Consensus Ranking Approach to Understanding the Underlying Mechanism With QSAR

Li Shao, Leihong Wu, Xiaohui Fan,* and Yiyu Cheng*

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

Received August 9, 2010

Constructing a highly predictive model and exploiting the underlying mechanism associated with a specific property of chemicals are the two main goals of quantitative structure−activity relationship analysis (QSAR). However, the latter has long been carried out as a byproduct of model construction. Here we confirmed for the first time in this study that conventional descriptor selection methods designed to develop a best predictive model are likely not suitable for mechanistic analysis, i.e., the selected descriptors strongly depended on the selection of chemicals in the training sets. As an alternative, a consensus ranking protocol was proposed to select a robust descriptor set for mechanistic analysis, which can successfully overcome the above shortcoming. Moreover, the consistently inferior model performance using descriptors selected for mechanistic analysis suggested the irreplaceable role of model development in achieving models with the best predictive capability.

INTRODUCTION

Conceiving that the structural formula of an organic compound, in principle, contains codes within it all of the information which predetermines the chemical, biological, and physical properties of the compound, a major goal of quantitative structure−activity relationship analysis (QSAR) studies is to find a mathematical relationship between the property under investigation (e.g., LD_{50}, pK_a, etc.) and the one or more descriptive descriptors related to the structure of the molecule.[1,2] Thus acceptable QSAR models, defining the correlation and the quantitative prediction of chemical and physical properties from a structure, finally can guide the synthetic chemist in the choice between alternative hypothetical structures. More fundamentally, such studies can illuminate and even elucidate the underlying mechanism by which the property in question is related to the chemical structure.[2] Nowadays, QSAR has been widely utilized as a useful strategy in both research and regulatory applications, such as drug design and high-throughput activity/toxicity screening.[3−5]

With the advancement of QSAR studies as well the descriptor-generating software packages, conducting QSAR investigations on the basis of a relatively small number of chemicals with thousands of descriptors is not a trivial work.[6−8] To date, many demonstrable successes[9−12] on the QSAR study have driven extensive research over the past decades. However, a majority of these studies focused on the development of an accurate predictive model with thousands of descriptors, making descriptor-based mechanism interpretation a byproduct of them. In other words, most if not all of the studies carried out mechanism interpretation based on descriptors selected for developing the so-called best predictive model, whereas few efforts have been made to evaluate whether those descriptors selected by the single model solely tuned by prediction accuracy were reliable ones for further mechanistic analysis. More importantly, how to select robust and informative descriptors to interpret the underlying mechanism of chemicals remains to be a very challenging problem.

Thus we made the first step in this study to evaluate whether the current application of mechanistic analysis on the basis of descriptors aiming to develop a best model is suitable by comparing the overlap of descriptors selected from the different training sets resampled from the same data set. Widely used machine learning algorithms, including the enhanced replacement method (ERM)[13,14] and the stepwise multivariate linear regression (MLR)[15] have been employed in this study. These results show that the list of descriptors selected by current methods was highly unstable and strongly depended on the selection of chemicals in the training sets, which will absolutely result in different interpretations in terms of mechanism. We further proposed a novel consensus ranking method to select robust descriptors for mechanistic analysis, which could successfully overcome the above shortcomings. Moreover, the inferior performance of models using descriptors selected for mechanistic analysis confirmed the irreplaceable role of model development in achieving models with the best predictive capability.

MATERIALS AND METHODS

Experimental Data and Calculation of Molecular Descriptors. Six QSAR data sets are selected and utilized in this study, consisting of three toxicity and three activity end points referred to as QT1, QT2, QT3, DVEGF, DEGF, and DHIV in this paper, respectively.[15−23]

The toxicity data sets have been taken from ref 16, which in total represents a comprehensive toxicity data set consisting of 1093 compounds. The toxicity of each compound is expressed as the growth inhibition of the ciliated protozoan *T. pyriformis*, that is the inverse logarithm of 50% inhibition of growth concentration ($pIGC_{50}$) values. Considering the computation burden, the training set (644 compounds) and

* Corresponding authors. E-mail: fanxh@zju.edu.cn (X.F.) and chengyy@zju.edu.cn (Y.C.). Telephone: +86-571-88208596 (X.F.) and +86-571-87951138 (Y.C.).

10.1021/ci100305g
Published on Web 11/04/2010

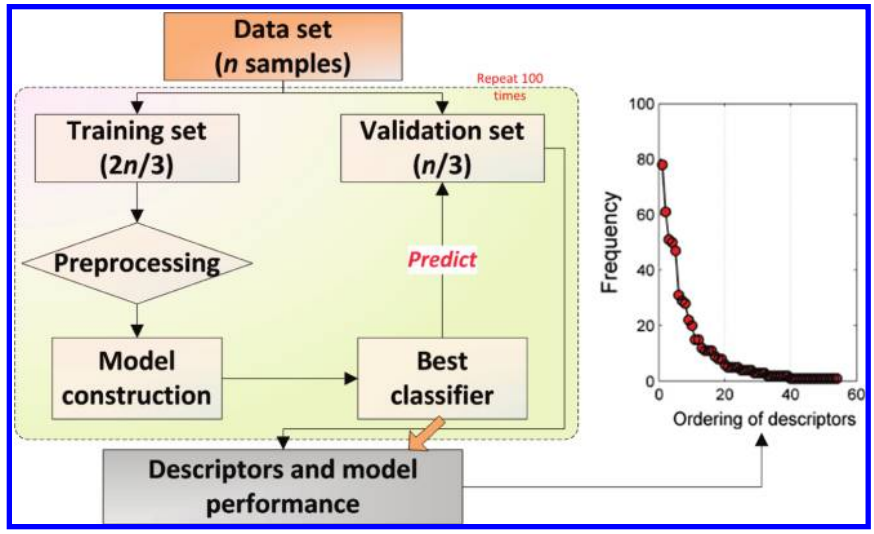

Figure 1. Work flow for the consensus ranking protocol.

two validation data sets (339 and 110 compounds) proposed in the original paper,[16] which consists of similar fractions of compounds with low, intermediate, and high toxicity values, have been utilized as three separate data sets (QT1, QT2, and QT3) in this study for simplicity. Based on the similes files provided in the original paper, Dragon version 5.4 was utilized to calculate 929 meaningful descriptors for each compound.

