comparison of different machine-learning methods, the same set of descriptors should be used. However, this is not always possible because some methods are more suitable for certain types of molecular structure representation than others. Here, we used the circular extended connectivity fingerprint with a diameter of 4 chemical bonds (ECFP_4) (Rogers and Hahn, 2010). The Tanimoto coefficient (T_c), calculated with the ECFP_4 fingerprint, is known to be an excellent measure of molecular similarity (Duan *et al.*, 2010). For υ-NN calculations, we used $1 - T_c$ as a metric of the distance between 2

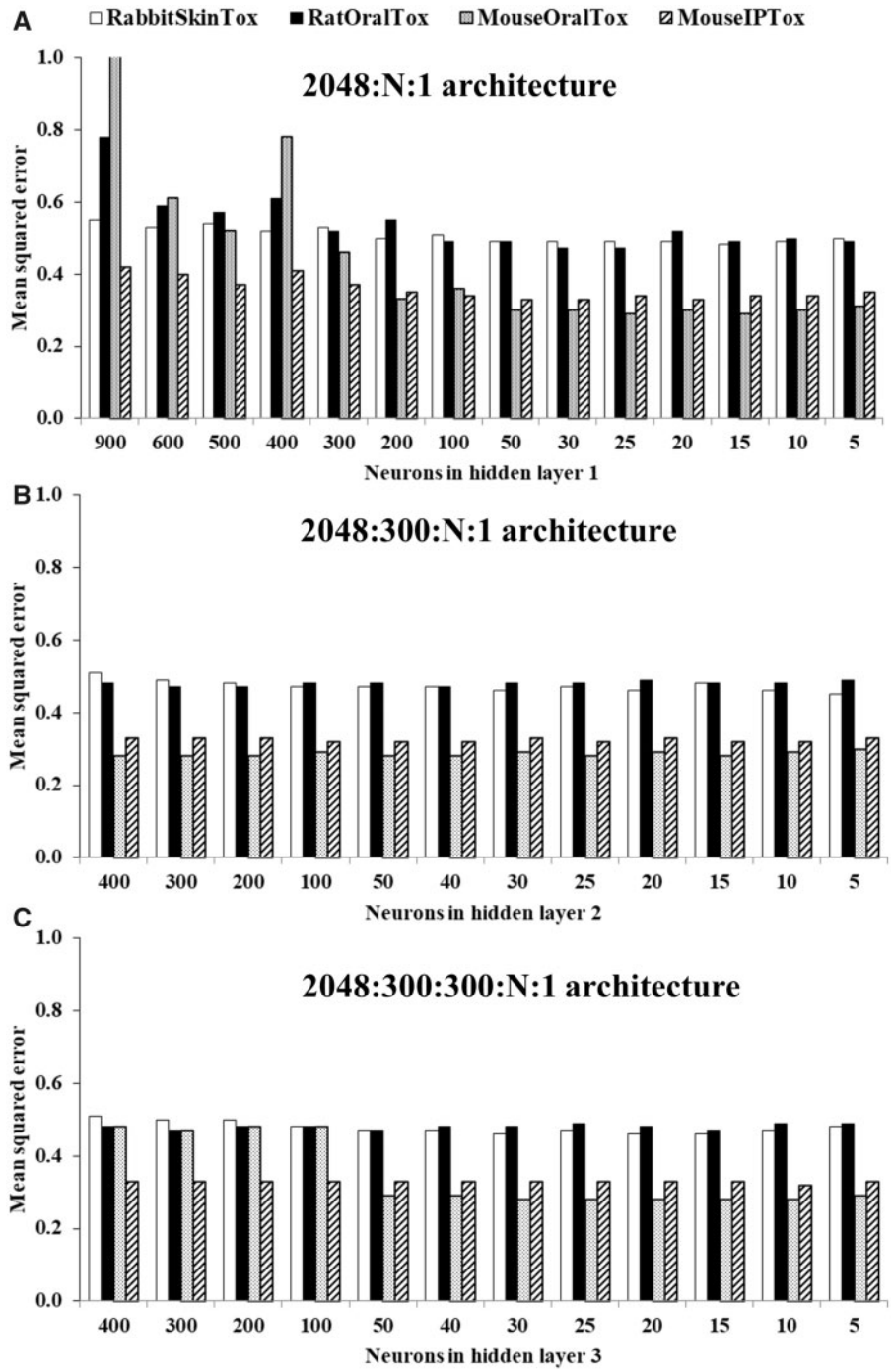

Figure 2. MSEs of increasingly deeper neural network architectures.

molecules. The T_c value between 2 molecules ranges from 0, for 2 molecules not sharing any structural features, to 1, for 2 molecules sharing all structural features. Accordingly, the Tanimoto distance ranges from 1 to 0. On the one hand, because there are thousands to tens of thousands of molecules in the acute toxicity datasets, the number of unique molecular structural features present in the datasets is potentially high. On the other, most molecules in the datasets are relatively small with a limited number of unique structural features. Therefore, the fingerprint of a given molecule is sparsely populated by the unique structural features present in the molecule. To speed up υ-NN computations, we folded the fingerprint to a fixed length of 2048 for Tanimoto distance calculations. Our calculations indicated that fingerprint folding has a negligible impact on the Tanimoto distance, because much longer or shorter fingerprints produce similar Tanimoto distances.

For RF models, we used ECFP_4 fingerprints as input descriptors 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 weights to be determined in a DNN model, the number of input descriptors has a marked

Table 2. MSE and Correlation Coefficients (R and R^2) Between Experimental and DNN-Predicted Log(LD50) Values Derived From 10-Fold Cross Validation of Acute Toxicity Datasets

	Merck DNN Parameters[b]			DNN Parameters This Work[c]		
In vivo datasets[a]	MSE	R	R^2	MSE	R	R^2
Rabbit skin toxicity	0.45	0.56	0.31	0.44	0.54	0.29
Rat oral toxicity	0.39	0.74	0.55	0.37	0.75	0.56
Mouse oral toxicity	0.23	0.68	0.46	0.21	0.69	0.48
Mouse intraperitoneal toxicity	0.26	0.72	0.52	0.23	0.74	0.55
Mouse intravenous toxicity	0.27	0.71	0.50	0.27	0.71	0.50
Rat subcutaneous toxicity	0.67	0.61	0.37	0.64	0.61	0.37
Mouse subcutaneous toxicity	0.46	0.70	0.49	0.45	0.70	0.49

[a]All datasets were downloaded from LeadScope Toxicity Database.
[b]Using Merck-recommended 2048: 4000: 2000: 1000: 1000: 1 architecture and hyperparameters.
[c]Using the selected 2048: 300: 300: 30: 1 architecture and hyperparameters.

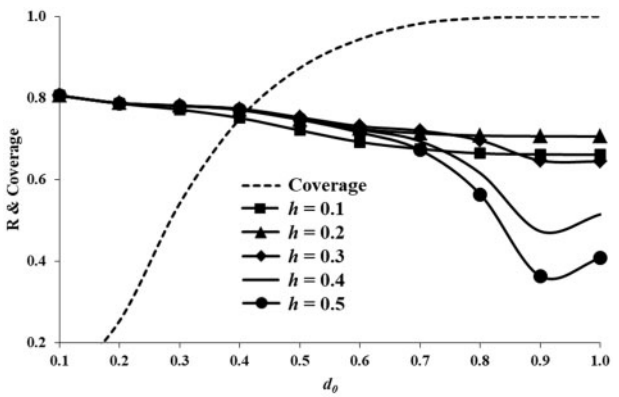

Figure 3. Correlation coefficient (*R*) between *v*-NN-predicted and experimental log(LD50) values and coverage, obtained from 10-fold cross validation calculations using different Tanimoto distance thresholds (d_0) and smoothing factors (h) for the mouse oral toxicity dataset.

impact on computational cost. To ensure maximal comparability between the DNN and *v*-NN approaches, we used a total of 2, 048 ECFP_4 fingerprint features as input descriptors for all datasets in performing DNN calculations. For each dataset, we selected the 2048 fingerprint features according to the following procedure:

1. Identify all unique fingerprint features present in the whole dataset;
2. Calculate the frequency of each fingerprint feature appearing in the molecules in the dataset;
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 de-selects 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 structure, directly from graph representation, led to learned features that were conceptually similar to circular fingerprints (Kearnes *et al.*, 2016).

