ACS OMEGA

Article

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

Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do

Linlin Zhao,[†] Wenyi Wang,[†] Alexander Sedykh,[*,‡] and Hao Zhu[*,†,§]

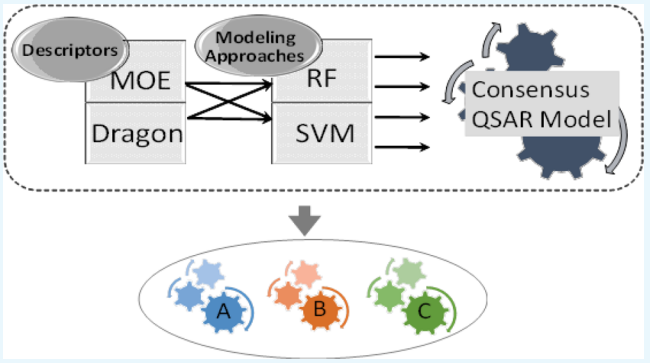
[†]The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, United States

[‡]Sciome LLC, Durham, North Carolina 27709, United States

[§]Department of Chemistry, Rutgers University, Camden, New Jersey 08102, United States

Supporting Information

ABSTRACT: Numerous chemical data sets have become available for quantitative structure−activity relationship (QSAR) modeling studies. However, the quality of different data sources may be different based on the nature of experimental protocols. Therefore, potential experimental errors in the modeling sets may lead to the development of poor QSAR models and further affect the predictions of new compounds. In this study, we explored the relationship between the ratio of questionable data in the modeling sets, which was obtained by simulating experimental errors, and the QSAR modeling performance. To this end, we used eight data sets (four continuous endpoints and four categorical endpoints) that have been extensively curated both in-house and by our collaborators to create over 1800 various QSAR models. Each data set was duplicated to create several new modeling sets with different ratios of simulated experimental errors (i.e., randomizing the activities of part of the compounds) in the modeling process. A fivefold cross-validation process was used to evaluate the modeling performance, which deteriorates when the ratio of experimental errors increases. All of the resulting models were also used to predict external sets of new compounds, which were excluded at the beginning of the modeling process. The modeling results showed that the compounds with relatively large prediction errors in cross-validation processes are likely to be those with simulated experimental errors. However, after removing a certain number of compounds with large prediction errors in the cross-validation process, the external predictions of new compounds did not show improvement. Our conclusion is that the QSAR predictions, especially consensus predictions, can identify compounds with potential experimental errors. But removing those compounds by the cross-validation procedure is not a reasonable means to improve model predictivity due to overfitting.



INTRODUCTION

Quantitative structure−activity relationship (QSAR) models are statistical models, which build correlations between the chemical structure information (represented by a set of molecular descriptors) of compounds and their target biological activities.[1] The data sets for QSAR modeling, which contain the structure information and activities of compounds, are generated by experimental scientists and available in various data sources. Along with the large chemical library and high-throughput screening technologies being developed, numerous data sets have become available for modelers.[2] Popular data sources include general data deposit portals, such as PubChem (http://pubchem.ncbi.nlm.nih.gov), and databases for specific research interests, such as Toxicity ForeCaster (ToxCast) (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data) and ACuteTox (http://www.acutetox.eu/). However, the quality of data may be different based on the nature of experimental protocols. The usefulness of public data sources is questionable due to lack of the necessary quality control.[3] General concerns have been raised regarding irreproducible experimental data,[4−6] which is relatively common in complex biological testing (e.g., animal models).

The major issues existing in the public data sources include (1) the incorrect representation of chemical structures (i.e., structural errors) and (2) inaccurate activity information (i.e., experimental errors). There have been many relevant works showing that noncurated chemical structures will result in models of poor accuracy and the curation of chemical structures will improve modeling predictivity.[7,8] The recent review[9] by Fourches et al. indicates a standardized workflow can be used to greatly decrease the structural errors in the public data sets. However, besides the chemical structure information, the quality of QSAR models also strongly depends on the target biological data. Because of the inevitable experimental errors, it is hard to know which compounds in the modeling set contain incorrect experimental data. Reliable biological data in data sets