Data set DVEGF was obtained from ref 15, which contained 74 molecules of diaryl ureas together with their inhibitory activities against vascular endothelial growth factor receptor (VEGFR-2) kinase. Data set DEGF was retrieved from refs 17−20 consisting of 69 molecules and corresponding epidermal growth factor receptor (EGFR) inhibitory activities, while the 101 molecules of data set DHIV belonged to the human inmmunodeficiency virus type-1 (HIV-1) integrase inhibitors.[21−23] Generally, the inhibitory activities are expressed as $\log(1/IC_{50})$. Detailed information about the data sets could be obtained in corresponding references.

The structures of the compounds in the above three data sets were drawn using ISIS/Draw implemented in the ISIS 2.5 package and optimized sequentially using the molecular mechanics force fields (MM+) encoded in HyperChem (version 8.04, Hypercube, Inc.) and the semiempirical method PM3 (Parametric Method-3)[24] with a gradient norm limit of 0.01 kcal/Å. Altogether 1664 meaningful descriptors were calculated for each compound using Dragon version 5.4,[25] encoding different aspects of the molecular structures, such as constitutional, topological, electronic, thermodynamic, geometric descriptors, etc.

For model construction and mechanistic analysis, common preprocessing procedures shown as follows are utilized: First, descriptors with the same entries for most of the training compounds (90%) were removed from the pool of variables considered; and then, pairs of variables with a correlation coefficient greater than 0.90 were classified as intercorrelated, and one of them with lower correlation with property end points was deleted. The resulting data matrices were used for further analysis.

Study Design. The study is designed as follows: First, we confirmed that the carrying out of the mechanism interpretation based on descriptors selected for developing the so-called best predictive model was not suitable. Then, a novel consensus ranking protocol was proposed in this study to accomplish a reliable descriptor selection for mechanistic analysis. Finally, model performance based on descriptors selected from model construction and mechanistic analysis procedures were evaluated and compared, with the aim to confirm that descriptors for mechanistic analysis may not be the best ones with respect to predictive capability. In other words, the model development procedure is indispensable in achieving models with the best predictive capability.

Assessment of Current Descriptor Selection Methods. The unfitness of carrying out mechanistic analysis as a byproduct of model construction has been evaluated in this study from two aspects. As a direct indication of the shortness of current descriptor selection methods, we first evaluated the overlap between best performed descriptor lists with different lengths obtained from the same data set; and then, the vulnerability of current descriptor selection methods to different training samples was assessed by evaluating the percentage of overlap for markers obtained from different training chemicals ($2n/3$) randomly selected from the original data set (n samples). Two widely utilized machine learning algorithms, including ERM and MLR, are selected in this study. Here the overlap degree represents the stability of the method.

Consensus Ranking Protocol for Reliable Mechanistic Analysis. Taking mechanistic analysis as an independent task in QSAR studies, we have proposed in this study a novel consensus ranking protocol for a reliable descriptor selection based on ERM[13](Figure 1). As a simulation of real-world scenario to apply the QSAR study, the data sets were divided into two sets, i.e., the training ($2n/3$) and validation ($n/3$) sets. Generally, the analysis protocol starts in developing a best classifier on the training set and ends in predicting the validation set, where corresponding prediction performance and descriptors selected in the best classifiers are recorded. To ensure the statistical validity and to provide an index of reliability, we repeat this procedure 100 times, resulting in 100 best classifiers and corresponding descriptors.

Detailed procedures about model construction and validation are illustrated as follows: First, common preprocessing procedures are carried out on the training set, which remove constant and highly related descriptors. Then the resulting data matrices were utilized to construct the best classifiers

using ERM.[13] Finally, the performance of the best classifiers is evaluated on the validation samples. Note that descriptors selected in the best classifiers and the model performance on validation sets in each repetition are recorded for following analysis.

For consensus ranking protocol, all the descriptors selected in the best classifiers with appropriate predictive capability are first rank ordered according to the frequency that each was selected as a signature descriptor in the validation processes. Considering that the higher the frequency a descriptor is selected and the more important the descriptor might be,[26] the descriptors with the highest frequencies would most probably be the potential markers for mechanistic analysis.

Prediction Capability of Descriptors Used for Mechanistic Analysis. It is common sense that a model construction aims to develop highly predictive models, while mechanistic analysis struggles to seek for potential markers to infer the structural characteristics associated with the end point property helpful for later molecular design and/or modifications. Thus models developed from descriptors for mechanistic analysis might not perform as well as those constructed mainly for prediction aims theoretically. Thus we further evaluated and compared the performance of ERM models using descriptors selected from model construction and mechanistic analysis, as a confirmation of the irreplaceable role of model construction in achieving models with best predictive capability.

Stepwise MLR. For simplicity and interpretability, stepwise MLR, a commonly used method in QSAR studies, was employed for model development in this study, where the best descriptors leading to the smallest standard deviations are added into the model step by step, until there is no other variable outside the equation that satisfies the selection criterion.

ERM. ERM[13] proposed by Andrew G. Mercader et al. aims to choose an optimal subset of d ($d < D$) descriptors from the pool of D descriptors or to minimize the following standard deviation S':

$$S' = \frac{1}{(N - d - 1)} \sum_{i=1}^{N} \mathrm{res}_i^2 \tag{1}$$

where N and res_i are the number of molecules in the training set and the residual for the ith molecule (i.e., the difference between experimental and predicted property), respectively. Generally, ERM is able to search for the global minimum more efficiently than the full search. The procedures are shown concisely as follows: First, this algorithm chooses an initial set of descriptors $\mathbf{d}_k$ (k is the total number of descriptors) at random, replaces one of the descriptors, say X_{ki} (the ith one in k descriptors) with all the remaining $D - d$ descriptors one by one, and keeps the set resulting in the smallest value of S'. That is what we called a 'step'; then, from the resulting set we substitute the descriptor with the greatest standard deviation in its coefficient, one by one, with all the remaining $D - d$ descriptors for it. This procedure is repeated until the set remains unmodified. Note that in each cycle, this algorithm does not modify the descriptor optimized in the previous one. Thus, we obtain the candidate $\mathbf{d}_m(i)$ that comes from the so-constructed path i. Note that if the replacement of the descriptor with the largest error by those in the pool does not decrease the value of S', the ERM chooses the next smallest value S', which is of great help for getting out of a local S' minimum.