Fingerprints for in vitro *dataset analysis*. Because Merck did not disclose molecular structures for the 15 *in vitro* datasets, we could not use ECFP_4 fingerprints for the corresponding *v*-NN comparison. To circumvent this problem, we converted the provided atom-pair descriptors into a fingerprint format by encoding only the presence or absence of atom-pairs and ignoring atom-pair counts. This results in loss of information to a certain degree, but allowed us to calculate and use different d_0s for comparison with the *v*-NN predictions.

DNN network architecture. In our initial investigation, we performed a large number of 10-fold cross validation calculations on the rabbit skin, rat oral, mouse oral, and mouse intraperitoneal toxicity datasets to probe the optimal number of hidden layers and hidden neurons. A seemingly counterintuitive observation from these calculations was that neither a larger number of hidden neurons nor a larger number of hidden layers necessarily leads to better neural networks for regression problems. This was previously observed in the Merck study on regression performance (Ma *et al.*, 2015), whereas the opposite was observed for classification studies where deeper and wider networks generally performed better than shallower and narrower ones (Koutsoukas *et al.*, 2017; Lenselink *et al.*, 2017). For the acute toxicity datasets, a single-hidden layer with a small number of hidden neurons performed as well as a DNN model with more hidden layers and a much larger number of hidden neurons. Figure 2 shows the MSEs between DNN-predicted and experimental log(LD50) values in mmol/kg for the 4 representative datasets. Figure 2A displays the results of single–hidden-layer neural networks with different numbers of hidden neurons. For all datasets, there were 2048 neurons in the input layers and only 1 in the output layer—(ie, the predicted log[LD50]). We represent the architecture of the neural networks in Figure 2A by 2408: N: 1, where N is the number of neurons in the hidden layer. Figure 2A shows that for all 4 datasets, neural networks with a very small number of hidden neurons outperformed those with a large number of hidden neurons. When the number of hidden neurons exceeds 300, the MSE can be markedly larger than when it is 300 or less.

Figure 2B displays MSEs for neural networks with a 2048: 300: N: 1 architecture, ie, 300 hidden neurons in the first hidden layer and a varying number of neurons in the second hidden layer. With this architecture, the number of hidden neurons in the second layer had a much smaller impact on prediction performance than the number of hidden neurons in the first layer. Furthermore, neural networks possessing a 2048: 300: N: 1 architecture outperformed those with the 2048: N: 1 architecture, indicating that 2 hidden layers are better than one.

Figure 2C shows MSEs for neural networks with a 2048: 300: 300: N: 1 architecture. For all 4 datasets, a third hidden layer

Table 3. MSE and Correlation Coefficient (*R*) Between the Predicted and Experimental Log(LD50) Values Derived From 10-Fold Cross Validation Calculations With the 3 Machine-Learning Methods for the 4 Acute Toxicity Datasets

Dataset	Rabbit skin toxicity				Rat oral toxicity				Mouse oral toxicity				Mouse intraperitoneal toxicity			
Method	RF	DNN	v-NN ($d_0 = 1.0$)	v-NN ($d_0 = 0.6$)	RF	DNN	v-NN ($d_0 = 1.0$)	v-NN ($d_0 = 0.6$)	RF	DNN	v-NN ($d_0 = 1.0$)	v-NN ($d_0 = 0.6$)	RF	DNN	v-NN ($d_0 = 1.0$)	v-NN ($d_0 = 0.6$)
MSE	0.39	0.45	0.44	0.41	0.36	0.37	0.46	0.35	0.21	0.21	0.27	0.18	0.24	0.23	0.33	0.19
R	0.60	0.54	0.55	0.58	0.76	0.75	0.71	0.77	0.69	0.69	0.66	0.74	0.73	0.74	0.68	0.78
Coverage	100%	100%	100%	75%	100%	100%	100%	86%	100%	100%	100%	94%	100%	100%	100%	95%

Coverage is the fraction of compounds for which predictions are given by the methods employed. DNN, deep neural network; RF, random forest; v-NN, variable nearest neighbor.

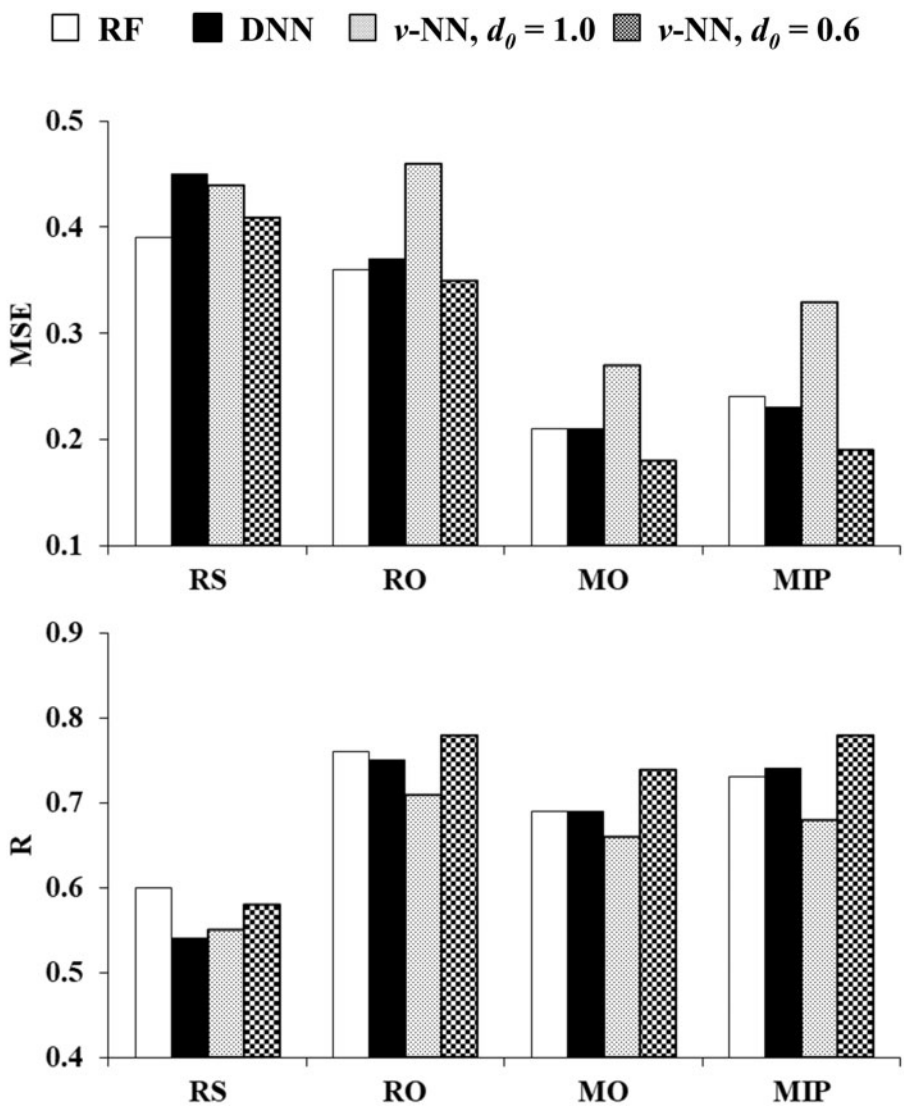

Figure 4. MSE and correlation coefficient (*R*) between the predicted and experimental log(LD50) values derived from 10-fold cross validation calculations using the 3 machine-learning methods for 4 acute toxicity datasets. Detailed data for the figure are provided in Table 3. DNN, deep neural network; RF, random forest; v-NN, variable nearest neighbor; RS, rabbit skin toxicity; RO, rat oral toxicity; MO, mouse oral toxicity; MIP, mouse intraperitoneal toxicity.

provided little to no performance enhancement. Hence, we decided to use the 2048: 300: 300: 30: 1 architecture in building our final DNN models for all 4 datasets. The total number of weights to be determined in the model training process was 713 430 for this architecture.

