molecular pharmaceutics

ARTICLE

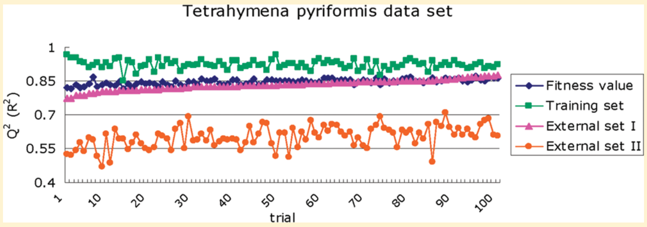
pubs.acs.org/molecularpharmaceutics

Why QSAR Fails: An Empirical Evaluation Using Conventional Computational Approach

Jianping Huang and Xiaohui Fan*

Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

S *Supporting Information*



ABSTRACT: Although a number of pitfalls of QSAR have been corrected in the past decade, the reliability of QSAR models is still insufficient. The reason why QSAR fails is still under hot debate; our study attempts to address this topic from a practical and empirical perspective, evaluating two relatively large toxicological data sets using a typical combination of support vector machine (SVM) and genetic algorithm (GA). Our results suggest that the vast number of equivalent models to be chosen and the insufficient validation strategy are primarily responsible for the failure of many QSAR models. First, a method often produces much more equivalent models than we might expect, and the corresponding descriptor sets show little overlap, indicating the unreliability of the conventional approaches. Moreover, although external validation has been considered necessary, validation on an arbitrarily selected independent set is still insufficient to guarantee the true predictability of a QSAR model. Therefore, more effective training and validation strategies are demanded to enhance the reliability of QSAR models. The present study also demonstrates that combinatorial or ensemble models can greatly reduce the variance of equivalent models, and that models built with the most frequently selected descriptors used by the equivalent models seem to yield more promising performances.

KEYWORDS: QSAR modeling, reliability, predictability, external validation, combinatorial model, ensemble method, support vector machine, genetic algorithm

1. INTRODUCTION

Quantitative structure−activity relationship (QSAR) models have been widely used in drug discovery and development, as well as in predictive toxicological studies and regulatory support.[1] Although the past decades have seen significant contributions of QSAR to these applications, e.g., reduction of cost, time and the use of animals,[2] the true predictability of QSAR models has not lived up to expectations,[3,4] and the reason why QSAR fails is still currently under hot debate.[3−6]

The reliability and robustness of QSAR models have captured much attention in the past decade,[7−10] aiming to improve the models' true predictability. For example, the modeling community has shown an increased awareness of the incorrect use of q^2 as a criterion for a model's predictability; and a consensus seems to have been reached that the high value of internal validation is not a sufficient criterion for the high predictive power of QSAR model, and that external validation is the only way to establish a reliable QSAR model.[11,12] With respect to the internal validation, especially when suffering from unavailability of external data, leave-one-out (LOO) and leave-many-out (LMO) cross validation and *Y*-randomization (i.e., permutation tests) are recommended as valuable validation steps. However, the validation and prediction should not be performed beyond the applicability domain of QSAR models.[9,12] Data quality and training sample size were also reported to have great impact on the reliability of QSAR models.[5,13]

Despite the effort devoted to investigating pitfalls of QSAR, the whole situation has not changed a lot: QSAR still often fails.[4] Recently, Maggiora[3] suggested that the chemical space is not as smooth as previously thought, and the existence of activity cliffs is responsible for the low predictability of most QSAR models. Different views were also present to explain the failure of QSAR models. Doweyko[6] attributed the failure to the improper selection of models likely caused by the overarching irrelevant or chance correlation. Johnson[4] also argued that "the manner in which QSAR is practiced is more responsible for its lack of success than any innate cause."

QSAR is known for establishing correlation between chemical structure and biological activity, and the success of a model greatly depends on how well it characterizes the structure. There are several known issues that restrict the QSAR model. Chemical structure is encoded in a numerical form in terms of molecular descriptors, in which information loss is unavoidably introduced. Additionally, the increasingly large number of molecular descriptors generated by the modern technique further confounds our selection of the most informative (or biologically relevant) ones. Theoretically, an ideal predictive QSAR model should capture the most informative molecular descriptors and represent a hypothesis regarding the underlying physical or biological phenomenon, because a model only possesses true predictability if it

Received: December 10, 2010
Accepted: March 3, 2011
Revised: February 24, 2011
Published: March 03, 2011

is biologically relevant. However, the wrong molecular descriptor set is often chosen, resulting in an incorrect model.

Naturally, the question is raised of whether it is possible for us to reach a near optimal descriptor set and model by using the conventional computational approach. It is no doubt that, given hundreds or even thousands of descriptors available, some descriptors are more biologically relevant and informative than the others. Hence, it is understandable that to obtain a model with a certain predictive power and minimize descriptors has been a common goal in the field of QSAR, the underlying hypothesis of which is that the modeling technique used is able to capture this subset or at least a near-optimal subset. But is this true? Are the conventional approaches really able to capture the optimal descriptor set and model?

The primary objective of our study is to try to answer the above questions. First of all, we attempted to determine if it was possible to retrieve the most informative descriptors using only the conventional approach. Assuming that the best fitting model is able to capture the most informative descriptors, or at least some of them, removing these descriptors from the descriptor pool and rebuilding the model using the remaining descriptors should result in a decreased fitting and prediction accuracy. Thus, it is rational to anticipate a downward trend in fitting and prediction accuracy if these steps are repeated many times. With this in mind, two relatively large toxicological data sets with more than 400 molecular descriptors were investigated. A typical combination of support vector machine and genetic algorithm (SVMGA) were used to obtain a best fitting model, and the corresponding descriptor set was then removed from the descriptor pool; the same process was repeated based on the remaining descriptors. Interestingly, the results of this study demonstrated that the number of equivalent descriptor sets was higher than we expected, and no obvious downtrend could be observed for even one of the data sets. This suggests that the most informative subset of descriptors may not exist, or that the SVMGA method employed here is unable to identify the most important subset of descriptors.