Received: March 8, 2017
Accepted: April 27, 2017
Published: June 19, 2017

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812

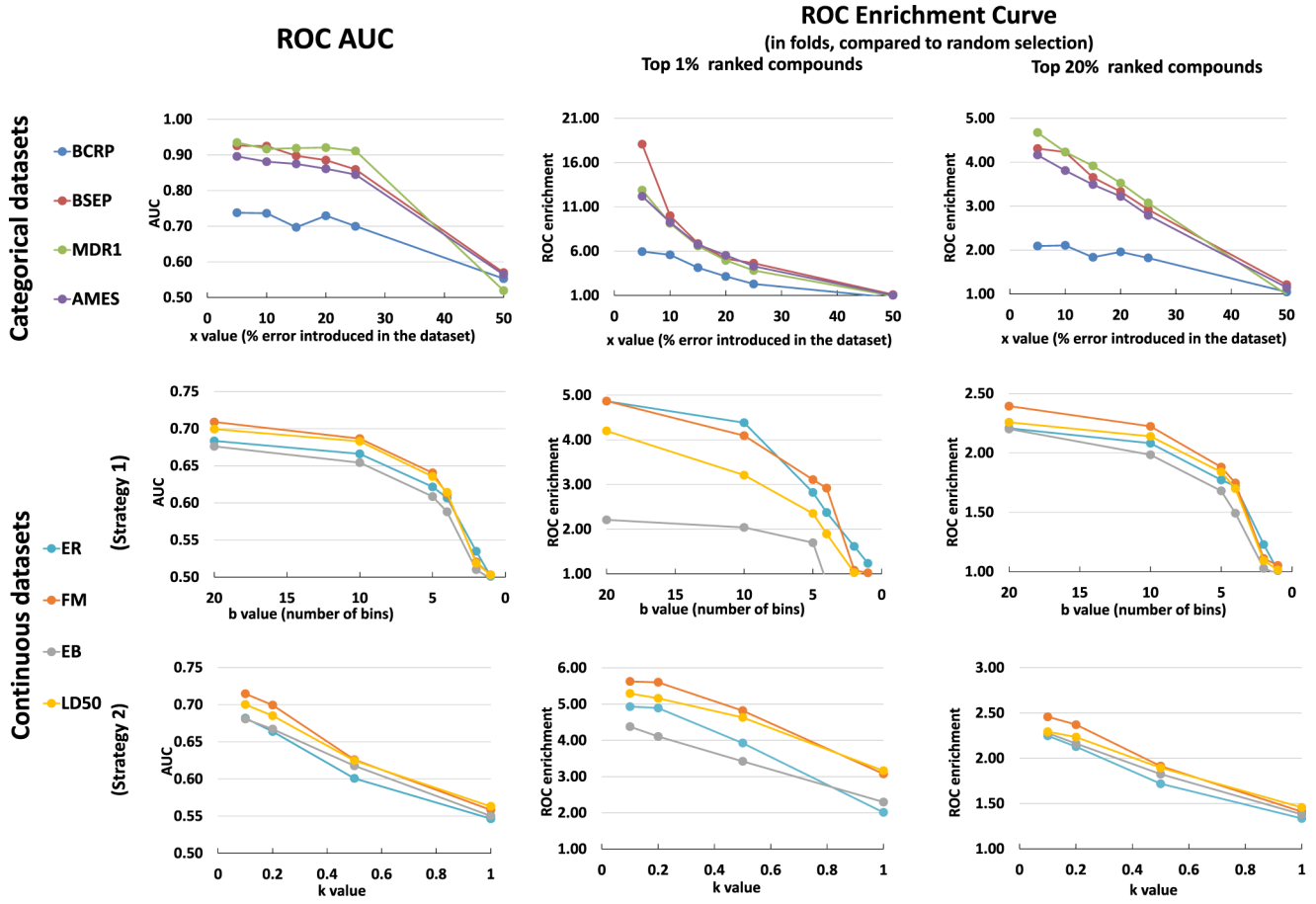


Figure 1. ROC AUC and ROC enrichment plots for each data sets.

are usually obtained by taking the average of multiple measurements (assuming that there is no systematic error in each measurement)[10] and/or testing the compounds under multiple concentrations.[10,11] Experimental errors normally occur when testing compounds just a single time and/or under a single concentration. Modeling data sets defined by a single measurement containing experimental errors will decrease the predictivity of the resulting QSAR models, according to a previous study.[12] Recently, Cortes-Ciriano et al.[13] simulated the experimental errors in QSAR modeling sets, and then compared the influence of different QSAR approaches on predictive accuracy. This study provides a practical reference for making a better decision about which modeling approach should be chosen depending on the quality of modeling sets. Roy et al.[14] have studied the relationship between systematic errors in the predictions and the applicability domain (AD) of QSAR modeling. They also exposed the flaw of using normal correlation coefficients to describe model predictivities.[15] These previous studies mainly focus on the relationship between the predictivity of QSAR models and the quality of modeling sets or the selection of modeling approaches. However, there is no systematic study on how to obtain a reliable QSAR model from an error-ridden modeling set (either a continuous set or a categorical set). Two relevant questions that have not been answered are (1) whether we can identify large experimental errors in the data sets, and (2) what can we do to improve models based on data sets with such errors.

The goal of this study is to address the above two challenges by designing a practical workflow, which can be used to identify potential experimental errors in QSAR modeling sets, providing a guide to improve the predictivities of the models from low-quality modeling sets. In this study, eight in-house data sets, which consisted of various numbers of compounds, were modified by introducing different levels of simulated experimental errors. Four types of QSAR models were generated with the original/modified modeling sets and the model performances were evaluated by a fivefold cross-validation process. Finally, all of the models were also evaluated by predicting one external set, which was set aside at the beginning of the modeling procedure.

RESULTS AND DISCUSSION

Overview. In this study, eight data sets with various bioactivities were used for modeling purposes. Some of them (e.g., AMES) have been extensively used in previous QSAR studies.[16−19] For this reason, the QSAR models developed in this study with the original modeling sets (without introducing simulated experimental errors or removing any compounds) have similar performances compared to those of previous studies. Furthermore, according to our previous studies, the consensus predictions (i.e., averaging predictions of all individual models) showed significant advantages compared to those of individual models, especially for external predictions.[20−23] Similarly, the consensus predictions obtained the highest accuracy for almost all models in this study (Tables S1 and S2). To avoid the complexity of comparing hundreds of different individual models, we only compared the consensus model performances for each data set in the following

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812

discussions. External prediction results of all consensus models are reported in Tables S3 and S4.

Three methods (one for categorical data sets and two for continuous data sets) were used to simulate experimental errors in the modeling sets (see Materials and Methods section for details). Several new modeling sets were generated with different levels of experimental noise added. For each categorical data set, there are six new modeling sets generated. For each continuous data set, there are six and four new modeling sets generated, using the two methods for introducing experimental errors. After the simulated experimental errors were introduced into the modeling sets, the model performance in the fivefold cross-validation for all data sets deteriorated (data not shown).