Comparison of DNN models. In order to examine the choice of hyperparameters, we compared our DNN model results for all 7 *in vivo* acute toxicity datasets with DNN models built using the hyperparameters from the Merck Challenge datasets (Ma *et al.*, 2015). Table 2 shows the correlation coefficients and MSEs

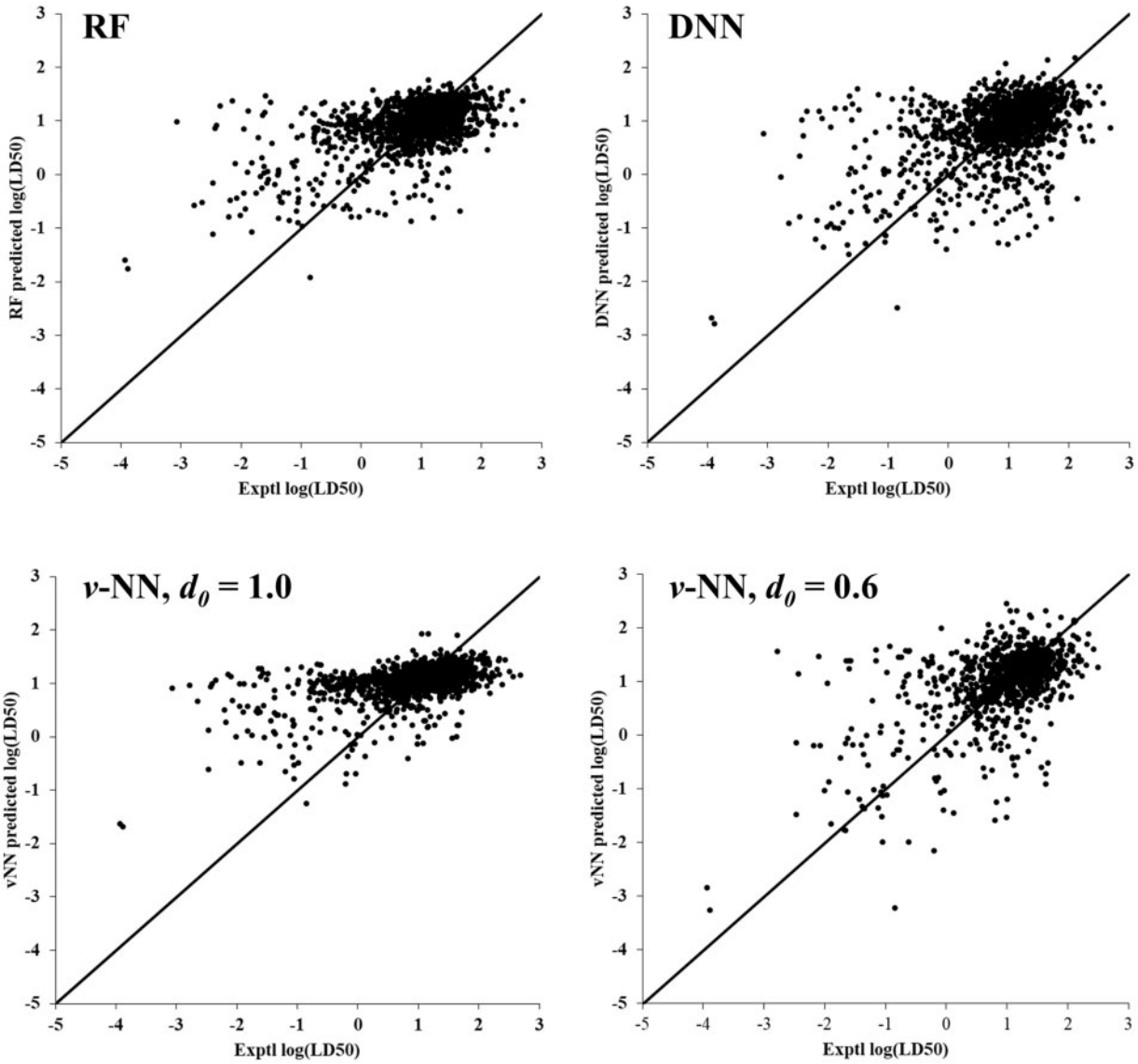

Figure 5. Predicted versus experimental log(LD50) values for the rabbit skin toxicity dataset. The straight line is the identity line with a slope of unity. A data point on this line indicates a perfect prediction. DNN, deep neural network; RF, random forest; v-NN, variable nearest neighbor.

between DNN-predicted and experimental log(LD50)s of the 7 datasets. Even though our DNN architecture (2048: 300: 300: 30: 1) is shallower and narrower than that of Merck's (2048: 4000: 2000: 1000: 1000: 1), the results are nearly the same, indicating that our DNN models are robust with respect to these variations.

v-NN model parameters. In contrast to the large number of weights to be determined for a DNN model, the total number of parameters for a v-NN model is only 2: the Tanimoto distance threshold d_0, and the smoothing factor h in Equation (1). To determine their optimal values, we performed 10-fold cross validation calculations with all 4 datasets, using different combinations of d_0 and h. Figure 3 shows the results obtained with the mouse oral toxicity dataset. The results obtained with the other datasets are similar and therefore not presented here. Figure 3 shows the correlation coefficient between the v-NN-predicted and experimental log(LD50) values, obtained with different values of h, plotted against d_0. It also plots coverage (ie, the fraction of compounds for which v-NN can make predictions) as a function of d_0. The results indicate that h has a smaller impact on prediction performance than does d_0, and that a value of 0.2 or 0.3 for h gives the best performance. Based on these results, we set h to 0.3 for all v-NN calculations used in this study.

RESULTS AND DISCUSSION

Mean Metrics

To compare the predictions of the DNN, RF, and v-NN models, we performed 10-fold cross validation calculations with each of the 3 machine-learning methods for the 4 datasets. Because both the DNN and RF models gave predictions for all input molecules, for the sake of comparison, we also set d_0 to 1.0 to provide v-NN predictions for every input molecule. As shown in Equation (1), the prediction is simply a Tanimoto distance-weighted average of all training set samples. In addition, we also made v-NN predictions by setting d_0 to 0.6. Figure 3 shows that at this Tanimoto distance threshold, the method