This study illustrated that a high accuracy of a randomly selected external set does not necessarily imply a high predictive power of a QSAR model. Much research has emphasized the importance of external validation for QSAR modeling. As a result, applying an independent set, separated from training samples or collected from other sources, for external validation has become an indispensible step in many studies. However, can this external validation process guarantee the true predictability of a QSAR model? Unfortunately, our study illustrated that there was no obvious relationship between the accuracies of two randomly selected external sets. In other words, a QSAR model with a high accuracy on an external set does not necessarily also have a high accuracy on another external set. Thus, it is not proper to use the accuracy of an arbitrarily selected external set as a criterion for the true predictability of a QSAR model.

Despite the uncertainty mentioned above, there are still ways to reach a more reliable descriptor set and model. Two possible approaches were proposed in this study, and their results were recorded. We illustrated that combining equivalent models can greatly reduce the variance, and that building models with the most frequently selected descriptors used by the equivalent models yields more promising performances.

2. MATERIALS AND METHODS

2.1. Data Sets. The two relatively large toxicological data sets used in this study were taken from the literature, i.e., the hERG data set[14] for the classification problem, and the *Tetrahymena pyriformis* (*T. pyriformis*) data set[15] for the regression problem, respectively. The summary of these data sets is given in Table 1.

Table 1. Summary of Data Sets

data set	no. of training sets	no. of external sets I	no. of external sets II	no. of descriptors
hERG	561[a]	1795[b]		437
T. pyriformis	644	339	110	407

[a] Active, 213; inactive, 348. [b] Active, 220; inactive, 1575.

hERG Data Set. The human Ether-a-go-go Related Gene (hERG) potassium channel, which is expressed in cardiac muscle cells, is one of the major causes responsible for QT interval prolongation and cardiac arrhythmia. The unwanted blockade of the hERG channel has led to the withdrawal of many drugs from the market. Thus, weeding out potential hERG channel inhibitors in the early stages of drug discovery circle is of interest.

The hERG data set used here was collected from Li et al.'s study.[14] In their study, a SVM classifier was trained using 495 samples combined with pharmacophore-based GRIND descriptors. The classifier was then applied to two external sets containing 66 and 1948 compounds, in which 72% and 73% of compounds were correctly predicted, respectively.

In our study, 495 training samples and 66 external samples taken from the literature were combined as the training set, among which 213 were active and 348 were inactive. The other 1948 compounds were assigned to the external set. However, due to limitations of the DRAGON software,[16] the generation tool for molecular descriptors used in this study, 153 compounds of the external set could not be correctly identified; therefore, the final external set used here contained 1795 compounds, among which 220 were active and 1575 were inactive.

T. pyriformis Data Set. The *T. pyriformis* data set was used by Zhu et al.[15] to address the robust and predictive models of chemical toxicity through collaborative and consensual mean. It consists of a training set of 644 compounds, an external set I of 339 compounds, and an external set II of 110 compounds. The training set and external set I were divided randomly from a compiled collection. The external set II, on the other hand, was taken from another study by Schultz et al.[17] The prediction accuracy (linear regression coefficient R_{abs}^2) reported for the external validation set I and II ranged from 0.71 to 0.85.[15]

2.2. Molecular Descriptors. Molecular descriptors were calculated using the 5.4 version of the DRAGON software,[16] which can generate as many as 1664 molecular descriptors. To simplify the process, we considered only the 2D molecular descriptors in this study, which included a total 929 of descriptors from 13 categories. After removing those with invalid values, those with too many zeros (greater than 90%), and those with small standard deviations (less than 0.5), the final number of descriptors used in the hERG data set and *T. pyriformis* data set were 437 and 407, respectively. All data were scaled to (−1, 1) to avoid predominant features.

2.3. Methods and Implements. Support vector machine (SVM), presented by Vapnik et al.,[18] has been extensively applied to QSAR modeling as a benchmark machine learning method due to its high performance and resistance to overfitting. The Java package of libsvm (version 2.8),[19] a free support vector machine tool, was used in this study.

Genetic algorithm (GA)[20] is one of the most popular feature selection methods occurring in QSAR literature. The Java genetic