Can QSAR Modeling Identify Potential Experimental Errors in Modeling Sets? The major goal of this study is to identify experimental errors in a modeling set using QSAR approaches. To this end, we performed a fivefold cross-validation for each model and consensus predictions were made based on the results of the fivefold cross-validation of all individual models. The compounds in each data set can be then sorted in decreasing order by their apparent prediction errors. The topmost compounds with the largest prediction errors can then be checked for the amount of introduced experimental noise. Plots on the left in Figure 1 shows the area under the receiver operating characteristic curve (ROC AUC) plot for each data set when prioritizing compounds with simulated experimental errors by their cross-validation prediction errors between experimental data and consensus predictions (ROC plots can be found in Figures S1 and S2). After sorting the compounds by their prediction errors, it is noticeable from the ROC enrichment plots on the right in Figure 1 that the compounds with simulated experimental errors can be prioritized in most data sets. For example, in categorical data sets, the top 1% compounds from the MDR1-*x*5 modeling set obtained about 12.9 (in folds, compared with that of the random selection) of ROC enrichment and the top 20% compounds from MDR1-*x*5 modeling set obtained about 4.7 (in folds, compared with that of the random selection) of ROC enrichment. The other two categorical data sets, BSEP and AMES, have similar results compared to those of MDR1. However, the ROC enrichment in BCRP data sets is not as significant as that in the others. The BCRP set is the smallest data set, which only contains about 300 compounds in the modeling set. The prediction accuracy of BCRP models is also worse than that for the other three data sets. It is thus reasonable to conclude that the impact of experimental errors on the QSAR modeling is stronger for small data sets than that for large data sets.

In continuous data sets, due to the nature of the two methods used to simulate experimental errors, every compound contains a certain level of simulated error. The ROC AUC plots for continuous data sets are based on the ratio of prioritized simulated experimental errors in the whole data set (i.e., the sum of simulated experimental errors in the prioritized compounds divided by the total error amount). Not surprisingly, the prioritization of compounds with simulated experimental errors is not as efficient as for categorical data sets because every compound carries some simulated experimental errors. The largest ROC AUC for continuous data sets is about 0.70, which is lower than that of categorical data sets. But the ROC enrichment plot of all continuous data sets still shows the ability of the cross-validation of the modeling sets themselves to prioritize compounds with large errors. For example, in the case of strategy 1 (experimental error simulation strategy 1, details are in the method part below), the top 1% compounds from the LD50-b20 modeling set obtained about 4.2 (in folds, compared with that of the random selection) of ROC enrichment and the top 20% compounds from LD50-b20 modeling set obtained about 2.3 (in folds, compared with that of the random selection) of ROC enrichment. In the case of strategy 2 (experimental error simulation strategy 2, details are in the method part below), the top 1% compounds from LD50-b20 modeling set obtained about 5.3 (in folds, compared with that of the random selection) of ROC enrichment and the top 20% compounds from LD50-b20 modeling set obtained about 2.3 (in folds, compared with that of the random selection) of ROC enrichment compared with that of the random selection.

For both categorical and continuous data sets, when the level of simulated experimental errors increases (e.g., the ratio of compounds with simulated errors rises), the prioritization of compounds with simulated errors using QSAR modeling became less efficient (ROC enrichment plots in Figure 1, ROC enrichment heatmaps in Supporting Information). For example, in the categorical data sets (Figure 1), the top 1% compounds from MDR1-*x*25 modeling set obtained about 3.8 (in folds, compared with that of the random selection) of ROC enrichment, which is much lower than that from the MDR1-*x*5 modeling set (12.9). A similar situation was found in the continuous data sets, the top 1% compounds from the FM-b5 modeling set obtained about 3.11 (in folds, compared with that of the random selection) of ROC enrichment, which is lower than that from the FM-b20 modeling set (4.9). And the top 1% compounds from the FM-*k*0.5 modeling set obtained about 4.8 (in folds, compared with that of the random selection) of ROC enrichment, which is lower than that from the FM-*k*0.1 modeling set (5.6). When modeling sets contain a large amount of simulated experimental errors (e.g., MDR1-*x*50, EB-*n*1, and EB-*k*1.0, etc.), the prioritization of compounds with simulated errors using QSAR modeling is not better than random selection. Our results indicate that the cross-validation of modeling sets themselves is capable of prioritizing compounds with experimental errors when (1) the modeling set is large enough and well curated; and (2) the level of experimental noise present in the data set is not too high. These conditions are essential for obtaining good models, that is, those capable of capturing true data relationships.

Can We Improve QSAR Models Predictions? Previous studies showed that applying the AD can improve model predictivity by removing compounds with unique structures (i.e., structure outliers[17,23,24]). There is a recent report demonstrating the importance of checking model AD before comparing their predictivities.[14] In this study, we also applied AD, which is defined by calculating the Euclidean distance between an external compound to its nearest neighbor in the modeling set, to all of the model predictions. The external model predictivities have moderate improvements after applying AD (Tables S3 and S4). However, it is clear that the implementation of AD could not significantly improve the predictivity of the models based on modeling sets with simulated experimental errors. Similar to what has been shown in the above section, the external predictivities of these models are still much lower than the models based on the original modeling sets.