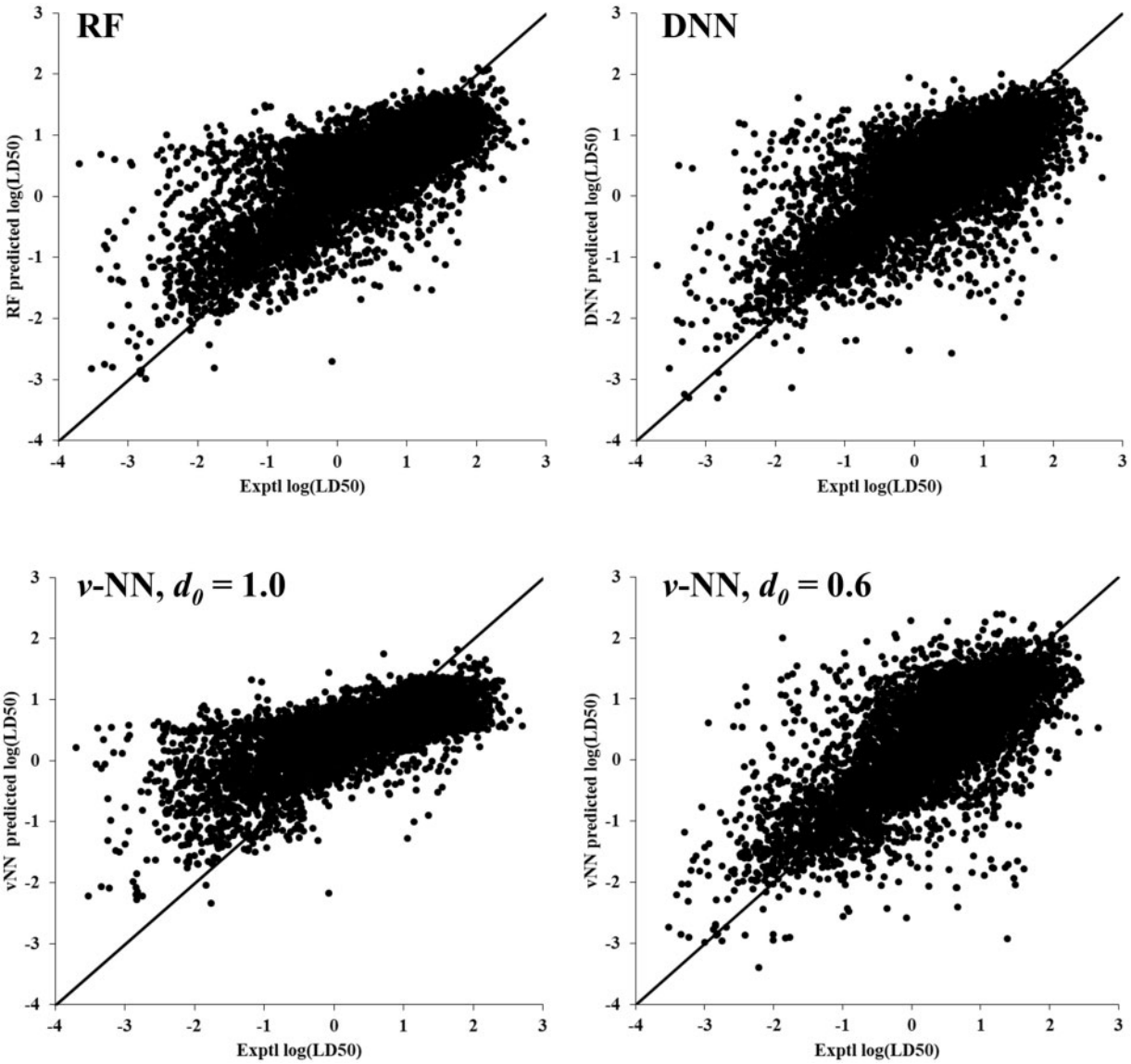

Figure 6. Predicted versus experimental log(LD50) values for the rat oral toxicity dataset. The straight line is the identity line with a slope of unity. A data point on this line indicates a perfect prediction. DNN, deep neural network; RF, random forest; υ-NN, variable nearest neighbor.

demonstrated both reasonably high coverage and reasonably good prediction performance, as measured by the correlation coefficient, R, between the predicted and experimental values.

Table 3 and Figure 4 show the MSEs and *R* values between the predicted and experimental values derived from 10-fold cross validation calculations for the 4 datasets. Of the 3 models giving predictions for all compounds, the RF model showed the lowest MSE for the rabbit skin and rat oral toxicity datasets, an MSE comparable to that of the DNN model for the mouse oral toxicity dataset, and an MSE slightly higher than that of the DNN model for the mouse intraperitoneal toxicity dataset. The worst performer was the υ-NN model with a Tanimoto distance threshold of 1.0. However, setting the Tanimoto distance threshold to 0.6 considerably improved the performance of the υ-NN model, as indicated by both the MSE and *R* values, especially for larger datasets. In fact, none of the methods performed well for the smallest dataset, and all tended to improve in performance as the dataset size increased. The results shown in Table 3 and Figure 4 corroborate a general observation that RF models outperform DNN models for small datasets, but DNN models improve with dataset size and outperform RF models for larger datasets (mouse intraperitoneal toxicity in this study). The requirement of large datasets arises from the large number of parameters that need to be optimized in DNN models.

Detailed Analysis of Regression Predictions

As we stated in the introduction, most studies evaluating machine-learning methods with a large number of datasets rely on global performance metrics. However, performance measures calculated from an entire dataset may obscure the finer details of prediction performance for datasets with an uneven distribution of samples. Figures 5–8 display the calculated versus predicted log(LD50) values derived from 10-fold cross validation for all of the datasets. The predictions all showed a clear tendency to underestimate toxicity for highly toxic compounds, ie, the predicted log(LD50) values were considerably higher than

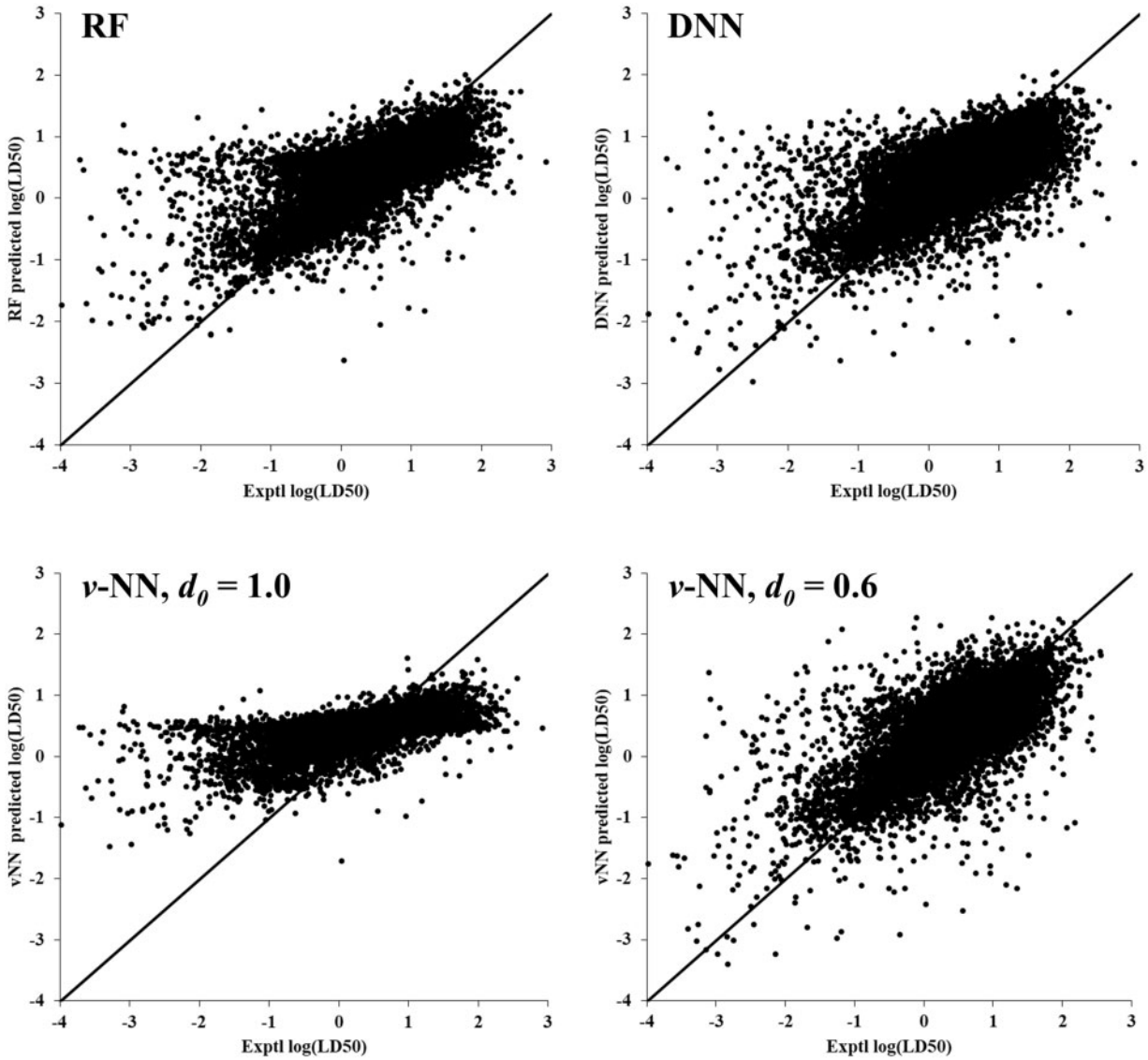

Figure 7. Predicted versus experimental log(LD50) values for the mouse oral toxicity dataset. The straight line is the identity line with a slope of unity. A data point on this line indicates a perfect prediction. DNN, deep neural network; RF, random forest; υ-NN, variable nearest neighbor.

the experimental values when the latter were < –1.0. Conversely, all of the predictions overestimated the toxicity of nontoxic compounds, ie, the predicted log(LD50) values were lower than the experimental values when the latter were >1.0. This trend is best demonstrated by the υ-NN results derived with a Tanimoto distance threshold of 1.0.