Performance Assessment. The overall performance of models was evaluated by measuring the prediction R^2 (explained variance), root-mean-square error (RMSE), and standard deviation error of prediction (SPRESS), calculated from the following equations:

$$R^2 = 1 - \frac{\sum (y_{\mathrm{exp}} - y_{\mathrm{pred}})^2}{\sum (y_{\mathrm{exp}} - y_{\mathrm{m}})^2} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n_s} (y_{\mathrm{exp}} - y_{\mathrm{pred}})^2}{n_s}} \tag{3}$$

$$\mathrm{SPRESS} = \sqrt{\frac{\sum (y_{\mathrm{exp}} - y_{\mathrm{pred}})^2}{n_s - k - 1}} \tag{4}$$

In the above equations, y_{exp}, y_{pred}, and y_{m} represent the experimental, predicted, and mean of the dependent variable, respectively, while n_s is the number of molecules in data set, and k is the number of independent variables in regression equation.

RESULTS

Unfitness of Current Mechanistic Analysis. For mechanistic analysis, the most important role of feature selection is to select robust and informative descriptors that are related to the property end point. Thus we assessed the fitness of current mechanistic analysis by evaluating the robustness of descriptors selected from the same and different training samples. Considering the statistical rule that the number of descriptors should not be larger than 5 times the number of samples[27,28] and the computation complexity, lists of 2−14 descriptors were selected for each of the data sets.

As the most direct way to assess the capability of conventional feature selection methods in selecting robust markers for mechanistic analysis, we first evaluated the stability of descriptor lists with best prediction performance and different lengths selected from the same data set. As shown in Table 1, where descriptors were selected from the six original data sets using ERM, an apparent difference was observed for descriptor lists with different lengths. In other words, descriptor sets with the best performance would vary with the length of lists, implying the shortness of the current feature selection methods designed for model construction in selecting markers for mechanistic analysis.

Figure 2a and b illustrates the percentage of overlap versus the different length of descriptor lists between two sets of compounds ($2n/3$) randomly retrieved from it based on ERM and MLR, respectively. At least the following information could be obtained: First, the overlap between features selected from the two sets of randomly retrieved training chemicals is consistently low, evidenced from the percentage less than 50% for most cases; and then, contrast to the consistently poor overlap for lists with seven or more descriptors, the overlap seems to be much more variable for feature lists with fewer than six descriptors, with the overlap skipping from 0

Table 1. Descriptors Lists with Different Lengths Selected from Six Original Data Sets using ERM

	ALOGP	AAC										
	ALOGP	TPSA(Tot)	Mp									
	ALOGP	TPSA(Tot)	AMW	Br-094								
	ALOGP	SIC0	VEp1	nRNCS	nRNH2							
	ALOGP	SIC0	nArOH	MAXDP	nDB	Sv						
	ALOGP	SIC0	nArOH	MAXDP	nDB	N-066	BEHv7					
QT1	ALOGP	BIC0	nArOH	MAXDP	nDB	nRNH2	CIC4	ATS1p				
	ALOGP	BIC0	nArOH	MAXDP	nDB	nRNH2	ESpm01r	Br-091	SIC4			
	ALOGP	BIC0	nArOH	MAXDP	nDB	nRNH2	ESpm01r	Br-091	IC4	MATS2v		
	ALOGP	BIC0	nArOH	MAXDP	nDB	nRNH2	ESpm01r	Br-091	IC4	MATS2v	C-018	
	ALOGP	BIC0	nArOH	MAXDP	nDB	nRNH2	ESpm01r	Br-091	SIC4	MATS2v	C-018	MLOGP2
	ALOGP	BIC0	nArOH	MAXDP	nDB	N-066	ESpm01r	Br-091	CIC4	MATS2v	C-018	MLOGP2
	ALOGP	BIC0	nArOH	MAXDP	C-038	nRNH2	ESpm01r	Br-091	IC4	ATS2v	C-018	nRNCS
QT2	BLTF96	TPSA(Tot)										
	BLTD48	nArOH	nDB									
	BLTF96	nArOH	nDB	Br-091								
	BLTF96	nArOH	nDB	Br-091	Hy							
	BLTF96	nArOH	TPSA(Tot)	Br-091	H-052	C-018						
	BLTF96	nArOH	TPSA(Tot)	Br-091	H-052	C-018	nRNCS					
	ALOGP	BIC0	C-040	WA	O-056	C-018	ATS3m	S-108				
	ALOGP	BIC0	C-040	WA	O-056	C-018	ATS3m	S-108	GATS1m			
	BLTD48	nArOH	nRCO	S2K	O-056	C-018	ATS3m	S-108	H-051	MATS1p		
	BLTF96	GATS1e	C-040	Ms	O-056	C-018	XMOD	S-108	H-051	GATS1m	H-046	
	MLOGP	GATS1e	C-040	WA	O-056	C-018	ATS1m	S-108	H-051	GATS1m	H-046	nH
	MLOGP	GATS1e	C-040	Mv	O-056	C-018	ATS1m	S-108	H-051	GATS1m	H-046	TI2
	MLOGP	GATS1e	nROR	Rww	O-056	C-018	ATS1m	S-108	H-051	GATS1m	H-046	nH
	BLTF96	nDB										
	BLTD48	nDB	nTB									
	BLTD48	nH	BELp4	T(F···Cl)								
	BLTD48	nH	BELp4	nR#C-	MATS2p							
	ALOGP	nDB	EPS1	nR#C-	ATS1p	F-084						
	ALOGP	nArNH2	EPS1	nCRX3	nRNCS	BEHp2	Ss					
QT3	ALOGP	nArNH2	EPS1	nR#C-	nRNCS	BEHp2	Ss	F-083				
	ALOGP	nArNH2	EPS1	nR#C-	nRNCS	BEHp2	Ss	nRNHO	T(F···Cl)			
	ALOGP	nArNH2	EPS1	nR#C-	nRNCS	BEHp2	Ss	nRNHO	MATS7v	nCRX3		
	ALOGP	piPC09	BEHm5	nR#C-	nRNCS	BEHp2	Ss	IDM	MATS7v	T(Cl···Cl)	F-083	
	ALOGP	nRNHO	S1K	MPC10	O-056	DELS	nRNCS	AMR	MATS7v	nCRX3	C-022	MSD
	ALOGP	piPC08	S1K	nR#C-	nRNCS	DELS	nRNHO	AMR	MATS7v	PCR	F-083	AECC
	ALOGP	piPC08	S1K	nR#C-	nRNCS	DELS	nROH	AMR	GATS7v	T(F···Cl)	piPC09	nR=Cp
DVEGF	PCD	C-027										
	GATS6p	ALOGP	JGI8									
	MATS3p	G(N···N)	PCR	RDF075m								
	GATS6p	ALOGP	PCR	RDF120p	TIC2							
	GATS6p	C-027	PCD	RDF120p	BEHe5	LP1						
	GATS6p	C-027	PCD	RDF120p	BEHe5	LP1	IC3					
	GATS6p	C-027	PCD	RDF120p	BEHe5	LP1	IC3	RDF055m				
	MATS6v	G(N···N)	PCR	RDF075m	MATS3p	IDDE	IC5	HATS0m	inflammat-80			
	GATS6p	C-027	PCR	RDF075m	BEHe5	RDF120m	CIC3	RDF060v	inflammat-80	R3m+		
	GATS6p	C-027	PCR	RDF135p	BEHe5	RDF120m	CIC3	RDF145p	T(O···F)	ATS6m	BEHv3	
	GATS6p	C-027	PCR	RDF075m	BEHe5	RDF120m	CIC3	RDF060v	H3m	H8p	Mor22e	MATS3m
	GATS6p	C-027	PCR	RDF075u	BEHe5	RDF120m	CIC3	RDF145p	H3m	H8p	RDF135p	T(O···F)
	GATS6p	C-027	PCR	RDF115m	BEHe5	RDF120m	IC3	RDF135p	T(O···F)	Mor03e	X5sol	G(N···Cl)
DEGF	G(N···Cl)	Hy										
	G(N···Cl)	Hy	BIC2									