As shown in the above section, most compounds with simulated experimental errors in the modeling sets can be

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812

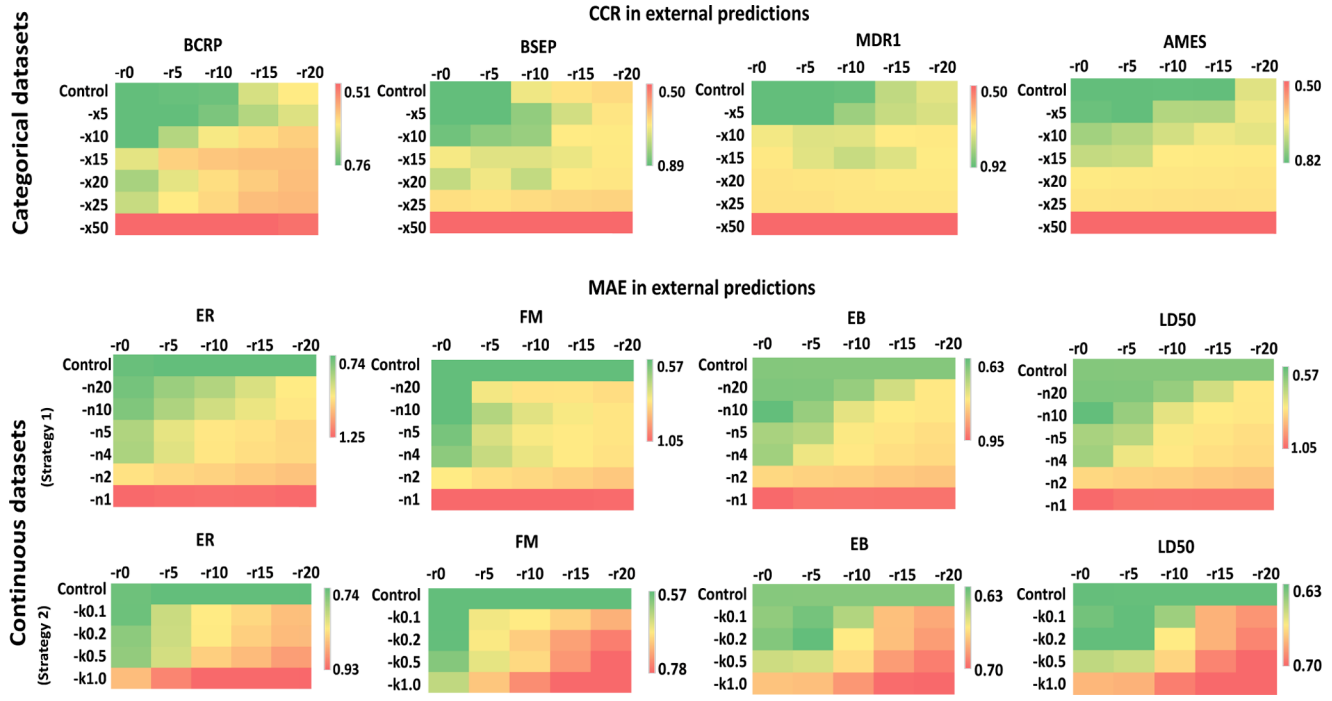


Figure 2. Comparison of external prediction results for each model from different modeling sets. In each heatmap, the *x* axis represents modeling sets with the top ranked 5, 10, 15, and 20% compounds removed by cross-validations, *y* axis represents modeling sets with different ratios of simulated experimental errors.

prioritized by the cross-validation procedure. It is noticeable that most of the compounds with simulated experimental errors can be excluded by removing the top 10−20% compounds, ranked by their prediction errors, from modeling sets. Therefore, different amounts of top-ranked compounds were removed from the sets, and the resulting new modeling sets with different, reduced sizes were used to redevelop QSAR models using the same approaches. For each data set, the top 5, 10, 15, and 20% compounds, which contain the highest cross-validation errors were removed to form four new modeling sets and the relevant QSAR models were developed accordingly. Not surprisingly, the cross-validation results with reduced modeling sets showed better statistics (e.g., higher correct classification rate (CCR)) than in those with all simulated experimental errors (data not shown).

In Figure 2, the external validation results of these new QSAR models, generated using reduced modeling sets, were shown and compared with those generated using all compounds. The predictivity of the same external compounds can truly reflect the model predictivity power for new (unseen to the model) compounds. The external prediction results of these new QSAR models are presented in Tables S3 and S4. The results of the above section showed that when the fraction of compounds with simulated experimental errors increased in the modeling set, the external predictivity deteriorated (the first column on the left of each heatmap). However, although most of these experimental errors can be removed by ranking the modeling set compounds by cross-validation results, the external predictivity of all of the models showed no improvement. For example, after removing 15% compounds from BSEP-x10 modeling sets, the ratio of compounds with experimental errors drops from about 10 to about 2%. But the CCR of external predictivity has no significant change (Figure 2). Because R^2 is not always suitable to describe model predictivity, especially for external compounds,[15] here we used mean absolute error (MAE) as a criteria to compare the predictivities of continuous models and a similar situation was obtained in the continuous data sets. For example, the external validation deteriorated after removing 10% compounds from the ER-*n*10 modeling set (the third row in the ER-*n* heatmap) and the ER-*k*0.2 modeling set (the third row in the ER-*k* heatmap). The MAE of external prediction increased from 0.75 to 0.80 for ER-*n*10 data sets and to 0.79 for ER-*k*0.2 data sets.