The sample distributions shown in Figure 1 demonstrate that for all 4 datasets, a majority of the samples are in the –1.0 to 1.0 log(LD50) range. Regardless of the method used, all predict most of the compounds to be marginally toxic, because they underestimate the toxicity of many highly toxic compounds and overestimate that of nontoxic compounds—a trend that severely limits the utility of machine learning for this application. In this context, reducing the Tanimoto distance threshold from 1.0 to 0.6 markedly improved the performance of the υ-NN models. Their predictions, which were more symmetrically distributed along the identity line in Figures 5–8, were even better than those of the RF and DNN models. The trade-off is a lower coverage of 75%–95% depending on the dataset size.

Distribution of Prediction Errors

To examine the performance deterioration for highly toxic compounds in more detail, we grouped the compounds into intervals of 1 log unit along the experimentally measured toxicity ranges, and calculated the MSE between the predicted and experimental values for compounds in each interval. Figure 9 shows the resulting percentages of compounds in log(LD50) unit intervals. All 3 methods gave extremely good predictions for compounds in the most populated intervals. As the number of samples in a toxicity interval decreased; however, the predictions worsened; that is, the MSE was inversely correlated with the fraction of samples in the toxicity interval. This trend is a result of the optimization criterion used to determine the model parameters; when the criterion is designed to minimize overall prediction errors (eg, MSE, RMSE), the most efficient way to achieve it is to minimize the errors of groups with the most samples. In this sense, a dataset of uniformly distributed samples is critical for developing high-performance models. This does not pose a problem for some tasks, such as image

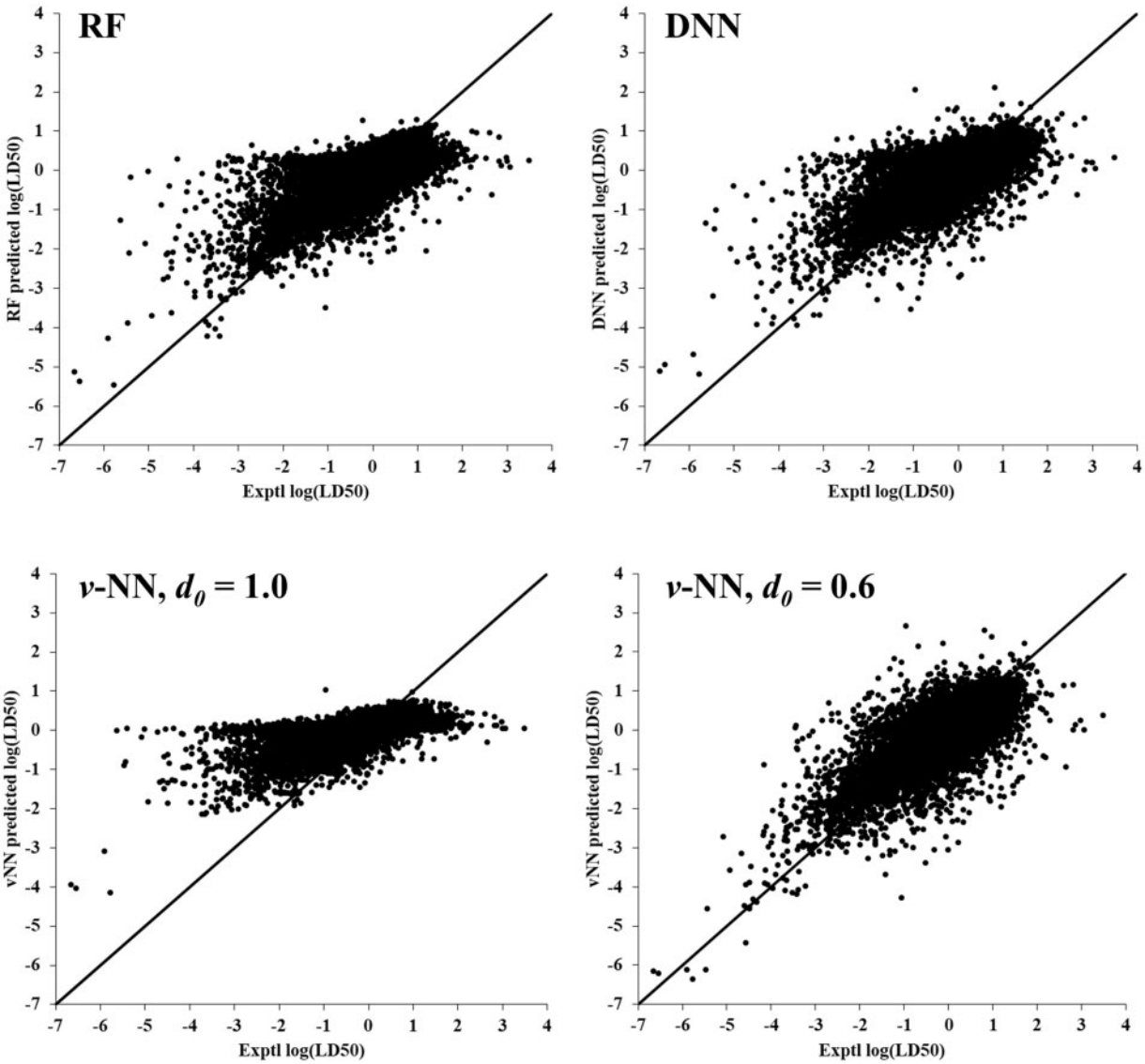

Figure 8. Predicted versus experimental log(LD50) values for the mouse intraperitoneal toxicity dataset. The straight line is the identity line with a slope of unity. A data point on this line indicates a perfect prediction. DNN, deep neural network; RF, random forest; υ-NN, variable nearest neighbor.

recognition; if an object is under-represented in an image dataset, one can always supply more images of that object to enhance its representation. However, this approach is not feasible for toxicity datasets because the number of highly toxic compounds is small relative to that of all compounds available or tested. An analogous case is the search for potent compounds at a drug target, a process vividly likened to looking for a needle in a haystack. No matter how the chemicals are selected, as long as they are structurally diverse with previously unknown toxicity, toxicity tests will demonstrate that most of the compounds are marginally toxic and only a small number are highly toxic. When all available compounds with known toxicities are used to build a prediction model, the resulting model will inevitably be biased toward the majority class of nontoxic or marginally toxic samples and underperform for other samples.

Effect of Resampling on DNN Performance

Given the highly uneven sample distribution, in which marginally toxic compounds greatly outnumber highly toxic ones, we further examined the impact of under-sampling of compounds with marginal toxicity as a means to achieve a relatively even sample distribution.

To achieve a relatively even sample distribution, we grouped the samples within each half-log-LD50 increment along the toxicity spectrum. We then selected a maximum of 500 compounds by structural diversity (as measured by ECFP_4 fingerprints) from each group to make up the resampled training sets. Given that the number of highly toxic compounds is extremely small and most of the compounds have marginal toxicity in all of the datasets (Figure 1), the resampled training sets have far fewer compounds than the full datasets. Because the rabbit skin toxicity dataset was too small, we performed under-sampling only for the 3 larger datasets. This resulted in 3914, 3704, and 4233 compounds for the rat oral, mouse oral, and mouse intraperitoneal datasets, which are considerably lower than the original sizes of 10 363, 21 776, and 29 476, respectively.