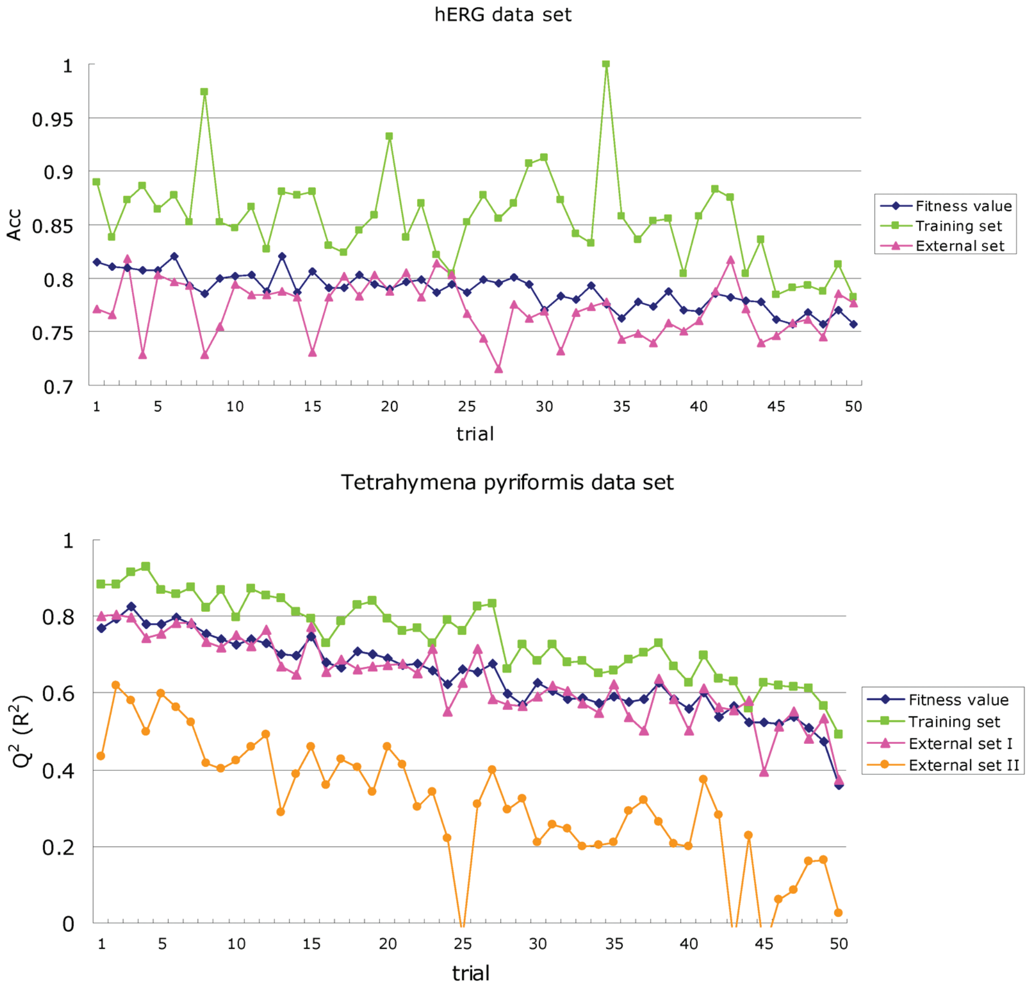


Figure 1. Results based on recursively removing the best fitting descriptors of hERG and *T. pyriformis* data set (50 trials). The x axis denotes the trials, and the y axis denotes the accuracy. The fitness values of GA denote the 5-fold cross-validation accuracy of the best fitting model obtained from the nth trial. In each trial, the fitness descriptor set and SVM parameters were used to train a SVM model based on the whole training set. The results of the training set and external set denote the fitting and external prediction accuracies predicted by using this SVM model, respectively.

algorithm package, Jgap (version 3.4.4),[21] was used for descriptor selection and parameter optimization.

Other methods used in this study for comparison included the following:

(1) k-nearest neighbors (kNN): The kNN algorithm classifies an unseen sample by a vote of k-nearest training instances as determined by some distance metric, typically Euclidean distance. In this study, the kNN method was implemented in the Java programming language, and its parameter k was determined using a nested 5-fold cross-validation from 1 to 15 and a step of 2.

(2) J48: A variant of C4.5 decision tree implemented in Weka (version 3.5.8),[22] a famous software package implemented in Java and available at http://www.cs.waikato.ac.nz/~ml/weka/ under the GNU General Public License; the default settings were used.

(3) CART: A classification and regression tree version implemented in Weka; the default settings were used.

(4) Random Forests (RF): Default settings recommended by Breiman[23] were used, and 1000 trees were built in a classifier.

(5) A SVM without feature selection (compared to the SVMGA): RBF kernel was adopted, and the cost C and parameter γ were tuned by a grid search on $[10^{-3}, 10^{-2}, ..., 10^{2}]$ and $[10^{-5}, 10^{-4}, ..., 10^{2}]$, evaluated using a nested 5-fold cross-validation on each training data set. All Java programs were running under a Java SE Runtime Environment (build 1.6.0_11-b03).

2.4. Validation. One of the most important tasks of the modeling process was defining of GA fitness function. For the classification problem, although measurements such as receiver operating characteristic (ROC) or area under the curve (AUC) were highly recommended,[24] we used the accuracy as our performance metric for simplicity since it was sufficient to our analysis. For the regression problem, the same performance metrics used in Zhu et al.'s study[15] were adopted:

$$Q_{\mathrm{abs}}{}^{2} = 1 - \sum(Y_{\mathrm{exp}} - Y_{\mathrm{cv}})^{2} / \sum(Y_{\mathrm{exp}} - \langle Y\rangle_{\mathrm{exp}})^{2} \tag{1}$$

$$R_{\mathrm{abs}}{}^{2} = 1 - \sum(Y_{\mathrm{exp}} - Y_{\mathrm{pred}})^{2} / \sum(Y_{\mathrm{exp}} - \langle Y\rangle \mathrm{exp})^{2} \tag{2}$$