Table 1. Continued

	G(N···Cl)	Hy	H-051	RDF120u									
	G(N···Cl)	Hy	H-051	RDF120u	LAI								
	G(N···Cl)	Hy	H-051	RDF120u	LAI	PHI							
	G(N···Cl)	Hy	H-051	RDF120u	Mor24m	E1s	HTm						
	G(N···Cl)	Hy	H-051	RDF120u	Mor24m	E1s	HTm	LAI					
	G(N···Cl)	G2e	H-051	RDF120m	Mor24m	E1s	Dm	Mor27p	GATS2e				
	G(N···Cl)	AAC	N-072	RDF150m	Mor24m	C-027	E1s	Mor06e	TPSA(NO)	RDF020e			
	G(N···Cl)	AAC	O-056	RDF150m	Mor24m	C-027	E1s	Mor06e	TPSA(NO)	RDF020e	N-072		
	G(N···Cl)	Hy	Mor27p	Mor17e	Mor24m	G3p	E1s	Mor06e	BIC2	EEig02d	RDF120u	Mor23m	
	G(N···Cl)	Hy	Mor27p	Mor17e	Mor24m	G3p	E1s	GGI10	SPH	MATS1m	RDF120u	Mor23m	Mor10p
	T(O···Cl)	G2e	Mor27p	Mor23p	Mor24m	C-027	E1s	MAXDP	MATS6e	nROR	H-051	Mor05u	GATS2e
DHIV	D/Dr10	LP1											
	PW5	EEig02d	nCIR										
	D/Dr10	MATS5p	GATS8m	C-003				Mor24m					
	D/Dr10	MATS5p	GATS8m	C-003	RDF130e			Mor24m	BELm2				
	D/Dr10	MATS5p	GATS8m	C-003	RDF130e	RDF135m			GATS8p	Me			
	D/Dr10	MATS5p	GATS8m	C-003	RDF130e	RDF135m	RDF085m						
	D/Dr10	MATS5p	GATS8m	C-003	RDF130e	RDF135m	RDF085m	Mor24m					
	D/Dr10	MATS5p	GATS8m	C-003	RDF130e	RDF135m	RDF085m	Mor24m	BELm2				
	D/Dr10	GATS5e	GATS8m	C-003	RDF130e	RDF135m	RDF065e	BELv2	GATS8p	Me			
	D/Dr10	GATS5e	GATS8m	C-003	RDF130e	RDF135m	RDF065e	BELv2	GATS8p	Me	N-071		
	D/Dr10	GATS5e	GATS8m	C-003	RDF130e	RDF135m	RDF065e	BELv2	GATS8p	Me	LP1	infective-50	
	D/Dr10	GATS7p	GATS8m	R6u+	RDF020m	RDF135m	Mor24m	EEig04d	GATS8p	N-071	LP1	infective-50	neoplastic-50
	D/Dr10	GATS7p	GATS8m	C-003	RDF020m	RDF135m	Mor24m	EEig04d	GATS8p	N-071	LP1	infective-50	neoplastic-50

to 1 in different data sets. In a word, the descriptors selected from different training samples are quite inconsistent, which will absolutely result in different interpretations in terms of mechanistic analysis.