Figure 10 presents the results of 10-fold cross validation using the DNN method for the 3 under-sampled datasets. It shows that most highly toxic compounds were still predicted to be less

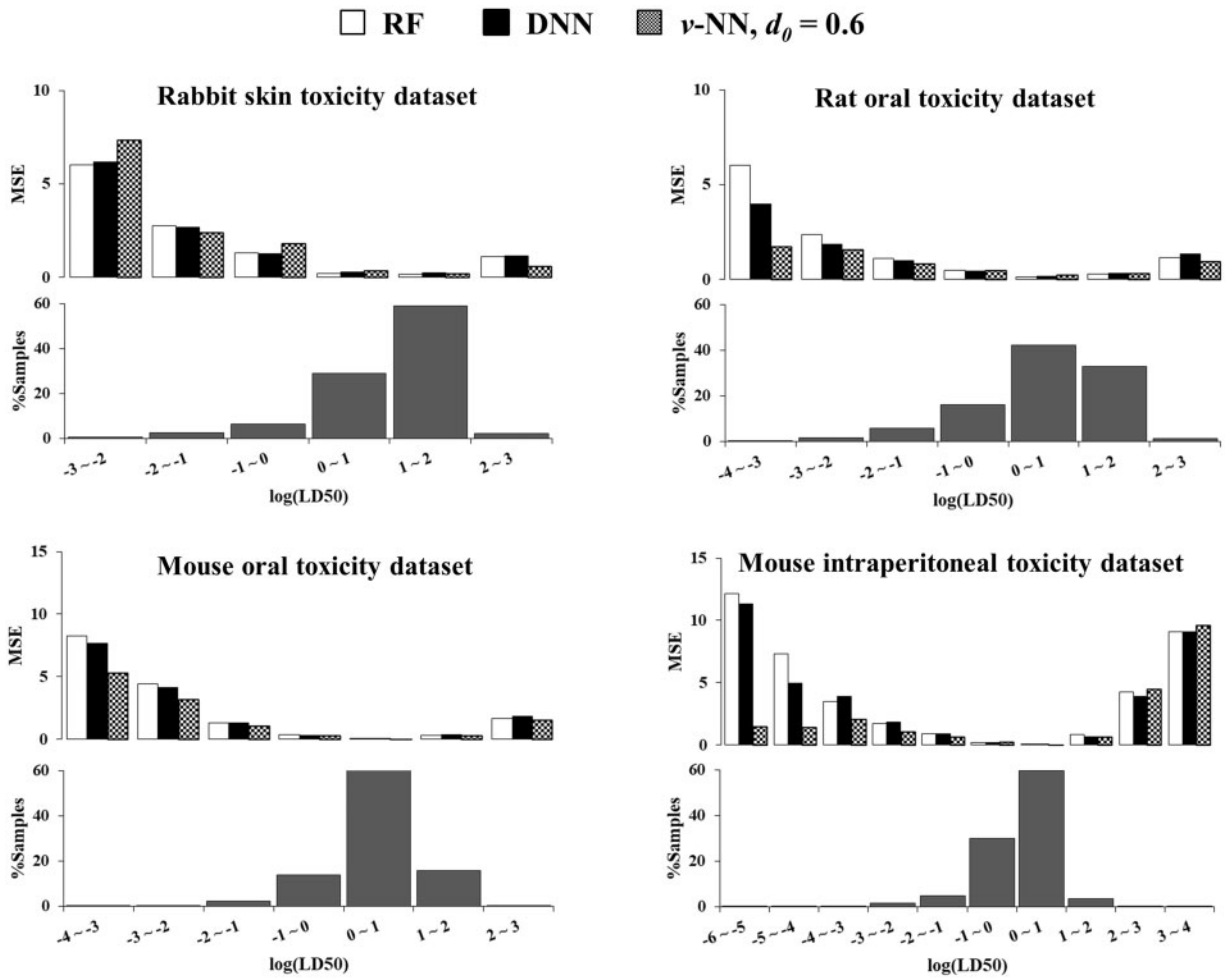

Figure 9. MSE between predicted and experimental log(LD50) values and the percentage of compounds in each log unit interval. DNN, deep neural network; RF, random forest; υ-NN, variable nearest neighbor.

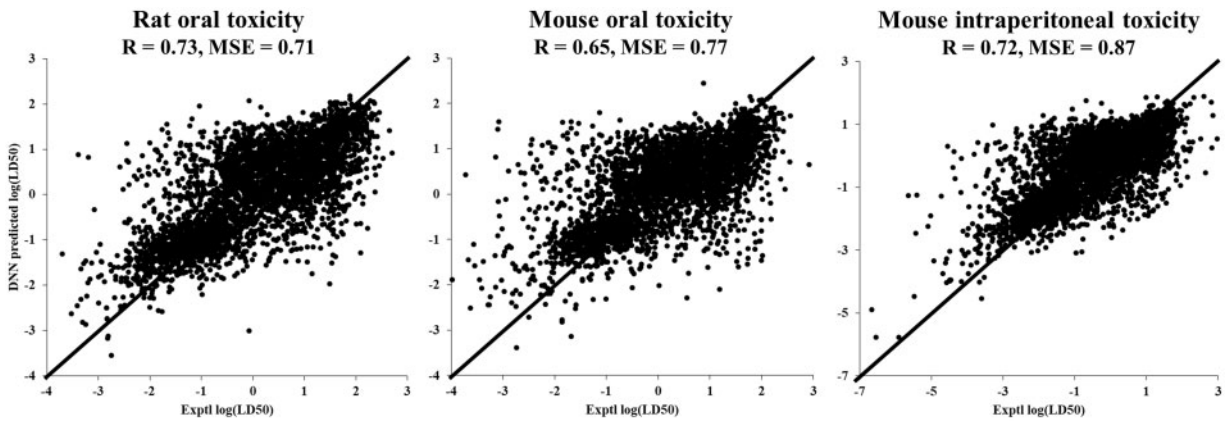

Figure 10. Experimental versus DNN-predicted log(LD50)s derived from 10-fold cross validation using re-sampled datasets. The straight line is the identity line with a slope of unity. A data point on this line indicates a perfect prediction.

toxic than their measured values, and the MSE values were 2 to 3 times larger than those of the original datasets (Table 3), primarily because the performance for marginally toxic compounds deteriorated markedly. Thus, the under-sampling approach resulted in a considerable reduction in the number of training samples, and did not improve the model performance for highly toxic compounds, but rather resulted in an overall deterioration in performance.

Global Versus Local Models

A contentious topic in QSAR research is whether to use a global or a local model approach (Helgee *et al.*, 2010; Sheridan, 2014;

