JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Recursive Random Forests Enable Better Predictive Performance and Model Interpretation than Variable Selection by LASSO

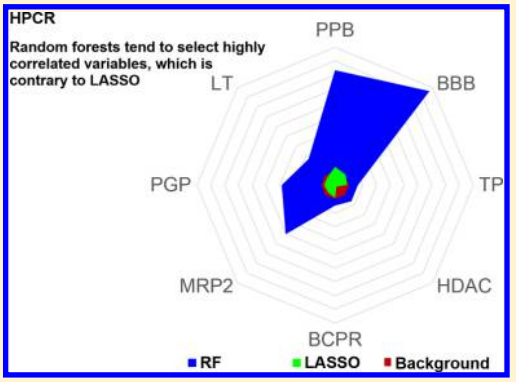
Xiang-Wei Zhu,*[,†,‡] Yan-Jun Xin,[†] and Hui-Lin Ge[§]

[†]College of Resource and Environment and [‡]Qingdao Engineering Research Center for Rural Environment, Qingdao Agricultural University, Qingdao, 266109 Shandong, China

[§]Hainan Provincial Key Laboratory of Quality and Safety for Tropical Fruits and Vegetables, Analysis and Testing Center, Chinese Academy of Tropical Agricultural Sciences, Haikou, 571101 Hainan, China

Supporting Information

ABSTRACT: Variable selection is of crucial significance in QSAR modeling since it increases the model predictive ability and reduces noise. The selection of the right variables is far more complicated than the development of predictive models. In this study, eight continuous and categorical data sets were employed to explore the applicability of two distinct variable selection methods random forests (RF) and least absolute shrinkage and selection operator (LASSO). Variable selection was performed: (1) by using recursive random forests to rule out a quarter of the least important descriptors at each iteration and (2) by using LASSO modeling with 10-fold inner cross-validation to tune its penalty λ for each data set. Along with regular statistical parameters of model performance, we proposed the highest pairwise correlation rate, average pairwise Pearson's correlation coefficient, and Tanimoto coefficient to evaluate the optimal by RF and LASSO in an extensive way. Results showed that variable selection could allow a tremendous reduction of noisy descriptors (at most 96% with RF method in this study) and apparently enhance model's predictive performance as well. Furthermore, random forests showed property of gathering important predictors without restricting their pairwise correlation, which is contrary to LASSO. The mutual exclusion of highly correlated variables in LASSO modeling tends to skip important variables that are highly related to response endpoints and thus undermine the model's predictive performance. The optimal variables selected by RF share low similarity with those by LASSO (e.g., the Tanimoto coefficients were smaller than 0.20 in seven out of eight data sets). We found that the differences between RF and LASSO predictive performances mainly resulted from the variables selected by different strategies rather than the learning algorithms. Our study showed that the right selection of variables is more important than the learning algorithm for modeling. We hope that a standard procedure could be developed based on these proposed statistical metrics to select the truly important variables for model interpretation, as well as for further use to facilitate drug discovery and environmental toxicity assessment.



INTRODUCTION

Quantitative structure activity relationship (QSAR), as it was first proposed by Dr. Hansch,[1] used physicochemical substituents or global parameters to represent chemical structures at the early stage. Variable selection was not so highlighted in view of the relative small set of compounds with simple predictor variables in those studies. The number of molecular descriptors has hugely increased over time, and nowadays thousands of descriptors have been developed to describe different aspects of a molecule ranging from simple bulk properties to sophisticated three-dimensional formulations.[2,3] However, when modeling a particular biological activity, it is reasonable to assume that only a small number of descriptors is actually correlated to the experimental responses. Unneeded predictors may jeopardize model performance and may also lead to worse decisions in model application.[4] The selection of the truly important variables from hundreds of thousands of descriptors is key to the success of model development.[5] New methods such as evolutionary algorithms[6−8] and systematic search[9,10] have been gradually introduced into QSAR fields. Given the fact that many theoretical descriptors are hard to explain explicitly, as well as the incredibly flexible machine learning algorithms which are tolerant to noisy variables, the emphasis on predictive capabilities overwhelmed variable selection as the primary concern in QSAR modeling.

The advent of high-throughput screening (HTS)[11] and toxicogenomic[12] technologies pushed the computational toxicology into the Big Data era.[13] Variable selection shows promise in interpreting the relationship between *in vitro* HTS endpoints and *in vivo* bioactivity profiles[14] and in identifying biomarkers (e.g., sensitive genes) of exposure and/or toxicity using the toxicogenomics data in risk assessment.[12,15]

Received: December 1, 2014
Published: March 6, 2015

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

Furthermore, the so-called hybrid model,[16,17] which incorporates the *in vitro* biological data with chemical structural information in the toxicological models, needs extensive variable selection to screen relevant information for the understanding and predicting chemicals' toxicity. The progress of many projects relied heavily on selecting relevant bioassays to a response, identifying key genes or gene families for toxicity pathway-based risk assessment, and investigating the mechanistic explanation of chemical-biological interaction. Generally, variable selection will play more vital roles in the era of computational toxicology than ever before.