Consensus Ranking Protocol for Mechanistic Analysis. Conforming to the rules mentioned above, we proposed a novel consensus ranking protocol in this study for mechanistic analysis, which is based on multiple repetitions of sample splitting and model construction. To ensure the capability of selected descriptors in covering the properties of the whole data set, we omit descriptors in models with validation R^2 lower than 0 for further analysis. In this protocol, the frequency of a descriptor selected as predictive signatures indicates the reliability of it as a marker for mechanistic analysis.

Figure 3 illustrates the frequency of each descriptor selected as potential markers in the 100 repetitions for 6 data sets. Generally, a drastic decrease is observed, especially for the top 10 descriptors. Most strikingly, we could see that the top two and three descriptors in QT1 and QT2 would be much more robust and reliable to be selected as markers for mechanistic analysis, evidenced from the significantly higher frequency (≥ 60). After a close inspection, we further found that the top descriptors listed above are all involved in Table 1, the descriptors selected in feature lists with different lengths, confirming the confidence of utilizing these top descriptors as markers for mechanistic analysis.

The Indispensable Role of Model Construction. The above findings confirmed the unfitness of carrying out mechanistic analysis as a byproduct of model construction. In this section, we further evaluated whether descriptors selected for mechanistic analysis are the best ones with respect to predictive capability, by comparing the performance of models using descriptors selected from current methods and consensus ranking protocol. For brevity, a set of six descriptors was used for model comparison in this study, a number widely utilized in QSAR studies.[11]

Table 2 illustrates the leave-one-out cross validation results including R^2, RMS, and SPRESS for the six data sets using descriptors selected from the conventional method and the consensus ranking protocol. Corresponding information for the utilized descriptors is given in Table 1 of the Supporting Information. We can see that models using descriptors selected for mechanistic analysis are consistently superior to those based on descriptors aimed for model construction.

DISCUSSION

Constructing highly predictive models and selecting reliable descriptors for mechanistic analysis are two main aims in QSAR studies. Considering that descriptor-based mechanism analysis has long been carried out as a byproduct of model construction, we evaluated and confirmed for the first time in this study that the current application of mechanistic analysis on the basis of descriptors designed to develop a best predictive model are not suitable for mechanistic analysis. As an alternative, a consensus ranking protocol was proposed to select a robust descriptor set for mechanistic analysis. Moreover, the consistently inferior model performance using descriptors selected for mechanistic analysis suggested the irreplaceable role of model development in achieving models with the best predictive capability.

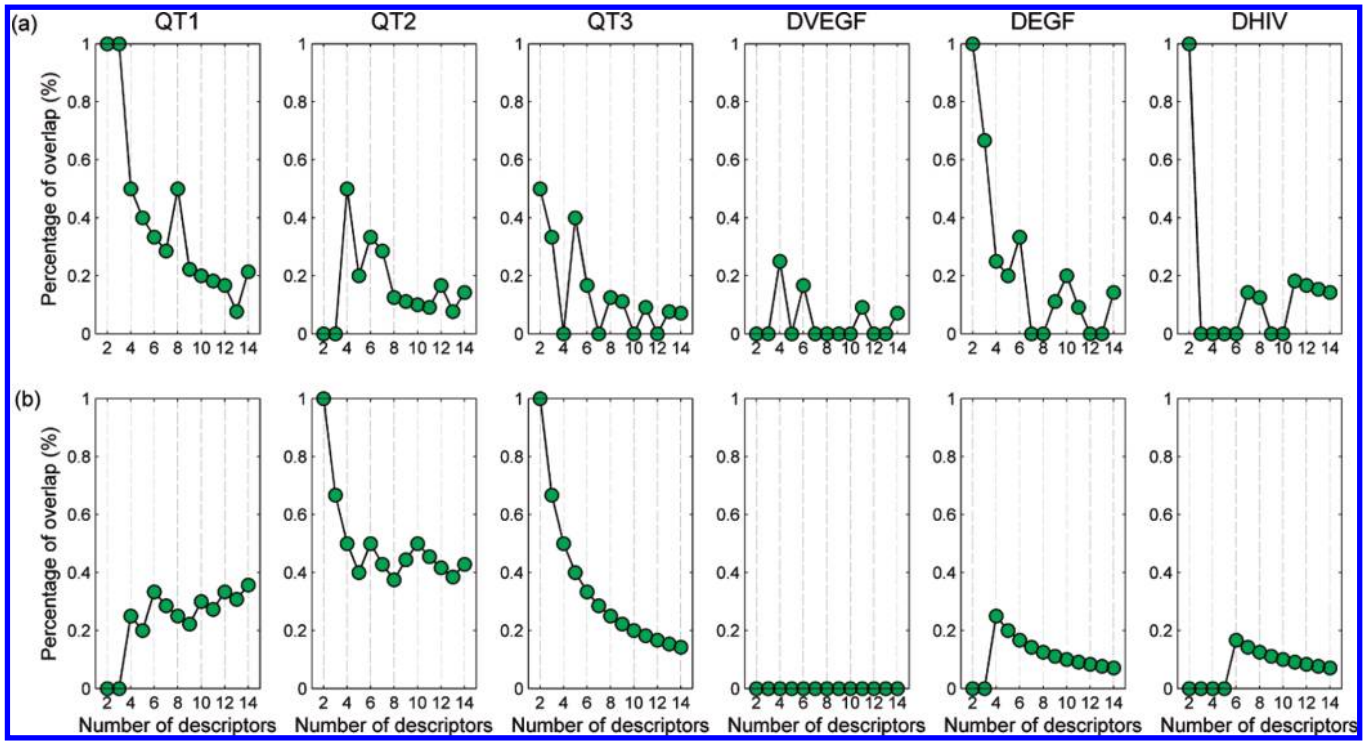

Figure 2. Percentage of overlap for descriptor lists with different lengths selected from two sets of chemicals randomly retrieved from original data sets using: (a) ERM and (b) MLR.