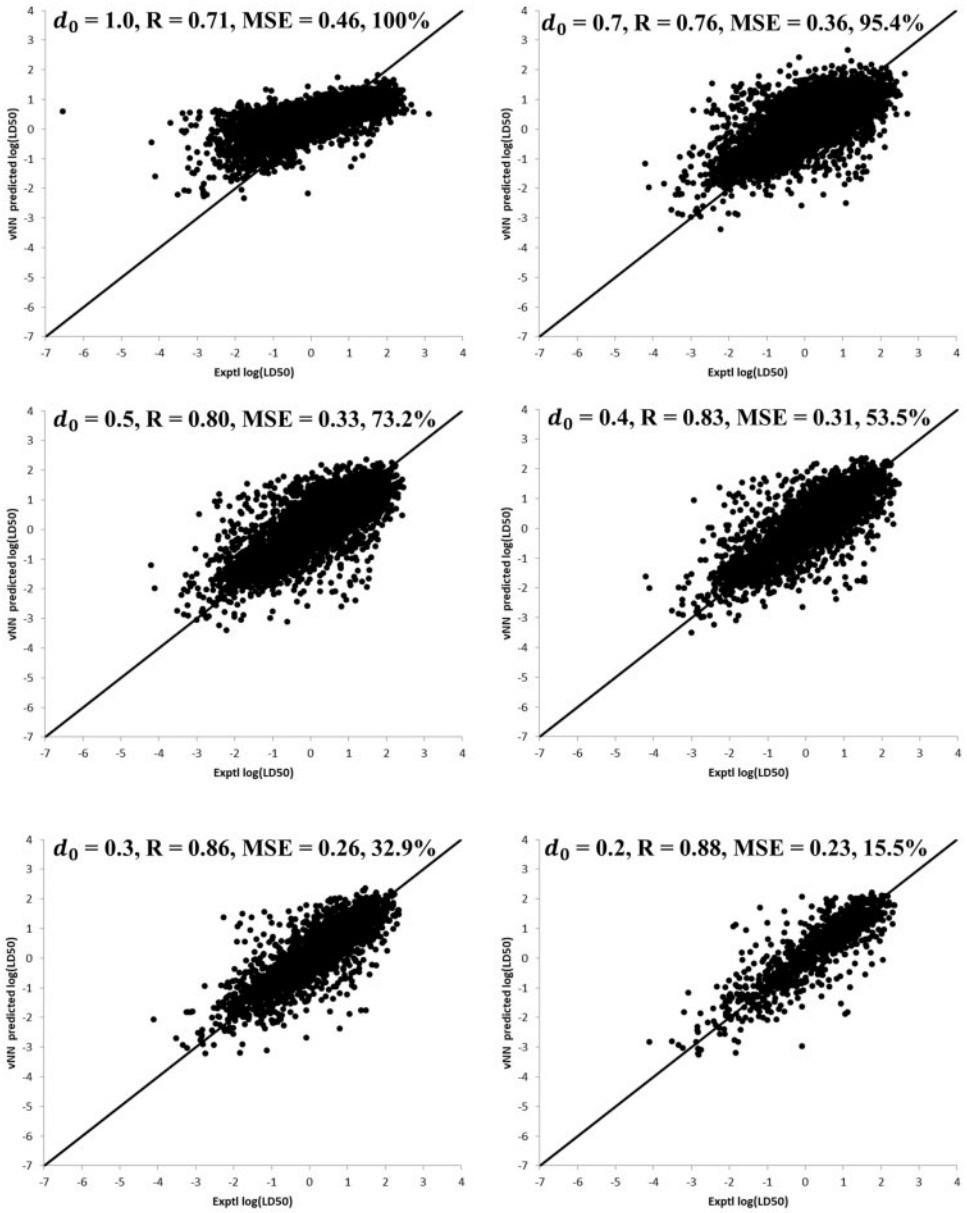

Figure 11. υ-NN-predicted versus experimental log(LD50) values for the rat oral toxicity dataset. We made the predictions with different values for the Tanimoto distance threshold (d_0), which resulted in differences in coverage (%), correlation coefficient (R), and MSE.

Yuan *et al.*, 2007). A global model approach uses all available samples to build a single model that is supposed to be applicable to all compounds. In contrast, a local model approach builds multiple models based on the structural similarity of training molecules, and makes predictions using the models most appropriate for the target structures (Yuan *et al.*, 2007). In principle, local models can be reasonably expected to give more accurate predictions. However, large training sets with well-formed structural clusters are required to develop a sufficient number of reliable local models. In essence, the υ-NN method can be considered a local approach, because it is designed to use only information of compounds in the training set that are structurally similar to the target compounds in making predictions. The hypothesis that local models can provide more accurate predictions is corroborated by the results derived from our υ-NN calculations using different values for the Tanimoto distance threshold (d_0). As shown in Figures 5–8, reducing d_0 from 1.0 to 0.6 led to a small reduction in coverage and a marked improvement in prediction performance. To fully understand the impact of d_0 on prediction performance, we performed 10-fold cross validation calculations using different values for d_0. Figure 11 shows the results for the rat oral toxicity dataset. The results for the 4 datasets examined in detail as well as for the 3 additional datasets are similar (Supplementary Figs. 1–4). They indicate successive improvement in prediction performance with successively lower values of d_0, albeit with successively reduced coverage. The loss of coverage was severe for the smallest datasets, but small datasets were also problematic for the DNN and RF models (Figs. 5–8).

In Vitro *Molecular Activity Models*

We further examined model performance for the Merck Challenge data consisting of 15 *in vitro* datasets (Ma *et al.*, 2015). Table 4 shows the RMSE values between DNN-predicted and

Table 4. RMSE[a] of DNN[b] and υ-NN Predictions for the Merck Challenge Test Set Compounds

Dataset	Molecules[c]	Published[d]	Posted[e]	This Work[f]	υ-NN ($d_0 = 1.0$)[g]	υ-NN ($d_0 = 0.3$)[h]
METAB	2029	**21.78**	23.19	22.89	27.43	23.40
HIVINT	2421	**0.44**	0.47	0.47	0.50	**0.38**
HIVPROT	4311	1.66	1.60	**1.52**	**1.14**	**0.95**
TDI	5559	0.40	0.41	**0.38**	0.46	**0.38**
THROMBIN	6924	2.04	2.10	**1.90**	**1.76**	**1.48**
OX1	7135	**0.73**	0.81	0.91	1.20	1.15
RAT_F	7821	**0.54**	0.55	0.55	0.62	**0.50**
DPP4	8327	**1.30**	1.68	1.45	**1.25**	**1.16**
PGP	8603	**0.36**	0.38	0.38	0.46	0.37
PPB	11622	**0.56**	0.57	0.57	0.76	0.62
CB1	11640	1.25	**1.21**	1.23	1.56	**0.99**
NK1	13482	**0.76**	**0.76**	0.77	0.79	0.77
OX2	14875	0.95	**0.93**	1.00	1.39	1.17
3A4	50000	**0.48**	0.50	0.49	**0.46**	**0.37**
LOGD	50000	**0.51**	0.51	0.55	0.94	0.72

[a]Among the RMSE values of the 3 DNN implementations, the lowest one for each dataset is indicated by a boldface value; the lowest RMSE of DNN and υ-NN predictions for each dataset is indicated by a underlined boldface value.
[b]With Merck-recommended DNN architecture and hyperparameters.
[c]Number of molecules in each dataset, including molecules in both the training (75%) and test sets (25%).
[d]Published RMSE of Merck DNN models.
[e]RMSE of Merck DNN models posted on GitHub.
[f]RMSE of our implementation of Merck DNN.
[g]RMSE of υ-NN calculated with a Tanimoto distance threshold of 1.0 for 100% coverage.
[h]RMSE of υ-NN calculated with a Tanimoto distance threshold of 0.3 for improved performance but with reduced coverage.

experimental activities of the test set compounds as reported in their paper, the results of reimplementing their published DNN models posted on GitHub, and the results of our implementation of their DNN models. Owing to the nature of the stochastic gradient descent optimization of network weights, each implementation of the same DNN model may result in a slightly different set of model parameters (weights). Nevertheless, the RMSE values of all 3 implementations are in overall agreement, with the published results having the lowest value for 7 datasets, the reimplementation posted on GitHub with the lowest value for 3 datasets (including one tie with the published results), and our implementation with the lowest value for 3 datasets. The observed minor differences—not unexpected because of the stochastic gradient decent optimization method used—are materially insignificant.

We also performed υ-NN calculations with different Tanimoto distance thresholds, d_0, on the same datasets using atom-pair fingerprints (see Materials and Methods section). We made υ-NN predictions for the test set compounds with d_0 set to 1.0 (for 100% coverage) or 0.3 (for reduced coverage and better performance). The RMSEs of the test set compounds obtained with d_0 set to 1.0 or 0.3 were compared with those of the DNN models in Table 4. Judging from the global performance metric RMSE alone, the υ-NN models with d_0 set to 1.0 for 100% coverage performed better than expected, as they achieved the lowest RMSEs compared with the DNN models, for 4 of the 15 datasets. Because a υ-NN prediction with d_0 set to 1.0 represents a distance-weighted average of all training molecule activities, the improvement is more a reflection of the unsatisfactory performance of the DNN method than the superiority of the υ-NN method with d_0 set at 1.0. Although employing a reduced d_0 of 0.3 markedly improved the RMSE of the υ-NN method, this came at the expense of reduced coverage, ie, predictions were not given for compounds without qualified neighbors in the training sets.

In addition to the global performance metrics, we also examined details of the predicted versus experimental results at different activity ranges. Supplementary Figure 5 shows the predicted activity plotted against the experimental activity for the test set compounds in the 15 datasets. The plots show that for 10 of the datasets, DNN predictions were clearly "compressed," ie, they underestimated the activity of highly active compounds and overestimated that of inactive compounds, resulting in predictions with an activity range narrower than that for the experimental results. This is the same trend we observed in DNN predictions of the acute *in vivo* toxicity datasets. Figure 12 shows the actual versus predicted values for a few extreme examples of *in vitro* molecular activity endpoints: time-dependent 3A4 inhibition (TDI), log(rat bioavailability) at 2 mg/kg (RATF_F), inhibition of dipeptidyl peptidase 4 (DPP4), and orexin 2 receptor (OX2) inhibition. In summary, our calculations using Merck DNN models on Merck *in vitro* datasets showed a pattern of DNN predictions similar to that observed in the *in vivo* animal toxicity data—underestimation of the potency of highly active compounds, regardless of DNN model construction or choice of hyperparameters.