$$\mathrm{MAE} = \sum|Y - Y_{\mathrm{pred}}|/n \tag{3}$$

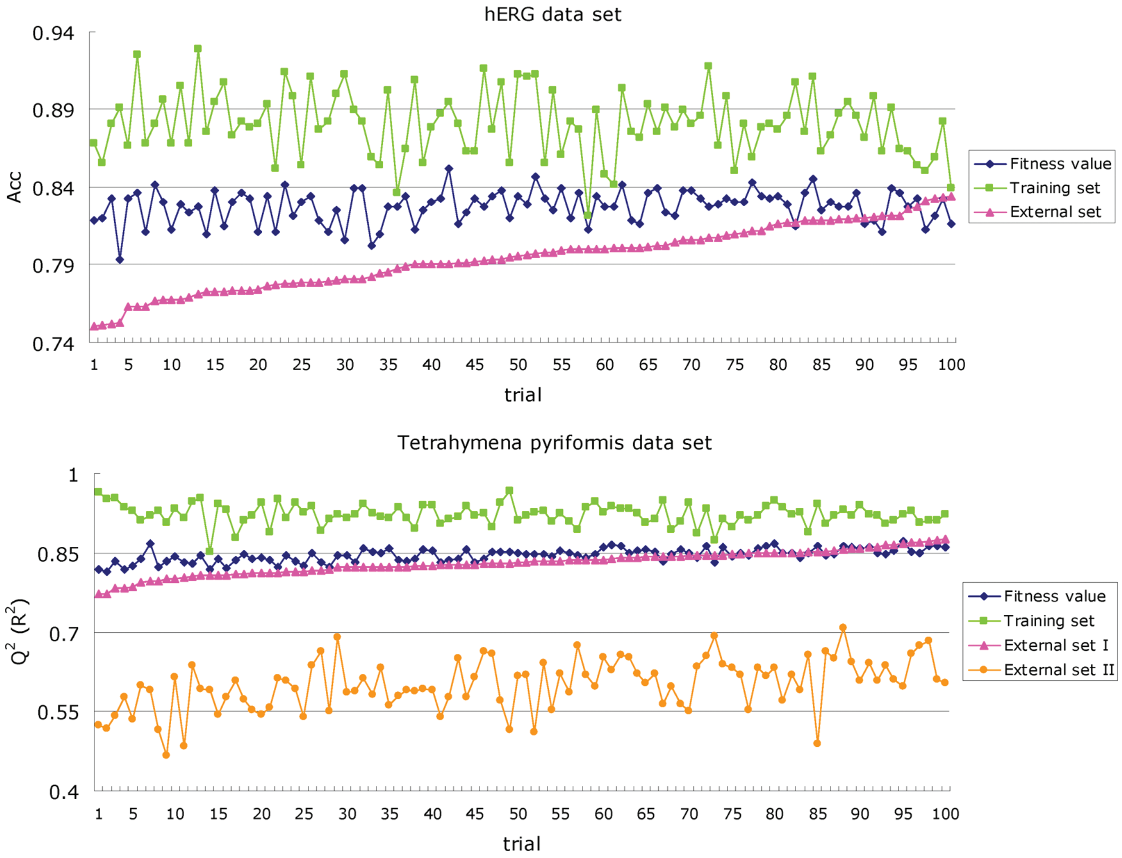


Figure 2. Results based on 100 trials of evaluation using a typical combination of SVM and GA on the whole hERG and *T. pyriformis* data sets, ordered by the prediction accuracy of external set I. The *x* axis denotes the trials, and the *y* axis denotes the accuracy. The fitness values of GA denote the 5-fold cross-validation accuracy of the best fitting model obtained from the *n*th trial. In each trial, the fitness descriptor set and SVM parameters were used to train a SVM model based on the whole training set. The results of the training set and external set denote the fitting and external prediction accuracies predicted by using this SVM model, respectively.

Here, $Q_{abs}{}^2$ represents the squared 5-fold cross-validation correlation coefficient for the fitting function, $R_{abs}{}^2$ is the coefficient of determination for the external validations sets, and MAE is the mean absolute error for the linear correlation between the predicted value Y_{pred} and the experimental value Y_{exp}.

3. RESULTS

3.1. Investigation on Recursive Removal of the Best Fitting Descriptors. The typical combination method of SVM and GA has been a benchmark approach for QSAR modeling. To assess how well SVMGA captured the most informative descriptors, we recursively removed the descriptor set occurring in the best fitting model from the descriptor pool, and rebuilt the model using the remaining descriptors. We repeated this process 50 times (trials) and evaluated the corresponding fitting and prediction accuracy in each trial.

Here, the GA was set to 50 population and 30 evolutions. By adjusting the gene and mutation probability, we restricted the GA to select only 3−10 descriptors. The fitness values were set to zeros for those chromosomes with fewer than 3 descriptors or more than 10 descriptors being selected. We called it a trial after 30 evolutions were finished, and the corresponding best fitness value, descriptor set, and SVM parameters were recorded. Subsequently, a final SVM model was trained on the whole training set, using these fitness descriptor sets and the SVM parameters, and applied to the same training set and external data sets. The fitting and external prediction accuracies were also recorded. This process was repeated for 50 trials for each of the two data sets (Figure 1).

Interestingly, despite that more and more best fitting descriptor sets were removed from the descriptor pool, little influence was observed on the fitness and external prediction accuracies for the hERG data set. Even in the 50 trial where only 48 descriptors were left, the total reduction of its fitness value was no more than 6%, and there was no obvious reduction observed for external prediction accuracy, either. A slight downward trend of fitness value and external prediction accuracy can be observed for the *T. pyriformis* data set, in which there were only 6 descriptors left after 50 trials (Figure 1). However, in the trials from 1 to 12, in which 101 descriptors were removed from the descriptor pool, the reduction of the fitness value was only around 3%, and the total reduction of external prediction accuracy was also not significant. This phenomenon indicates that the most informative subset of descriptors may not exist, or that the SVMGA employed here tends to miss some of the most important subset of descriptors in each trial. Here, we prefer the latter explanation, with reasons illustrated in the coming section.

3.2. Investigating the Relationship of Multiple Equivalent Outcomes. We investigated the relationship of the multiple equivalent outcomes yielded by the SVMGA method based on

Table 2. The List of Most Frequently Selected Descriptors of 100 Trials

	hERG		*T. pyriformis*	
no.	descriptors	frequency	descriptors	frequency
1	H-047	37	nDB	47
2	nF	31	O-057	44
3	MAXDN	20	ALOGP	43
4	nX	17	CIC0	42
5	T(F..F)	17	MAXDP	42
6	H-053	16	AMW	38
7	nO	16	piPC03	33
8	C-006	15	nROH	27
9	TPSA(Tot)	14	nArOH	26
10	TPSA(NO)	13	MLOGP	22
11	EEig05x	12	H-046	20
12	nR=Cs	11	nO	20
13	ZM1 V	11	SEigp	19
14	C-001	10	BLTA96	18
15	Ss	10	SEigZ	18
16	T(N..F)	10	ALOGP2	17
17	EEig14r	9	O-058	15
18	Pol	9	Ms	14
19	CIC0	8	TPSA(Tot)	14
20	DELS	8	BLTD48	13

the whole descriptor set, especially focusing on the overlap of multiple equivalent descriptors and the variance produced by corresponding models. Here, the same settings used in the previous section were employed, except that the restriction on the number of descriptors selected by GA was widened to 3−20, which enabled us to involve more possible informative descriptors at each trial. We repeated the same process for 100 trials without removing the descriptor set from the descriptor pool, which resulted in 100 equivalent descriptors and models. The best fitness value of GA and the fitting and external prediction accuracies based on the SVM model trained on the whole training set using the best fitting parameters were given (Figure 2). The trials were reordered and shown in ascending order of the prediction accuracy of external set I. We also examined the overlap of the descriptors selected in the 100 trials to determine the ability of SVMGA to capture the most informative descriptors. The top 20 frequently selected descriptors are shown in Table 2, and their meaning can be referred to the DRAGON software.