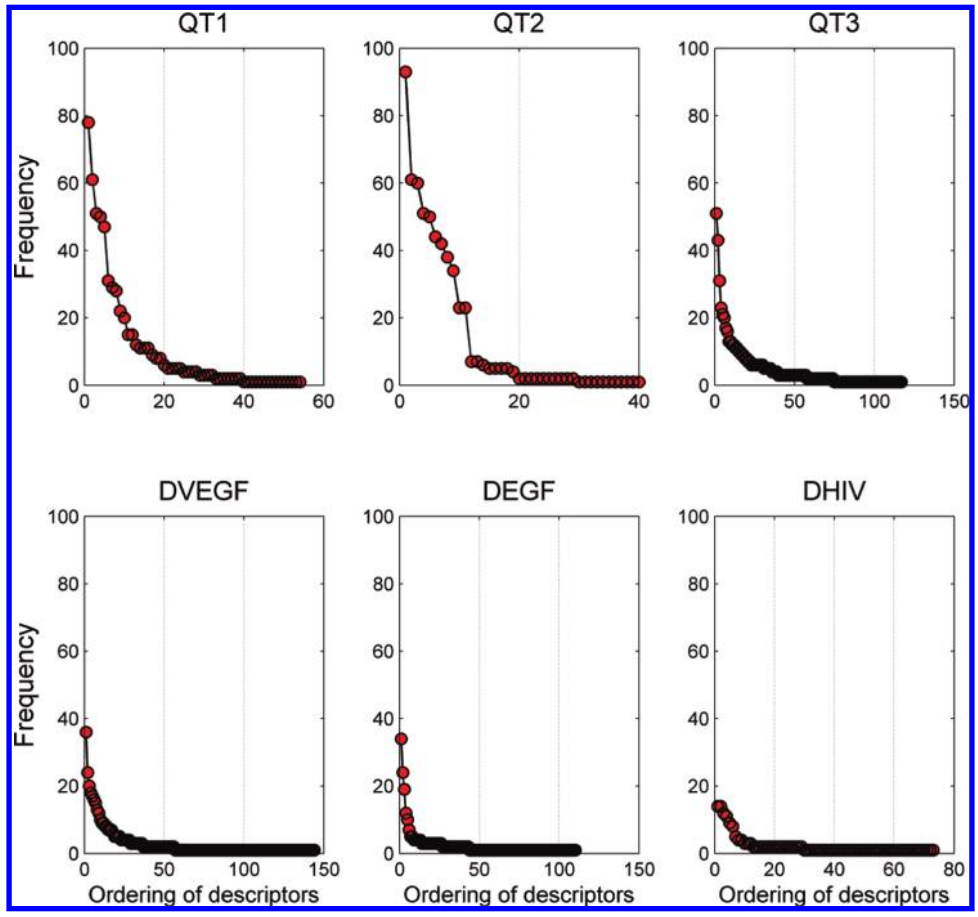

Figure 3. Frequency of descriptors selected as potential markers in the consensus ranking protocol for six data sets.

From Figure 2a and b, we can see that current feature selection methods designed for model construction could result in highly variable descriptors and thus are not suitable for mechanistic analysis. Such a conclusion was further confirmed by Table 1, where apparent variability exists for descriptors selected from the same data set. The exact reason for this phenomenon is out of the scope of this study. However, we suppose that it could to some extent be

Table 2. Leave-One-out Cross Validation Results for Six Data Sets Using Descriptors Selected from Single ERM and the Consensus Ranking Protocol

	single ERM			consensus ranking protocol		
	R^2	RMS	SPRESS	R^2	RMS	SPRESS
QT1	0.791	0.482	0.485	0.774	0.502	0.504
QT2	0.812	0.455	0.459	0.807	0.460	0.465
QT3	0.756	0.440	0.455	0.654	0.523	0.541
DVEGF	0.792	0.347	0.365	0.607	0.477	0.502
DEGF	0.663	0.361	0.381	0.656	0.365	0.385
DHIV	0.540	0.430	0.446	0.439	0.475	0.492

attributed to the high-variable dimension and relative small sample size, which may result in multiple combinations of descriptors with similar model performance. Furthermore, the inconsistency of descriptors selected from different training samples implies also the heterogeneous nature of training samples which should be considered as an important factor in selecting reliable markers for mechanistic analysis. After a close inspection, we found that despite the poor overlap for feature lists with different lengths, lists containing six or fewer descriptors are much more variable with the overlap skipping from 0 to 1 in different data sets. Since considerable QSAR studies retain six or fewer descriptors for further analysis to ensure the statistical capability of models in real applications,[7,9,11] the above conclusions emphasized the importance of developing novel and robust marker selection protocols aimed for mechanistic analysis, which should also take the heterogeneous nature of training samples into consideration simultaneously.

How to select reliable biomarkers associated with specific biological status such as diseases has long been the major difficulty in QSAR and also omics studies, where poor overlap of biomarkers selected from different studies have been reported.[29,30] Although not many pragmatic protocols have been proposed to date due to the small sample size and the high-variable dimension, some studies[31] strived to interpret the reliability of a biomarker using a robustness measure, i.e., the frequency of a gene or a metabolite being selected as a biomarker. In other words, the higher the frequency a feature is selected, the more reliable it might be as a biomarker. Inspired from it, we proposed a consensus ranking protocol in this study, which has been successfully applied to select robust biomarkers in metabolomic studies.[32]

The superiority of the consensus ranking protocol could be summarized as follows: First, it can take the information involved in all compounds into consideration rather than that in randomly selected training samples, leading to more robust and reliable descriptors for mechanistic analysis. Taking QT1 in Figure 3 as an example, a list of 54 descriptors would be selected as signature features within different descriptor lists resulting from the best training models. If the conventional protocols are utilized, then the final descriptors used for mechanistic analysis would be random subsections of the 54 descriptors, leading to the highly variable nature of the selected descriptors. Second, it could provide an index of robustness based on frequency through the repetitive split of training samples and following descriptor selection. Such information is also of great help when considering the relative contribution of the selected descriptors to the studied end point property. Thus we could conclude that besides a practicable index of the reliability of a descriptor as potential markers for mechanistic analysis, the novel consensus ranking protocol proposed in this study could also take into consideration the heterogeneous property of modeling samples and the associated impact of random splitting of training and validation sets. In other words, multiple repetitions utilized in this protocol could also preclude the impact caused by random selection of training samples used by a majority of QSAR studies.[7,9]

We further evaluated whether descriptors selected for mechanistic analysis could perform as well as those selected in model construction by comparing the performance of models using descriptors selected in the above two procedures. Based on the novel consensus ranking protocol proposed in this study, consistently inferior predictive capability was observed for models constructed from descriptors for mechanistic analysis. This is not surprising to us, since the model construction procedure should theoretically select the best set of descriptors in terms of model performance. Thus we could conclude that mechanistic analysis could not replace the model construction procedure in achieving models with the best predictive capability either.