All of the results above indicate that, although most compounds with simulated experimental errors can be identified using the prioritization strategy based on the cross-validation results, simply removing the suspicious compounds from the modeling sets did not improve the external predictivity of QSAR models. When the top-ranked compounds are removed as described above, a certain number of compounds with the correct experimental values are removed as well. This step will not only decrease the AD of model, which normally depends on the size of the modeling set, but will also result in the overfitting issue, as reported previously.[25]

What Can We Do to Modeling Sets with Suspicious Data? Another interesting finding is that the external predictivity of QSAR models seems unaffected when the ratio of simulated experimental errors is small in the modeling set (Figure 2). For example, among categorical data sets, the external predictivity of BSEP-*x*5 and BSEP-*x*10 models (CCR = 0.88 and 0.87, respectively) is similar to that based on the original BSEP modeling set (CCR = 0.89). Among continuous data sets, the external predictivities of the FM-*n*20 and FM-*n*10 models (R^2 = 0.67 and 0.68, MAE = 0.57 and 0.57) is similar to that based on the original FM modeling set (R^2 = 0.66, MAE = 0.57). Similar situations can be found with other models/modeling sets. We believe that two factors contribute to this observation. First, the models can tolerate and overcome the small amount of noise/errors in the data set, if it is sufficiently large. Second, the inherent amount of noise present in the

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812

Table 1. Information on Chemical Data Sets Used in This Study

	size	actives	inactives	description	sources
categorical sets					
BCRP[b]	395	178	217	inhibition of membrane transporters at 10 μM	Sedykh et al.[16]
BSEP[b]	725	303	422	bile salt efflux pump inhibition at 100 μM	Metrabase database[34]
MDR1[b]	1585	750	835	inhibition of membrane transporters at 10 μM	Sedykh et al.[16]
AMES	3979	1718	2261	bacterial mutagenicity Ames test	CCRIS database[35]

	size	[Min; Max]	mean ± SD	description	sources
continuous sets[a]					
ER	546	[−4.50; 2.81]	−0.03 ± 1.57	relative binding affinity to ERα	Zhang et al.[18]
FM[b]	675	[−5.94; 2.00]	−2.12 ± 1.35	LC_{50}, toxicity to fathead minnow at 96 h exposure	Klopman et al.[28]
EB[b]	899	[−2.18; 6.34]	3.19 ± 1.23	IC_{50}, toxicity to environmental bacteria (U.S. EPA MICROTOX test)	Pangrekar et al.,[26] Klopman et al.[27,28]
LD50	7332	[−0.34;10.21]	2.54 ± 0.96	LD_{50}, rat acute toxicity, oral	Zhu et al.[17]

[a]Continuous activity values were negative log 10 transformed. [b]Denotes proprietary data sets provided by Multicase Inc. (http://www.multicase.com/case-ultra-models) as accessed in 2015. For these, the "Source" column provides direct precursor publications.

original experimental data (this amount depends on the endpoint) sets the upper limit on the evaluation accuracy of models, so that models based on controls and noisy data sets will not be easily distinguishable by performance, if their accuracy is close to or exceeds that limit.

CONCLUSIONS

In this study, we used four continuous and four categorical data sets, which have been extensively curated in-house and by our collaborators to address two questions related to experimental errors in the modeling set: (1) Can we find a way to identify the experimental errors in the modeling sets? (2) What can we do to improve the QSAR models, which are generated from data sets containing a certain ratio of experimental errors?

By applying three experimental error simulation strategies on each data set, more than 1800 various QSAR models were generated from all of the modeling sets with different ratios of simulated errors. We described in detail the strategy for identification of experimental errors in modeling sets. The compounds with relatively large prediction errors in the cross-validation process are likely to be those with simulated experimental errors. Thus, the cross-validation of modeling sets is able to prioritize compounds with experimental errors. This strategy will work efficiently when (1) the modeling set is large and highly curated for the structure information; and (2) the experimental error level is not too high (e.g., the ratio of compounds with errors is lower than 5−15% for a categorical data set).

After identifying the experimental errors in the modeling sets by analyzing the cross-validation results, we noticed that most of the simulated experimental errors can be excluded by removing a certain percentage of compounds with a high ranking of prediction error. Therefore, various amounts of top-ranked compounds were removed from the modeling sets, and the resultant new modeling sets with different, reduced sizes were used to redevelop QSAR models. We performed external validations for these new models to evaluate their predictivities for new compounds. However, simply removing the suspicious compounds from the modeling sets did not improve the external predictivity of QSAR models. When the top-ranked compounds are removed, a certain number of compounds with true experimental values are also removed. This will not only decrease the prediction reliability but also result in the overfitting issue. Therefore, the suspicious compounds prioritized by cross-validation may be candidates for retesting to obtain the correct experimental values. If this is not possible, these sample points should be kept as they are, to allow model training to overcome these or at least to signify areas of chemical space, where prediction errors will be likely.

MATERIALS AND METHODS

Data Sets. The eight data sets used in this study (Table 1) were taken from public literature and extensively curated in-house or obtained from Multicase Inc. (Beachwood, OH 44122). These data sets include four categorical and four continuous bioactivity endpoints. The sizes of both the two types of data sets vary from hundreds to thousands. These data sets represent diverse biological properties useful for drug design and/or regulatory risk assessment. The BCRP, MDR1, and BSEP data sets represent inhibition of the respective membrane transporters. The AMES data set is a large bacterial mutagenicity collection from public sources. The ER data set was collected from previous estrogen receptor binding studies and specifically refers to the chemical binding affinity of ERα.[18] The EB data set contains the results of Microtox testing of environmental bacteria (aerobic heterotrophs, nitrosomonas, methanogens, and photobacteria) by U.S. EPA.[26,27] The remaining two data sets, FM and LD50, are whole animal toxicity endpoints, and represent the acute toxicity testing results against the fathead minnow and rat, respectively.[17,28]