It was very clear that there was no obvious relationship between external prediction accuracy, fitting value of GA, or fitting accuracy of training set in either of the two data sets. As the external prediction accuracy of hERG data set rose from about 75% to 84%, the corresponding fitness value and fitting accuracy of the training set were bounded in a certain range but without any significant trend. The *T. pyriformis* data set showed similar results; the R^2 of external set I ranged from about 0.77 to 0.88. It was surprising to see that as the prediction accuracy of external set I rose, there was no similar trend observed for the prediction accuracy of external set II. This suggests that the external prediction accuracy may not be a proper criterion for the true predictability of a QSAR model. Even a high Q^2 on training set and a high R^2 on an external set still seem insufficient to guarantee the high predictive power of a QSAR model on other data sets. For example, considering trial 100 with relatively high Q^2 (0.86) of GA and R^2 (0.88) of external set I, the model yielded a R^2 as low as 0.60 for the external set II. Contrary to this, considering trial 29 with relatively small Q^2 (0.84) of GA and R^2 (0.82) of external set I, the R^2 for external set II was as high as 0.69.

Another interesting observation is that there was a notable gap between the accuracies of external set I and external set II. The accuracies of external set I were much closer to the fitness values than those of external set II. A possible explanation is that the former was more similar to the training set than the latter. If this were the case, the accuracies of the two external sets should be alike after consideration of only those external samples within the applicability domain defined by the training set. However, we repeated the same process by considering applicability domain measured by Euclidean distance to the descriptor centroid of training set and found no significant difference (Figure S1 in the Supporting Information). As can be seen from Figure S1 in the Supporting Information, even only those samples in the applicability domain were taken into account, the gap between the two external sets still exists, thus, consideration of applicability domain alone is not sufficient for assessing a model's predictability on an external set.

To get a deeper insight into why the gap exists between the two external sets, we combined the training set with the two external sets, regenerating the three data sets based on a *k*-mean cluster algorithm, thus making the data sets more similar to each other (Figure S2 in the Supporting Information). Results of Figure S2 in the Supporting Information showed that the accuracies of external set I and external set II were closer than those of Figure 2; however, a high R^2 on the external set I still did not necessarily indicate a high R^2 on external set II. Here, the similarity to the training set was measured by the similarity of a molecular descriptor to the descriptor centroid of training set;[25] the applicability domain was defined by the maximal Euclidean distance to the descriptor centroid of training set estimated using those 95% training samples nearest to the centroid. The difference between Figure S1 and Figure S2 in the Supporting Information seems to suggest that the distribution of a data set is also an essential important factor that may account for its accuracy.

The overlap of descriptors of Table 2 showed that none of descriptors were found in more than half of all equivalent models, although 3 to 20 descriptors were allowed to be selected for each SVMGA model. For the hERG data set, all except 3 descriptors occurred in no more than 20 models, indicating there was very little overlap of descriptors occurring in these equivalent models. Once again, this further suggested that the SVMGA is apt to miss part of, or even all of, the most important subset of descriptors (if there is any). For the *T. pyriformis* data set, the overlap of descriptors was slightly greater, which may suggest that it is easier for SVMGA to capture some of the important descriptors for the *T. pyriformis* data set than for the hERG data set. This may be the reason why the downward tendency of the *T. pyriformis* data set was more notable than that of hERG data set as seen from the Figure 1.

As intercorrelation of descriptors may have a negative impact for descriptor selection and model development, the descriptor pairs with highest Pearson correlation coefficient (≥90%) were also given for the *T. pyriformis* data set (Table S1 in the Supporting Information) and for the hERG data set (Table S2 in the Supporting Information). Considering the top 10 most

Table 3. Prediction Accuracy of the hERG External Set Based on All Descriptors and the Most Frequently Selected Descriptors by Using Different Classification Methods

method	10 descriptors	20 descriptors	50 descriptors	all descriptors
KNN	0.812	0.816	0.805	0.769
J48	0.802	0.817	0.802	0.788
CART	0.828	0.795	0.812	0.795
RF	0.827	0.831	0.828	0.817
SVM	0.831	0.821	0.805	0.789
av	0.820	0.816	0.810	0.792

frequently selected descriptors, 4 descriptors (i.e., ALOGP, CIC0, piPC03 and MLOGP) for *T. pyriformis* data set and 3 descriptors (i.e., nF, TPSA(Tot) and TPSA(NO)) for *T. pyriformis* data set were found to have high correlation coefficients with some of the other descriptors. These results showed that the high correlations among descriptors are very common and may play an important role for the reliability of QSAR models. However, they do not seem to be the most important factor leading to the unreliability of QSAR models. For example, with respect to the hERG data set, since the descriptor nF had high correlation with T(N..F) and F-083 (Table S2 in the Supporting Information), we combined the frequencies of all these three descriptors (Table 2) together and got a total overlap of 47. Descriptors TPSA(Tot) and TPSA(NO) also showed high correlation; after combining their frequencies (Table 2) together we got a total overlap of 27. Both the total overlaps were less than half of the number of trials, therefore, they should not be the primary reason accounting for the little overlap of descriptors occurring in the 100 equivalent models.

3.3. Reaching More Reliable Descriptor Set and Model Based on Most Selected Descriptors. As can be seen from Figure 2, SVMGA produced very diverse descriptor sets and models in different trials. Moreover, zero overlap of descriptors can be found in two equivalent models generated by the same SVMGA approach with the same settings. The great variance of the results yielded by these equivalent models remarkably affected their reliability. Certainly, randomization is unavoidable in a modeling process subject to many unstable factors. For example, as a stochastic heuristic searching algorithm, GA tends to get trapped into local optimization and yield different solutions. However, how about compiling all these solutions together and considering only those most selected descriptors? It is reasonable to assume that the most important descriptors would have a greater chance to be selected. Is this a possible way to reduce the variance produced by chance? With this in mind, a more reliable descriptor selection and modeling method was proposed here based on the most frequently selected descriptors.