We proposed for the first time in this study that conventional descriptors selected in the single best predictive model are highly vulnerable to different training samples and could not afford reliable mechanistic analysis. The consensus ranking protocol could successfully overcome the above shortcoming, provide a way for obtaining robust descriptors, and also the relative contribution of the selected descriptors to the studied end point property. Furthermore, the inferior performance of descriptors selected by consensus ranking protocol confirmed the indispensable role of model construction in achieving highly predictive models. In contrast to the massive efforts paid in improving model performance, further studies should cast more attention to develop novel protocols for reliable mechanistic analysis.

CONCLUSION

We set the first step in this study confirming that mechanistic analysis should not be carried out as a byproduct of model construction from the high vulnerability of current descriptor selection methods designed to develop a best predictive model for different training samples. As the first endeavor, the consensus ranking protocol proposed in this study could successfully overcomes the shortcoming of conventional descriptor selection methods and provides a reliable way for mechanistic analysis. In contrast to the massive studies struggling to improve model predictive capabilities, more attention should be paid to reliable mechanistic analysis and to develop corresponding descriptor selection protocols in future studies.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China (no. 30801556), the Zhejiang Province Science and Technology Plan Project (no. 2008C23065 and R2080693), and the Research Fund for the Doctoral Program of Higher Education of China (no. 20090101110126).

Supporting Information Available: Corresponding information for the utilized descriptors is given. This material

is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Katritzky, A. R.; Karelson, M.; Lobanov, V. S. QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure Appl. Chem.* **1997**, *69*, 245–248.

(2) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR - The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287.

(3) Fjodorova, N.; Novich, M.; Vrachko, M.; Smirnov, V.; Kharchevnikova, N.; Zholdakova, Z.; Novikov, S.; Skvortsova, N.; Filimonov, D.; Poroikov, V.; Benfenati, E. Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. *J. Environ. Sci. Health, Part C: Environ. Carcinog. Ecotoxicol. Rev.* **2008**, *26*, 201–236.

(4) Salum, L. B.; Andricopulo, A. D. Fragment-based QSAR strategies in drug design. *Expert Opin. Drug Discovery* **2010**, *5*, 405–412.

(5) Tsygankova, I. G. Variable selection in QSAR models for drug design. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 132–142.

(6) Fatemi, M. H.; Malekzadeh, H. Prediction of Log(IGC(50))(−1) for Benzene Derivatives to Ciliate Tetrahymena pyriformis from Their Molecular Descriptors. *Bull. Chem. Soc. Jpn.* **2010**, *83*, 233–245.

(7) Garkani-Nejad, Z.; Ahmadi-Roudi, B. Modeling the antileishmanial activity screening of 5-nitro-2-heterocyclic benzylidene hydrazides using different chemometrics methods. *Eur. J. Med. Chem.* **2010**, *45*, 719–726.

(8) Perez-Garrido, A.; Helguera, A. M.; Rodriguez, F. G.; Cordeiro, M. QSAR models to predict mutagenicity of acrylates, methacrylates and alpha, beta-unsaturated carbonyl compounds. *Dent. Mater.* **2010**, *26*, 397–415.

(9) Duchowicz, P. R.; Vitale, M. G.; Castro, E. A.; Autino, J. C.; Romanelli, G. P.; Bennardi, D. O. QSAR modeling of the interaction of flavonoids with GABA(A) receptor. *Eur. J. Med. Chem.* **2008**, *43*, 1593–1602.

(10) Li, J. Z.; Du, J.; Xi, L. L.; Liu, H. X.; Yao, X. J.; Liu, M. C. Validated quantitative structure-activity relationship analysis of a series of 2-aminothiazole based p56(Lck) inhibitors. *Anal. Chim. Acta* **2009**, *631*, 29–39.

(11) Mercader, A. G.; Duchowicz, P. R.; Fernandez, F. M.; Castro, E. A.; Bennardi, D. O.; Autino, J. C.; Romanelli, G. P. QSAR prediction of inhibition of aldose reductase for flavonoids. *Bioorg. Med. Chem.* **2008**, *16*, 7470–7476.

(12) Sharma, S.; Prabhakar, Y. S.; Singh, P.; Sharma, B. K. QSAR study about ATP-sensitive potassium channel activation of cromakalim analogues using CP-MLR approach. *Eur. J. Med. Chem.* **2008**, *43*, 2354–2360.

(13) Mercader, A. G.; Duchowicz, P. R.; Fernandez, F. M.; Castro, E. A. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 138–144.

(14) Sun, M.; Zheng, Y. G.; Wei, H. T.; Chen, J. Q.; Cai, J.; Ji, M. Enhanced Replacement Method-based Quantitative Structure - Activity Relationship Modeling and Support Vector Machine Classification of 4-Anilino-3-quinolinecarbonitriles as Src Kinase Inhibitors. *QSAR Comb. Sci.* **2009**, *28*, 312–324.

(15) Sun, M.; Chen, J. Q.; Wei, H. T.; Yin, S. Q.; Yang, Y.; Ji, M. Quantitative Structure-Activity Relationship and Classification Analysis of Diaryl Ureas Against Vascular Endothelial Growth Factor Receptor-2 Kinase Using Linear and Non-Linear Models. *Chem. Biol. Drug Des.* **2009**, *73*, 644–654.