Experimental Error Simulation. Different levels of experimental errors were simulated and introduced into each modeling set in this study. We used three different strategies to simulate experimental errors based on the data type. For each categorical data set, we randomly selected x% (x = 5, 10, 15, 20, 25, 50) compounds from the two classes and exchanged their activity categories and then obtained six new modeling sets. Each new modeling set was labeled based on their levels of simulated experimental errors. For example, the AMES-x5 modeling set is the new AMES modeling set, when x% = 5% of modeling set compounds have simulated experimental errors. For continuous modeling sets, there are two strategies used in this study to simulate experimental errors: (1) progressive scrambling, in which compounds were sorted by their activities, and were assigned to n bins (n = 1, 2, 4, 5, 10, or 20), thus forming n subsets based on activities. We randomly shuffled activity values among compounds within each bin and obtained six new modeling sets; (2) the standard deviation of the activity

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812

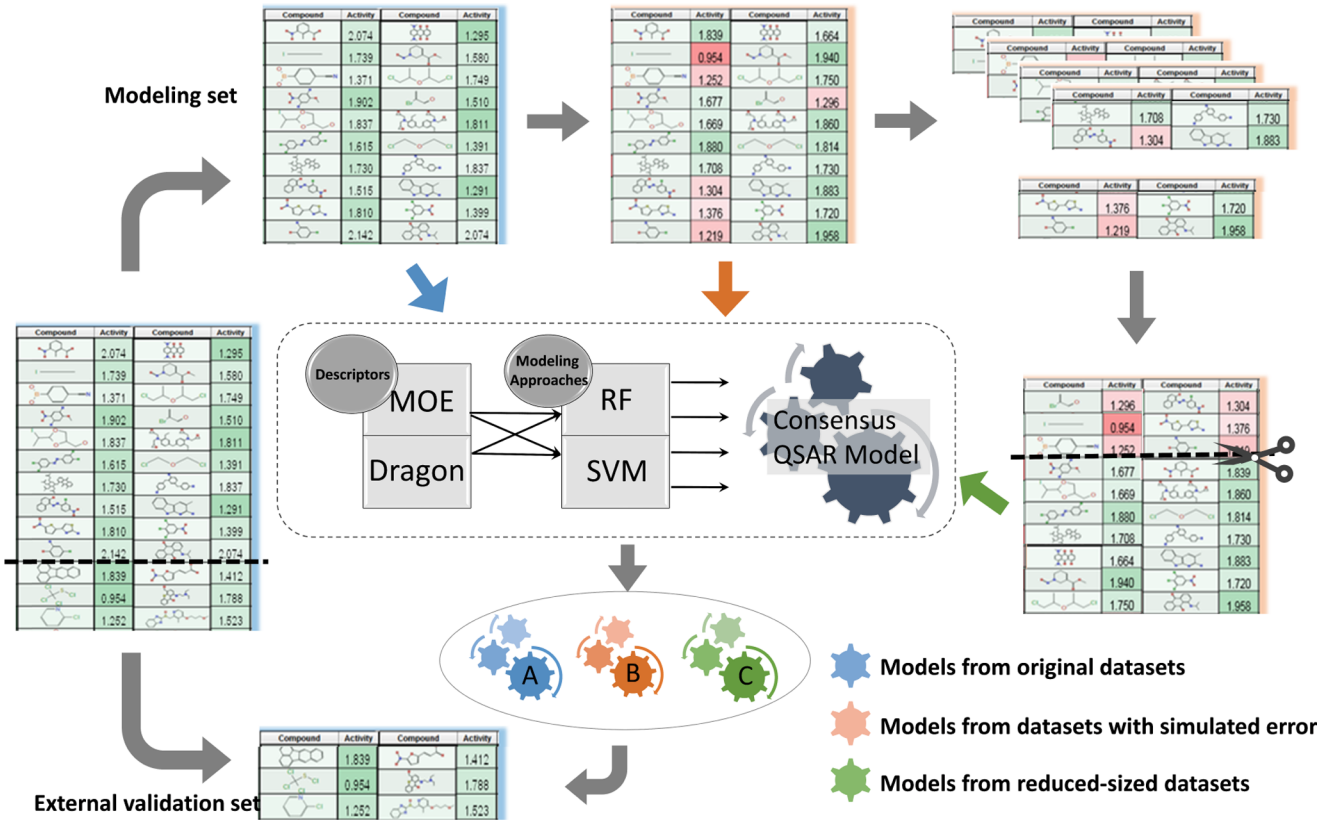


Figure 3. Modeling workflow.

was first derived in each data set. Then, the standard deviation of each data set was multiplied by a parameter k (k = 0.1, 0.2, 0.5, 1.0), and this result was denoted as sigma. We generated random values from zero-centered normal distributions with each sigma, added these values as errors to the activity value of each compound in the original modeling sets and finally got four new modeling sets. Again, we named the new modeling set as LD50-n1, when n is 1, and LD50-k0.1, when k is 0.1. The first approach will generate relatively larger experimental errors than the second approach. We used both methods to cover various types of existing continuous data sets (e.g., some data sets with relatively larger experimental errors). All of the experimental error simulation work was repeated five times. (The experimental error simulations for AMES and LD50 can be found in Supporting Information.) The results presented in this study were the averages of all of the five trials.