Five different classification methods were also employed to the hERG data set to build different models based on the top 10, 20, and 50 descriptors (Only the top 20 were shown in Table 2). The models were trained on the training set and applied to the external set. Their accuracies were compared with those obtained by using all the descriptors; the results are recorded in Table 3. Clearly, there were improvements when the classification methods were based on only the top 10, 20, or 50 descriptors; an average accuracy of 82% was achieved when using only the top 10 descriptors. SVM obtained the highest accuracy (83.1%), about 4% higher than the average accuracy based on all descriptors. It should be noted that this accuracy was very close to the maximum prediction accuracy on the external set (83.4%), as can be seen from Figure 2. Similarly, for the *T. pyriformis* data set, improvements can also be observed when using the top 10, 20, and 50 descriptors, especially in terms of MAE. The overall best results were those based on the top 20 descriptors, with $R^2 = 0.87$ and MAE = 0.25 for external set I, and $R^2 = 0.67$ and MAE = 0.36 for external set II. These values were similar to the maximum external prediction accuracies from Figure 2, which were $R^2 = 0.86$ and MAE = 0.33 for external set I, and $R^2 = 0.7$ and MAE = 0.42 for external set II. In short, models built with the most frequently selected descriptors tend to yield higher or at least comparable performance, and thus seem to be more reliable.

3.4. Reaching a More Reliable Model Based on Combination. Another approach presented here to reach a more reliable model was through the combination of equivalent models. We demonstrated that the combinatorial or ensemble model can reduce the variance produced by their component models, and thus enhancing the reliability. Combinations of 5, 10, 20, and 50 models were investigated based on the 100 equivalent models obtained in section 3.2; the external prediction accuracies and corresponding standard deviations (stdev) were also provided for the hERG data set (Table 5) and *T. pyriformis* data set (Table 6). The accuracy and stdev were obtained based on 1000 trials' evaluations. For example, for the combinations of 5 models, we randomly retrieved 5 of the 100 equivalent models without replacement and evaluated the prediction results separately; the final prediction accuracy was then evaluated by voting. This evaluation process was repeated 1000 times, and the results were averaged.

It is safe to say that the performance of combinatorial model can be greatly enhanced. As the number of models combined together increased, the accuracy steadily rose from 79.5% to 82.2% and reached a plateau at around 20 models, while standard deviation dropped sharply from 2% to 0.3% (Table 5). A similar trend can also be observed for the *T. pyriformis* data set in terms of R^2 and MAE (Table 6). In short, the combinatorial models produced performance better than those obtained by the best single models with significantly less variance.

It is worth noting that, for the hERG data set, our methods outperformed the SVM method used in the original literature[14] by about 10%. For the *T. pyriformis* data set, more promising results yielded by our methods can also be observed, comparing with all 15 models developed in the original literature.[15] It is interesting to see that the consensus model developed in the literature[15] also showed improved accuracy over other single models. This is consistent with the idea that the combinatorial or ensemble model is usually more reliable than its component models.

4. DISCUSSION

Numerous issues have been found to be associated with the reliability of QSAR models, which can be categorized into data issues (e.g., data quality and sample size), modeling issues (e.g., overtraining, chance correlation), and prediction issues (e.g., domain applicability).[26] In this study, we focused on the training and validation process and emphasized why the conventional approach often arrives at the wrong model.

Two main objectives of QSAR models are (1) to predict biological activity of untested and possibly unavailable compounds and (2) to seek which physicochemical or structural properties of compound are correlated with the biological activity and to discover their relationships.[27] In either case, a good QSAR model should represent a hypothesis regarding the

Table 4. Prediction Accuracy of the *T. pyriformis* External Set Based on All Descriptors and the Most Frequently Selected Descriptors by Using SVM

data set	validation criterion	10 descriptors	20 descriptors	50 descriptors	all descriptors
external set I	Q^2	0.841	0.868	0.885	0.843
	MAE	0.283	0.254	0.241	0.299
external set II	Q^2	0.671	0.672	0.653	0.670
	MAE	0.366	0.363	0.379	0.384

Table 5. Prediction Accuracy and Standard Deviation of the hERG External Set by Combining Different SVMGA Models

no. of models	accuracy	stdev
1	0.795	0.020
5	0.809	0.009
10	0.816	0.007
20	0.821	0.005
50	0.822	0.003

underlying physical or biological phenomenon. Even if a QSAR model is designed for predicting biological activity only (e.g., for screening), we still need to ensure that the molecular descriptors used in the model have some biological relevance and are not just selected by chance: only if this condition is met can we guarantee that the model possesses true predictability. As a result, the model's mechanistic interpretation has been seriously addressed in the OECD principles that at least some possible associations between the descriptors used in the model and the predicted end point should be considered.[28] Thus, selection of the best or near-best relevant or informative molecular descriptor set during the modeling or feature selection process is essential to reach a powerful model.

Although it is a common goal for many researchers to retrieve as few explanatory descriptors as possible to facilitate interpretation of the QSAR models,[29] whether the available feature selection methods are capable of capturing those most informative descriptors still remains a question upon which little emphasis has been given. It is rational for us to believe that, among the great number of available descriptors, some are more relevant to the mechanism than the others. Therefore, it is more likely that we often arrive at the irrelevant descriptors by using the available feature selection methods. Involving many irrelevant descriptors into the model can greatly degrade predictive power.

The investigation results based on the recursive removal of best fitting descriptors (Figure 1) suggest that the typical SVMGA approach often fails to capture many of the informative descriptors, letting them slip into subsequent modeling trials. Thus, those informative descriptors are apt to be equally distributed over whole trials, which explains why the fitting and prediction accuracy shown in Figure 1 changes smoothly. The overlap study (Table 2) further supported this idea; the low overlap of different equivalent models suggests that the SVMGA approach may capture only a very small part, or even none, of the most informative subset of descriptors. The up and down of the external prediction accuracy as seen in Figure 2 may also indicate that some of the equivalent models are unable to capture the informative descriptors very well and thus lead to the degradation in accuracy. This suggests that the conventional approach often tends to include irrelevant descriptors and does not always capture the optimal informative descriptor set or reach the best model.