In the Big Data era of chemical toxicity modeling, it is common that the number of predictors (p) is greater than that of response variables (n). One method widely used to deal with this high dimension, low sample size data situation (i.e., $n < p$) is the least absolute shrinkage and selection operator (LASSO).[18] It is a constrained version of ordinary least-squares (OLS) regression (eq 1) and typically used for regression of a single response variable y on a predictor matrix X.

$$\hat{\beta}^{ols} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2\right\} \quad (1)$$

LASSO shrinks the regression coefficients (β_j) toward each other (and toward zero) by imposing a penalty λ on the size of the regression coefficients (eq 2).

$$\hat{\beta}^{lasso} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{m}\left|\beta_j\right|\right\} \quad (2)$$

The absolute size of the regression coefficients β is constrained. The higher the penalty λ, the more regression coefficients are shrunk toward zero. Penalty $\lambda = 0$ gives the OLS solution (eq 1) if $n > p$. The regularized regression is quite sensitive to the selection of the parameter λ. In order to appropriately tune this parameter, the usual approach is to estimate the performance with different λ using cross-validation.[19] Comparing with genetic algorithm or heuristic subset selection, LASSO is fast and accurate with the advantage of avoiding overfitting automatically.

Random forests (RF),[20] an ensemble of decision trees, also received a lot of attention[21] for they can handle large numbers of variables with a relatively small number of observations ($n < p$). In contrast with LASSO, RF does not select variables during the course of model development. However, it provides an assessment of variable importance using the mean decrease in the predictive accuracy as a metric.[22] The predictive accuracy will fluctuate when an individual variable (descriptor) is permuted, while all other variables are left unchanged. For example, the variable importance of predictor variable in a tree is then determined by the difference in predictive accuracy of the out-of-bag observations before and after permuting it in the same tree.[23] Based on variable importance of predictor, a series of recursive variable selection have been proposed in the literature.[24,25] Briefly, the variable importance in recursive RF for each predictor was first calculated. Then a portion of variables, say 20%, with the smallest importance will be eliminated, and a new model will be built using the remaining variables. Finally, the variables leading to the smallest out-of-bag error rate of a RF model will be selected as the optimal variable set.

Basically, both LASSO and random forests are efficient and accurate modeling techniques widely used in computational toxicology. Their predictive performance and distinct variable selection strategies were of particular interest in many published literatures. Studies mainly focused on their predictive performance and the ability to reduce noisy variables. Generally, LASSO and RF outperformed other modeling methods in the prediction accuracy.[26] Both of them could enable better model transparency (the fraction of variables selected by each variable selection method) insofar as the number of variables in the data set can be reduced without a significant deterioration of model predictive performance.[27] Furthermore, studies also showed random forests performed better than LASSO in capturing nonlinearity of the bioprocess data.[19] Meanwhile, several very recent studies indicated the feature of mutual exclusion of highly correlated pairwise predictors in LASSO variable selection may jeopardize model's predictive performance and interpretation.[28−30] Admittedly, LASSO is not designed for a data set with collinear variables and with response variables that are not linear. The major problem is that the more the number of variables, the more chances of collinearity. Given the near infinite number of theoretical structural descriptors or the *in vitro* HTS and *in vivo* toxicogenomic profiles, collinearity is the problem we will encounter and needs to be coped with in most QSAR modeling. Another issue is that we could not know whether the response variables are linear or not since the structure−activity optimization surface is not smooth for the existence of outliers such as activity cliff or even measurement errors.[31]

Consequently, all the questions can be reduced to one: among the variable subsets selected by different variable selection strategies, which one is reliable and can then be used for model interpretation and how to define the reliability. So many statistical metrics had been proposed to evaluate model performance; however, the metrics to evaluate whether a selected variable subset is optimal are rarely studied. Thus, our aims are to use new statistical metrics to compare the variable selection of LASSO and RF in combination with data sets with diverse biological endpoints. We put emphasis on the quantitative characterization of the ultimate optimal variable sets selected by LASSO and RF.