Molecular Descriptors. Molecular Operation Environment (MOE) software version 2015.10[29] and Dragon version 6.0[30] were used in this study for calculating 192 (MOE) and over 1500 (Dragon) 2D chemical descriptors for compounds in each data set. After that, for each data set, all of the descriptor values were normalized to the range from 0 to 1, and redundant descriptors were excluded by deleting descriptors with low variance (standard deviation <0.01), and/or randomly deleting one from any pairs of descriptors that have a high correlation (R^2 > 0.95). The remaining 120−140 MOE descriptors and 700−1300 Dragon descriptors (actual numbers are data set dependent) were used in the following modeling process.

Modeling Approaches. In this study, QSAR models were developed using two machine-learning algorithms random forest (RF) and support vector machines (SVMs). In the RF algorithm, which was developed by Breiman,[31] a random forest is a predictor that consists of many decision trees and makes a prediction that ensembles outputs from each individual tree. In this study, RF was implemented in R.2.15.1[32] using the package "randomForest". In the random forest modeling procedure, n samples were randomly drawn from the original data. These samples were used to construct n training sets and to build n trees. For each node of the tree, m descriptors were randomly chosen from the descriptors set. The best data split was calculated using these m descriptors for each training set. In this study, only the default parameter values (n = 500; m is the square root of the number of descriptors for category models and one-third of the number of descriptors for continuous models) were used for model development.

The SVM algorithm was first developed by Cortes and Vapnik.[33] In this study, SVM was implemented in R.2.15.1[32] using the package "e1071". Basically, the SVM algorithm attempted to find the optimal separating hyperplane between two classes by maximizing the margin. The support vectors are the points, which fall within this margin. The outlier data points (i.e., data points on the "wrong" side of the margin) are weighted down to reduce their influence. In the nonlinear case, the data points are usually projected into a higher-dimensional space (to make them linearly separable) using kernel techniques. There are many types of SVM extensions in the package "e1071" based on different types of kernels. In this study, we used the eps-regression SVM approach and its kernel type is radial basis.

Applying AD. In this study, the AD was calculated from the distribution of Euclidean distances between each compound and its nearest neighbor in the modeling set using the relevant chemical descriptors. The threshold value to define AD for a QSAR model places its boundary at one-half of the standard

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812

deviation calculated for the distribution of distances between each compound in the modeling set and its nearest neighbor in the same set. If the distance of the external compound from any of its nearest neighbors in the modeling set exceeds the threshold, the prediction is considered unreliable and excluded.

Modeling Workflow. The overall modeling workflow is as shown in Figure 3. Each data set was divided into a modeling set (83.3% of the overall set) and an external validation set (16.7% of the overall set). The modeling sets were then modified by introducing different levels of simulated experimental errors (see the next section for details) and the external validation sets were set aside and used to test the predictivity of each model. Multiple QSAR models were first created using the original modeling sets, and then a consensus model A (shown in blue in Figure 3) was generated by averaging the results of all individual QSAR models that were developed using a combination of a single modeling approach (either RF or SVM) and a single type of descriptor (either MOE or Dragon). Then, QSAR models were also developed using modeling sets with different ratios of simulated errors, and a consensus model B (shown in orange in Figure 3) was generated as well. The fivefold cross-validation was carried out to show the performance of the resulting models (Tables S1 and S2). In the fivefold cross-validation process, each modeling set was randomly divided into five equivalent subsets. Each time, four subsets (80% of the modeling set compounds) were combined and used to develop QSAR models and the remaining one subset (20% of the modeling set compounds) was used as a test set for validating purposes (Figure 3). This procedure was repeated five times so that each modeling set compound was used for prediction once.

We tested the performance of the models by applying AD (Tables S3 and S4). Then, we tested the performance of the models by removing the modeling set compounds with large prediction errors in the fivefold cross-validation process. By removing different ratios (i.e., ratio = 5, 10, 15, and 20%) of the modeling set compounds based on their prediction errors, the QSAR models were redeveloped by the reduced size modeling set. This effort resulted in the consensus model C (shown in green in Figure 3). Eventually, all QSAR models were compared to each other using the same excluded validation set.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b00274.

Error simulation examples and ROC enrichment heatmaps (XLSX)

AUC plots for categorical data sets and continuous data sets; fivefold cross-validation results for categorical data sets and continuous data sets; external prediction results for categorical data sets and continuous data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: alex.sedykh@sciome.com (A.S.).

*E-mail: hao.zhu99@rutgers.edu. Tel: (856) 225-6781 (H.Z.).

ORCID

Linlin Zhao: 0000-0002-7283-7681

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Multicase Inc. (Beachwood, OH) for providing the five data sets (BCRP, BSEP, MDR1, FM, and EB) used in this study. This research was supported in part by the National Institutes of Health (NIH) grants P30ES005022 and R15ES023148, and the Johns Hopkins Center for Alternatives to Animal Testing (CAAT) grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and CAAT.

■ REFERENCES

(1) Sprous, D. G.; Palmer, R. K.; Swanson, J. T.; Lawless, M. QSAR in the Pharmaceutical Research Setting: QSAR Models for Broad, Large Problems. *Curr. Top. Med. Chem.* **2010**, *10*, 619−637.

(2) Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants. *Chem. Res. Toxicol.* **2014**, *27*, 1643−1651.

(3) Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discovery Today* **2011**, *16*, 747−750.

(4) Prinz, F.; Schlange, T.; Asadullah, K. Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets? *Nat. Rev. Drug Discovery* **2011**, *10*, 712.

(5) Ioannidis, J. P. A.; Allison, D. B.; Ball, C. A.; Coulibaly, I.; Cui, X.; Culhane, A. C.; Falchi, M.; Furlanello, C.; Game, L.; Jurman, G.; Mangion, J.; Mehta, T.; Nitzberg, M.; Page, G. P.; Petretto, E.; Van Noort, V. Repeatability of Published Microarray Gene Expression Analyses. *Nat. Genet.* **2009**, *41*, 149−155.