Table 6. Prediction Accuracy of the *T. pyriformis* External Set by Combining Different SVMGA Models

	external set I		external set II	
no. of models	R^2	MAE	R^2	MAE
1	0.831	0.294	0.604	0.431
5	0.867	0.267	0.659	0.378
10	0.872	0.263	0.665	0.374
20	0.874	0.261	0.668	0.371
50	0.875	0.260	0.670	0.370

Using different parameters for model training further increased the total number of equivalent models. Nevertheless, it is not sufficient to ascribe the failure of QSAR models exclusively to the existence of too many equivalent models; the central problem appears to be that we often choose the wrong model from among them, which is why, despite of the availability of so many validation strategies, the true predictive power of QSAR models still cannot be guaranteed.

Nowadays, as the insufficiency of internal validation has been realized, external validation for QSAR modeling has been regarded as the only way to determine a reliable QSAR model. Unfortunately, this concept is often misunderstood, and it is common practice to consider the corresponding prediction accuracy as a proof of the true predictability of the model after validating the model on an external data set. Furthermore, it is often the case that an arbitrary independent data set from training samples or collected from other sources is used for external validation without considering its nature. Although it is well-known that there is no relationship between internal and external predictability,[30] there has been little emphasis on the relationship between external predictability and true predictability. Our study based on the *T. pyriformis* data set (Figure 2) shows that there seems to be no relationship between accuracies of two different external sets. This suggests that a high internal accuracy combined with a high external accuracy based on an arbitrarily selected external set is still insufficient proof of the true predictability of a QSAR model.

The real problem seems to be determining what external data set should be chosen and how to validate the QSAR model with it. To address this, applicability domain was considered, which has already been a major topic in the modeling, validation, and prediction process. However, even when we repeated the process and considered only external samples within the applicability domain, the results changed little (Figure S1 in the Supporting Information).

One possible explanation for this phenomenon is that the prediction result of a QSAR model is dependent on not only the applicability domain but also the distribution of data set in the chemical space. It is more likely that data sets with similar distributions in the chemical space tend to exhibit similar prediction performance (Figure S2 in the Supporting Information). For example, the prediction results of external set I in the *T. pyriformis* data set seem closer to the fitness values of the training set (5-fold cross-validation) than those of external set II (Figure 2) because the compounds in external set I were chosen in a similar way as in the training set. On the other hand, Wold and Dunn have stated that QSAR models normally only have local validity.[31] As a result, local models are often superior to global models in terms of prediction accuracy for similar compounds;[26,32] a global model that covers a more diverse range of chemical space tends to lose detailed information of certain local spaces and may show poor performance even for similar compounds in those areas. As such, it is readily comprehensible that a global model has diverse predictive power over different regions of the whole chemical space; in other words, a QSAR model may show different performance for data sets having different distribution. In this regard, a model's predictability may be better assessed by regions other than the whole chemical space.

There are several methods to obtain a more reliable model. One method is to improve the training process by avoiding the generation of too many irrelevant descriptors and models, and by minimizing chance correlation as much as possible. In this regard, new methods that are capable of extracting those most informative descriptors are desirable. Another method is to take advantage of the available equivalent models and counteract their variance by considering combinations of various candidate models. A third method is to develop new validation strategies to pick out the most powerful model from all possible models.

Two possible approaches presented in this study try to make use of the available equivalent descriptor sets and models. One approach is based on the assumption that those most used descriptors extracted from the equivalent models are more likely to be informative,[33] and the other approach is based on the hypothesis that higher accuracy can be achieved by using combinatorial or ensemble models[34] (Tables 5 and 6).

We found that it can be beneficial to construct models based on the most frequently selected descriptors (Tables 3 and 4). In our study, a number of models trained with different machine learning methods were developed based on the hERG data set. All five of the different machine learning methods produced higher accuracies (with an average improvement of about 3%) by using only the top 10 descriptors than by using all descriptors (Table 3). This shows that frequently selected descriptors are more likely to have associations with the hERG channel inhibition than randomly selected descriptors. In fact, seeking consistent descriptor sets for different machine learning methods is of interest. Dutta et al.[35] developed an ensemble feature selection method to reach a consistent descriptor set for multiple QSAR models, but at the cost of their accuracies. In contrast, our method produced not only consistent descriptor sets but also higher accuracies.

Another possible way to improve QSAR is by using combinatorial or ensemble models, which often produce higher accuracies than single models. A combinatorial model can be created by combining numerous homogeneous or heterogeneous (e.g., generated by different learners) models, and both can significantly improve the performance while substantially reducing variance. In fact, combinatorial models have been successfully applied to QSAR modeling in the past.[15,35−39]

Both approaches proposed in this study yield promising accuracies that are better than those yielded by their most powerful component models. Therefore, as the general goal of external validation is to assess the predictive power of available models and ultimately to choose the most powerful one, these approaches are preferable especially in cases where there are not enough samples available for external use.

5. CONCLUSIONS

The failure of QSAR models has captured much attention. From a practical point of view, this failure can arise from two aspects. On the one hand, the conventional training process often produces too many possible solutions whose validation results vary in a wide range and are difficult to manage. On the other hand, the available validation strategies are still not strong enough to guarantee the true predictability of available models. In this study, two potential approaches are proposed from the side of training to obtain more successful models, and possible reasons are also presented from the side of validation to explain the limitations of currently available external validation strategy. In any case, caution should be exercised when external validation is used to assess the predictive power of QSAR models, and further study is suggested to explore more effective training and validation strategies.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Tables S1 and S2 and Figures S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*College of Pharmaceutical Sciences, Zhejiang University, 388 YuHangTang Rd, Hangzhou, China 310058. Tel: +86 571 88208596. Fax: +86 571 88208428. E-mail: fanxh@zju.edu.cn.