SUMMARY

Highly unbalanced datasets pose a significant challenge for machine-learning methods in developing classification models (Krawczyk, 2016). However, the impact of such datasets on regression model performance has not been carefully investigated. Most studies comparing the performance of machine-learning methods on regression problems rely on the overall deviation between the predicted and experimental values of all samples, as measured by the RMSE or MSE, without considering the appropriateness of using an overall performance metric for datasets with a highly uneven distribution of samples. In this study, we examined the performance of 3 machine-learning methods for 7 acute toxicity datasets varying in size and having

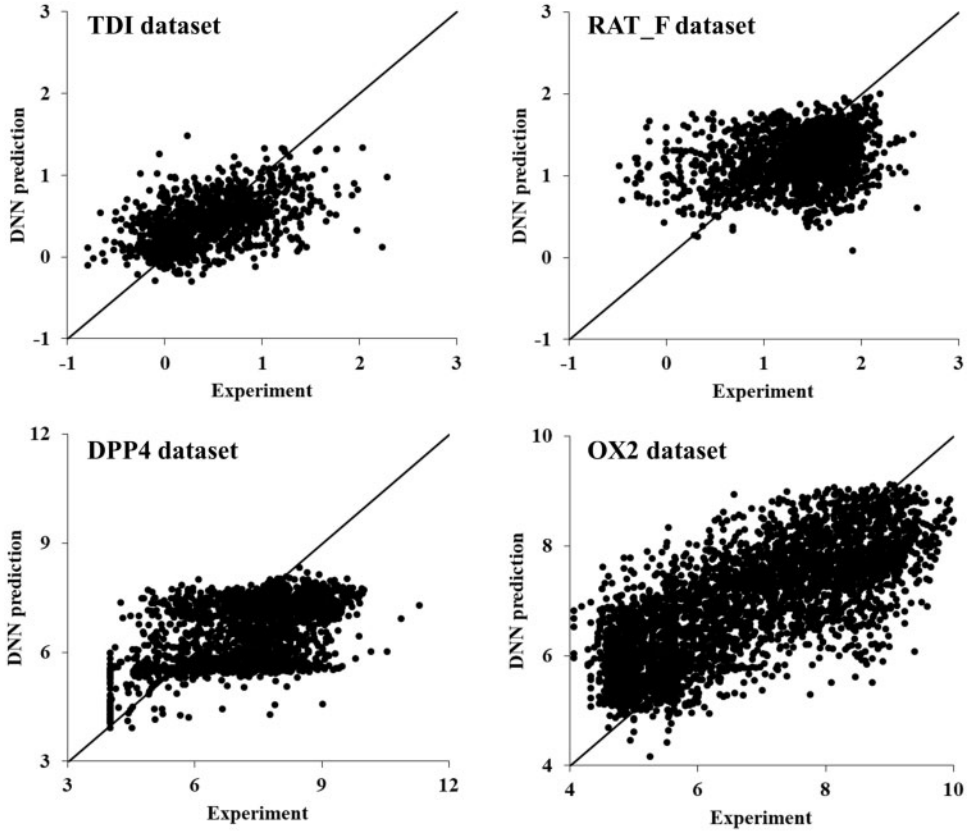

Figure 12. Predicted versus experimental activity values of the test set compounds for the *in vitro* molecular activity endpoints of TDI, log(rat bioavailability) at 2 mg/kg (RATF_F), inhibition of DPP4, and inhibition of OX2, using the Merck-recommended deep neural network models. All prediction models show severe underestimation of activity for highly active compounds and overestimation of activity for inactive compounds.

highly uneven distributions of samples across the toxicity spectrum. We demonstrated that a highly unbalanced dataset poses at least as big a challenge for regression as it does for classification. With a highly unbalanced dataset, the performance of models trained on all available samples is highly biased toward the most populous samples on the activity spectrum and against samples with the highest and lowest activities.

This bias presents a significant challenge for applications in predictive toxicology and drug discovery. Most datasets in these research areas are highly unbalanced, because highly active compounds for any desired target are rare and most compounds are marginally active or inactive. This challenge must first be surmounted before predictive toxicology can reliably identify highly toxic compounds. Because highly toxic compounds are rare relative to all compounds tested, models trained with all available data may make markedly worse predictions for highly toxic compounds than some overall performance metrics suggest.

In light of the achievements of DNNs in artificial intelligence, we expected them to perform better than shallow learning methods. To our surprise, the DNNs tested here did not show better performance for developing regression models. A plausible argument against this observation is that the datasets we used were too small, given that DNNs perform like shallow learning methods for small datasets but outperform other methods for large datasets. However, owing to the cost of testing chemicals, high-quality chemical datasets with more than tens of thousands of compounds are rare or not publically available.

Counterintuitively, the υ-NN approach—the simplest method among the 3 studied here—performed reasonably well in quantitative prediction of the toxicity of highly potent compounds. Because a υ-NN model does not provide predictions for compounds without qualified nearest neighbors in the training set, a plausible explanation for its better than expected performance is that the Tanimoto distance threshold defines a good applicability domain for the υ-NN model, and by excluding predictions for compounds outside this domain, the overall performance for the remaining compounds is elevated. This suggests that the Tanimoto similarity may serve as a basis for .defining a high-performance applicability domain, which is the subject of an ongoing study

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

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 Material Command (Ft Detrick, Maryland) 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 article has been approved for public release with unlimited distribution.

REFERENCES

Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., *et al.* (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv: 1605.02688v1*.

Ball, N., Bartels, M., Budinsky, R., Klapacz, J., Hays, S., Kirman, C., and Patlewicz, G. (2014). The challenge of using read-across within the EU REACH regulatory framework; how much uncertainty is too much? Dipropylene glycol methyl ether acetate, an exemplary case study. *Regul. Toxicol. Pharmacol.* **68**, 212–221.

Ball, N., Cronin, M. T., Shen, J., Blackburn, K., Booth, E. D., Bouhifd, M., Donley, E., Egnash, L., Hastings, C., Juberg, D. R., *et al.* (2016). Toward Good Read-Across Practice (GRAP) guidance. *Altex* **33**, 149–166.

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

Burden, N., Sewell, F., and Chapman, K. (2015). Testing chemical safety: What is needed to ensure the widespread application of non-animal approaches? *PLoS Biol.* **13**, e1002156.

Duan, J., Dixon, S. L., Lowrie, J. F., and Sherman, W. (2010). Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph Model* **29**, 157–170.

Helgee, E. A., Carlsson, L., Boyer, S., and Norinder, U. (2010). Evaluation of quantitative structure-activity relationship modeling strategies: Local and global models. *J. Chem. Inf. Model* **50**, 677–689.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: Moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608.

Koutsoukas, A., Monaghan, K. J., Li, X., and Huan, J. (2017). Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform.* **9**, 42.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **5**, 221–232.

Lenselink, E. B., ten Dijke, N., Bongers, B., Papadatos, G., van Vlijmen, H. W. T., Kowalczyk, W., IJzerman, A. P., and van Westen, G. J. P. (2017). Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* **9**, 45.

Liu, R., Tawa, G., and Wallqvist, A. (2012). Locally weighted learning methods for predicting dose-dependent toxicity with application to the human maximum recommended daily dose. *Chem. Res. Toxicol.* **25**, 2216–2226.

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

National Academy of Sciences. (2015). In *Application of Modern Toxicology Approaches for Predicting Acute Toxicity for Chemical Defense*. National Academies Press, Washington, DC.

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

Shao, C. Y., Chen, S. Z., Su, B. H., Tseng, Y. J., Esposito, E. X., and Hopfinger, A. J. (2013). Dependence of QSAR models on the selection of trial descriptor sets: A demonstration using nanotoxicity endpoints of decorated nanotubes. *J. Chem. Inf. Model* **53**, 142–158.

Sheridan, R. P. (2014). Global quantitative structure-activity relationship models vs selected local models as predictors of off-target activities for project compounds. *J. Chem. Inf. Model* **54**, 1083–1092.

Taylor, K., Stengel, W., Casalegno, C., and Andrew, D. (2014). Experiences of the REACH testing proposals system to reduce animal testing. *Altex* **31**, 107–128.

Vogelgesang, J. (2002). The EC white paper on a strategy for a future chemicals policy. *Altern. Lab. Anim.* **30**(Suppl 2), 211–212.

Yuan, H., Wang, Y., and Cheng, Y. (2007). Local and global quantitative structure-activity relationship modeling and prediction for the baseline toxicity. *J. Chem. Inf. Model* **47**, 159–169.