(16) 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.

(17) Mastalerz, H.; Chang, M.; Chen, P.; Fink, B. E.; Gavai, A.; Han, W. C.; Johnson, W.; Langley, D.; Lee, F. Y.; Leavitt, K.; Marathe, P.; Norris, D.; Oppenheimer, S.; Sleczka, B.; Tarrant, J.; Tokarski, J. S.; Vite, G. D.; Vyas, D. M.; Wong, H.; Wong, T. W.; Zhang, H. J.; Zhang, G. F. 5-((4-Aminopiperidin-1-yl)methyl)pyrrolotriazine dual inhibitors of EGFR and HER2 protein tyrosine kinases. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4947–4954.

(18) Mastalerz, H.; Chang, M.; Chen, P.; Dextraze, P.; Fink, B. E.; Gavai, A.; Goyal, B.; Han, W. C.; Johnson, W.; Langley, D.; Lee, F. Y.; Marathe, P.; Mathur, A.; Oppenheimer, S.; Ruediger, E.; Tarrant, J.; Tokarski, J. S.; Vite, G. D.; Vyas, D. M.; Wong, H.; Wong, T. W.; Zhang, H. J.; Zhang, G. F. New C-5 substituted pyrrolotriazine dual inhibitors of EGFR and HER2 protein tyrosine kinases. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2036–2042.

(19) Ducray, R.; Ballard, P.; Barlaam, B. C.; Hickinson, M. D.; Kettle, J. G.; Ogilvie, D. J.; Trigwell, C. B. Novel 3-alkoxy-1H-pyrazolo[3,4-d]pyrimidines as EGFR and erbB2 receptor tyrosine kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 959–962.

(20) Mastalerz, H.; Chang, M.; Gavai, A.; Johnson, W.; Langley, D.; Lee, F. Y.; Marathe, P.; Mathur, A.; Oppenheimer, S.; Tarrant, J.; Tokarski, J. S.; Vite, G. D.; Vyas, D. M.; Wong, H.; Wong, T. W.; Zhang, H. J.; Zhang, G. F. Novel C-5 aminomethyl pyrrolotriazine dual inhibitors of EGFR and HER2 protein tyrosine kinases. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2828–2833.

(21) Muraglia, E.; Kinzel, O.; Gardelli, C.; Crescenzi, B.; Donghi, M.; Ferrara, M.; Nizi, E.; Orvieto, F.; Pescatore, G.; Laufer, R.; Gonzalez-Paz, O.; Di Marco, A.; Fiore, F.; Monteagudo, E.; Fonsi, M.; Felock, P. J.; Rowley, M.; Summa, V. Design and synthesis of bicyclic pyrimidinones as potent and orally bioavailable HIV-1 integrase inhibitors. *J. Med. Chem.* **2008**, *51*, 861–874.

(22) Pace, P.; Di Francesco, M. E.; Gardelli, C.; Harper, S.; Muraglia, E.; Nizi, E.; Orvieto, F.; Petrocchi, A.; Poma, M.; Rowley, M.; Scarpelli, R.; Laufer, R.; Paz, O. G.; Monteagudo, E.; Bonelli, F.; Hazuda, D.; Stillmock, K. A.; Summa, V. Dihydroxypyrimidine-4-carboxamides as novel potent and selective HIV integrase inhibitors. *J. Med. Chem.* **2007**, *50*, 2225–2239.

(23) Gardelli, C.; Nizi, E.; Muraglia, E.; Crescenzi, B.; Ferrara, M.; Orvieto, F.; Pace, P.; Pescatore, G.; Poma, M.; Ferreira, M.; Scarpelli, R.; Homnick, C. F.; Ikemoto, N.; Alfieri, A.; Verdirame, M.; Bonelli, F.; Paz, O. G.; Taliani, M.; Monteagudo, E.; Pesci, S.; Laufer, R.; Felock, P.; Stillmock, K. A.; Hazuda, D.; Rowley, M.; Summa, V. Discovery and synthesis of HIV integrase inhibitors: Development of potent and orally bioavailable N-methyl Pyrimidones. *J. Med. Chem.* **2007**, *50*, 4953–4975.

(24) Stewart, J. J. P. Optimization of parameters for semiempirical methods. 1. method. *J. Comput. Chem.* **1989**, *10*, 209–220.

(25) *Dragon for Windows (Software for molecular Descriptor Calculation)*, version 5.4; Talete srl: Milano, Italy, 2006; http://www.talete.mi.it.

(26) Wang, J. M.; Krudy, G.; Xie, X. Q.; Wu, C. D.; Holland, G. Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. *J. Chem. Inf. Model.* **2006**, *46*, 2674–2683.

(27) Tute, M. History and objectives of quantitative drug design in advances in drug research. In *Comprehensive Medicinal Chemistry*; Sammes, P., Taylor, J., Eds.; Pergamon: Oxford, United Kingdom, 1990; Vol. 4, pp 1−32.

(28) Hansch, C.; Sammes, P. G.; Taylor, J. B. *Comprehensive Medicinal Chemistry: The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*,: Pergamon: New York, 1990; Vol. 6. pp 1−19.

(29) Tan, P. K.; Downey, T. J.; Spitznagel, E. L.; Xu, P.; Fu, D.; Dimitrov, D. S.; Lempicki, R. A.; Raaka, B. M.; Cam, M. C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **2003**, *31*, 5676–5684.

(30) Ein-Dor, L.; Kela, I.; Getz, G.; Givol, D.; Domany, E. Outcome signature genes in breast cancer: is there a unique set. *Bioinformatics.* **2005**, *21*, 171–178.

(31) Abeel, T.; Helleputte, T.; Van de Peer, Y.; Dupont, P.; Saeys, Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics.* **2010**, *26*, 392–398.

(32) Shao, L.; Shen, P.; Qu, H. B.; Fan, X. H. Multiple ranking protocol for identifying reliable biomarkers in metabonomics. *Anal. Chim. Acta*; submitted.

CI100305G