■ ACKNOWLEDGMENT

We thank Don Ding and Dr. Hong Fang (NCTR/FDA, USA) for their help during the manuscript preparation. This work was supported by National S&T Major Project (2009ZX09311-002 and 2009ZX09502-012) and the Ministry of Science and Technology of China (2008ZX09202-006).

■ REFERENCES

(1) Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure−activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112* (12), 1249.

(2) Verma, J.; Khedkar, V.; Coutinho, E. 3D-QSAR in Drug Design-A Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95–115.

(3) Maggiora, G. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535.

(4) Johnson, S. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **2008**, *48* (1), 25–26.

(5) Stouch, T.; Kenyon, J.; Johnson, S.; Chen, X.; Doweyko, A.; Li, Y. In silico ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **2003**, *17* (4), 83–92.

(6) Doweyko, A. QSAR: dead or alive?. *J. Comput.-Aided Mol. Des.* **2008**, *22* (2), 81–89.

(7) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Modeling Robust QSAR. *J. Chem. Inf. Model.* **2006**, *46* (6), 2310–2318.

(8) Scior, T.; Medina-Franco, J.; Do, Q.; Martínez-Mayorga, K.; Rojas, Y.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009**, *16* (32), 4297–4313.

(9) Tetko, I.; Bruneau, P.; Mewes, H.; Rohrer, D.; Poda, G. Can we estimate the accuracy of ADME-Tox predictions?. *Drug Discovery Today* **2006**, *11* (15−16), 700–707.

(10) Zvinavashe, E.; Murk, A.; Rietjens, I. Promises and Pitfalls of Quantitative Structure- Activity Relationship Approaches for Predicting Metabolism and Toxicity. *Chem. Res. Toxicol.* **2008**, *21* (12), 2229–2236.

(11) Golbraikh, A.; Tropsha, A. Beware of q2!. *J. Mol. Graphics Modell.* **2002**, *20* (4), 269–276.

(12) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26* (5), 694–701.

(13) Roy, P.; Leonard, J.; Roy, K. Exploring the impact of size of training sets for the development of predictive QSAR models. *Chemom. Intell. Lab. Syst.* **2008**, *90* (1), 31–42.

(14) Li, Q.; Jørgensen, F.; Oprea, T.; Brunak, S.; Taboureau, O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharmaceutics* **2008**, *5* (1), 117–127.

(15) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; O berg, T.; Dao, P.; Cherkasov, A.; Tetko, I. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* **2008**, *48* (4), 766–784.

(16) Talete SRL, DRAGON for Windows (Software for Molecular Descriptor Calculations). Version 5.4, 2006, http://www.talete.mi.it/.

(17) Schultz, T.; Hewitt, M.; Netzeva, T.; Cronin, M. Assessing applicability domains of toxicological QSARs: definition, confidence in predicted values, and the role of mechanisms of action. *QSAR Comb. Sci.* **2007**, *26* (2), 238–254.

(18) Vapnik, V. *The nature of statistical learning theory*; Springer: 2000.

(19) Chang, C.; Lin, C. LIBSVM: a library for support vector machines. 2001, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

(20) Hopfinger, A.; Patel, H. *Application of genetic algorithms to the general QSAR problem and to guiding molecular diversity experiments*; Academic Press: San Diego, CA, 1996.

(21) Meffert, K.; Rotstan, N.; Knowles, C.; Sangiorgi, U. JGAP—Java Genetic Algorithms and Genetic Programming Package. URL: http://jgap.sourceforge.net/.

(22) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. The WEKA data mining software: An update. *ACM SIGKDD Explorations* **2009**, *11* (1), 10–18.

(23) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(24) Hanley, J.; McNeil, B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143* (1), 29.

(25) Sheridan, R. P. The Centroid Approximation for Mixtures: Calculating Similarity and Deriving Structure- Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1456–1469.

(26) Feher, M.; Ewing, T. Global or local QSAR: is there a way out? *QSAR Comb. Sci.* **2009**, *28* (8), 850–855.

(27) Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M.; McDowell, R.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111* (10), 1361.

(28) OECD Report from the Expert Group on (Q)SARs on Principles for the Validation of (Q)SARs. http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=env/jm/mono-(2004)24&doclanguage=en (Dec 10).

(29) Huang, J.; Ma, G.; Muhammad, I.; Cheng, Y. Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. *J. Chem. Inf. Model.* **2007**, *47* (4), 1638–1647.

(30) Kubinyi, H.; Hamprecht, F.; Mietzner, T. Three-Dimensional Quantitative Similarity- Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41* (14), 2553–2564.

(31) Wold, S.; Martens, H.; Wold, H. The multivariate calibration problem in chemistry solved by the PLS method. *Matrix Pencils* **1983**, 286–293.

(32) Yuan, H.; Wang, Y.; Cheng, Y. Local and Global Quantitative Structure- Activity Relationship Modeling and Prediction for the Baseline Toxicity. *J. Chem. Inf. Model.* **2007**, *47* (1), 159–169.

(33) Shao, L.; Wu, L.; Fan, X.; Cheng, Y. Consensus Ranking Approach to Understanding the Underlying Mechanism With QSAR. *J. Chem. Inf. Model.* **2010**, *50* (11), 1941–1948.

(34) Dietterich, T. Ensemble methods in machine learning. *Mult. Classifier Syst.* **2000**, 1–15.

(35) Dutta, D.; Guha, R.; Wild, D.; Chen, T. Ensemble feature selection: consistent descriptor subsets for multiple QSAR models. *J. Chem. Inf. Model.* **2007**, *47* (3), 989–997.

(36) Agrafiotis, D.; Cedeno, W.; Lobanov, V. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (4), 903–911.

(37) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1971–1978.

(38) Sanchez-Martin, R.; Mittoo, S.; Bradley, M. The impact of combinatorial methodologies on medicinal chemistry. *Curr. Top. Med. Chem.* **2004**, *4* (7), 653–669.

(39) Wang, X.; Tang, H.; Golbraikh, A.; Tropsha, A. Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F. *J. Chem. Inf. Model.* **2008**, *48* (5), 997–1013.