(6) Bell, A. W.; Deutsch, E. W.; Au, C. E.; Kearney, R. E.; Beavis, R.; Sechi, S.; Nilsson, T.; Bergeron, J. J. A HUPO Test Sample Study Reveals Common Problems in Mass Spectrometry-Based Proteomics. *Nat. Methods* **2009**, *6*, 423−430.

(7) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.

(8) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* **2008**, *27*, 1337−1345.

(9) Fourches, D.; Muratov, E.; Tropsha, A. Curation of Chemogenomics Data. *Nat. Chem. Biol.* **2015**, *11*, 535.

(10) Hinkelmann, K. P. I.; Kempthorne, O. S. U. *Design and Analysis of Experiments, Introduction to Experimental Design*; 2nd ed.; John Wiley & Sons, Inc.: Hoboken, 2008; Vol. *1*.

(11) Feinberg, M.; Boulanger, B.; Dewé, W.; Hubert, P. New Advances in Method Validation and Measurement Uncertainty Aimed at Improving the Quality of Chemical Data. *Anal. Bioanal. Chem.* **2004**, *380*, 502−514.

(12) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 125−134.

(13) Cortes-Ciriano, I.; Bender, A.; Malliavin, T. E. Comparing the Influence of Simulated Experimental Errors on 12 Machine Learning Algorithms in Bioactivity Modeling Using 12 Diverse Data Sets. *J. Chem. Inf. Model.* **2015**, *55*, 1413−1425.

(14) Roy, K.; Ambure, P.; Aher, R. B. How Important Is to Detect Systematic Error in Predictions and Understand Statistical Applicability Domain of QSAR Models? *Chemom. Intell. Lab. Syst.* **2017**, *162*, 44.

(15) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be Aware of Error Measures. Further Studies on Validation of Predictive QSAR Models. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 18−33.

(16) Sedykh, A.; Fourches, D.; Duan, J.; Hucke, O.; Garneau, M.; Zhu, H.; Bonneau, P.; Tropsha, A. Human Intestinal Transporter Database: QSAR Modeling and Virtual Profiling of Drug Uptake, Efflux and Interactions. *Pharm. Res.* **2013**, *30*, 996−1007.

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812

(17) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative Structure-Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure. *Chem. Res. Toxicol.* **2009**, *22*, 1913−1921.
(18) Zhang, L.; Sedykh, A.; Tripathi, A.; Zhu, H.; Afantitis, A.; Mouchlis, V. D.; Melagraki, G.; Rusyn, I.; Tropsha, A. Identification of Putative Estrogen Receptor-Mediated Endocrine Disrupting Chemicals Using QSAR- and Structure-Based Virtual Screening Approaches. *Toxicol. Appl. Pharmacol.* **2013**, *272*, 67−76.
(19) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K. R.; Xi, L.; Liu, H.; Yao, X.; Öberg, 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.
(20) Wang, W.; Kim, M. T.; Sedykh, A.; Zhu, H. Developing Enhanced Blood-Brain Barrier Permeability Models: Integrating External Bio-Assay Data in QSAR Modeling. *Pharm. Res.* **2015**, *32*, 3055−3065.
(21) Kim, M. T.; Sedykh, A.; Chakravarti, S. K.; Saiakhov, R. D.; Zhu, H. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches. *Pharm. Res.* **2014**, *31*, 1002−1014.
(22) Solimeo, R.; Zhang, J.; Kim, M.; Sedykh, A.; Zhu, H. Predicting Chemical Ocular Toxicity Using a Combinatorial QSAR Approach. *Chem. Res. Toxicol.* **2012**, *25*, 2763−2769.
(23) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766−784.
(24) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, 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.
(25) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.
(26) Pangrekar, J.; Klopman, G.; Rosenkranz, H. S. Expert-System Comparison of Structural Determinants of Chemical Toxicity to Environmental Bacteria. *Environ. Toxicol. Chem.* **1994**, *13*, 979−1001.
(27) Klopman, G.; Stuart, S. E. Multiple Computer-Automated Structure Evaluation Study of Aquatic Toxicity. III. *Vibrio fischeri*. *Environ. Toxicol. Chem.* **2003**, *22*, 466−472.
(28) Klopman, G.; Saiakhov, R.; Rosenkranz, H. S. Multiple Computer-Automated Structure Evaluation Study of Aquatic Toxicity II. Fathead Minnow. *Environ. Toxicol. Chem.* **2000**, *19*, 441−447.
(29) *Molecular Operating Environment (MOE)*; Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2016.
(30) *Dragon (Software for Molecular Descriptor Calculation)*, version 6.0; Talete srl, 2013. http://www.talete.mi.it/.
(31) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(32) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013. http://www.R-Project.org/.
(33) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.
(34) Mak, L.; Marcus, D.; Howlett, A.; Yarova, G.; Duchateau, G.; Klaffke, W.; Bender, A.; Glen, R. C. Metrabase: A Cheminformatics and Bioinformatics Database for Small Molecule Transporter Data Analysis and (Q)SAR Modeling. *J. Cheminform.* **2015**, *7*, 31.
(35) *Chemical Carcinogenesis Research Information System (CCRIS) Database*; National Library of Medicine: Bethesda, MD. https://toxnet.nlm.nih.gov/newtoxnet/ccris.htm.

DOI: 10.1021/acsomega.7b00274
ACS Omega 2017, 2, 2805−2812