MATERIALS AND METHODS

Data Sets. To ensure the reliability and generalization of this study, we used diverse data sets with continuous and categorical endpoints (see Table 1 for details). Four data sets contain the continuous biological endpoints: (1) plasma protein binding (PPB);[32] (2) blood brain barrier (BBB),[33] (3) toxicity on *tetrahymena pyriformis* (TP);[34] and (4) human histone deacetylase (HDAC) inhibition.[35] Another four data sets with binary responses (active and inactive) include the following: (1) breast cancer resistance protein (BCPR);[36] (2) multidrug-resistance associated protein 2 (MRP2);[36] (3) P-glycoprotein (PGP) substrate;[37] and 4) liver toxicity (LT).[38] All that chemical information can be found in Supplementary Table S1. Previous studies already developed sound predictive models based on these eight data sets. Here we remodeled those data sets with different descriptors and learning methods. Our purpose was not to compare our modeling results with previous ones. Since the selection of descriptors affects the modeling results,[39] four different types of descriptors were

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

Table 1. Summary of Eight Data Sets Used for Performance Comparison[a]

data set	n	m	descriptor	property	active/inactive or $y_{min} \sim y_{max}$ (mean)
PPB[32]	1242	887	Dragon	protein binding (%)	0.5−99.5% (64.4%)
BBB[33]	159	253	Dragon	logBB	−2.15−1.64 (−0.05)
TP[34]	983	116	MOE	pIGC50	−2.67−3.34 (0.17)
HDAC[35]	59	88	MOE	pIC50	4−8.46 (6.13)
BCPR[36]	382	335	Mold2	substrate/nonsubstrate	167/215
MRP2[36]	96	273	Mold2	substrate/nonsubstrate	48/48
PGP[37]	195	114	CDK	substrate/nonsubstrate	108/86
LT[38]	292	120	CDK	toxic/nontoxic	156/136

[a]The number of chemicals (n) and descriptors (m) for eight data sets were listed with corresponding descriptor types and response endpoints. The number of active or inactive compounds were show for four categorical data sets. The minimum, maximum, and mean values for continuous responses were shown for another four data sets.

calculated for those eight data sets (Table 1): Chemistry Development Kit (CDK, v.1.4.13, GNU Lesser General Public License), Dragon software (v.5.5, Talete SRL, Milan, Italy), Molecular Operating Environment (MOE, v.2009.10, Chemical Computing Group, Montreal, Canada), and Mold2 (NCTR, US FDA) (Table 1). The biological endpoints and descriptors were paired up with no special arrangement as long as the final modeling data sets include both the $n < p$ and $n > p$ cases. After range scaling (from 0 to 1), we removed descriptors with low variance (standard deviation smaller than 0.001) and high redundancy (if pairwise R^2 larger than 0.95, one of the pairs was randomly removed).

Modeling Methods. To ensure the reliability and robustness of model performance, the standard 5-fold cross-validation procedure[40] was used for all the model development in this work. Briefly, the whole data set was randomly split into five parts. One part was kept for model assessment, and four parts were used in the model development. Consequently, an integrated model consisting of all five models will be used for external prediction as a whole.

Recursive Random Forests Modeling. Random forests were run through the package "randomForest" on the R language platform (v3.1.0). Referring to a previous study,[41] we set the number of descriptors randomly sampled for splitting at each node during tree induction (m_{try}) as one-third of the total number of descriptors for regression models and square root of the total number of descriptors for classification models. The default number of trees (n_{tree}) of 500 is used for all data sets. The mean decrease in the predictive accuracy[22] was used to rank the variable importance. According to the variable importance, we keep 3/4 of the most important descriptors and rebuild the models. Recursively removing one-fourth of the descriptors with the least importance until there are at least seven left. We call this variable selection procedure as recursive random forests (RRF). Our purpose here is to select optimal variable sets from a series of them to maximize the model's predictive power. Generated descriptors at each recursive step were recorded for further research.

LASSO Modeling. LASSO was run through the package "glmnet" on the R language platform (v3.1.0). Penalty λ (eq 2) is a key parameter to be optimized in the LASSO modeling. Besides the 5-fold cross-validation, an inner layer of cross-validation was designed to ensure the optimization of parameter λ. According to the 5-fold cross-validation, a whole data set was first split into five parts. Based on four parts of the whole data set, an inner loop of an additional 10-fold cross-validation was performed to optimize parameter λ and select an optimal one that gives minimum cross-validation error. When the optimal λ (i.e., optimal variables included in the LASSO model) was obtained, LASSO modeling will be performed based on the four parts used in the inner loop. The remaining one part will be used to validate the predictive performance of LASSO models. The whole procedure was then repeated for five times in the outer 5-fold cross-validation loop. Totally, five models with associated optimal penalty λ will be built for one data set.

SVM and GLM Modeling. The support vector machine (SVM)[42] and general linear modeling (GLM)[43] approaches were also used to build models using variables selected by RF and LASSO. SVM finds in the descriptor-activity space the narrowest band containing most of the data points.[44] Comparing with RF, it is like a black box machine learning methods without inherent variables selection function. Here we used SVM and GLM to build models for both the continuous and categorical data sets based on the variable sets selected by RF and LASSO. The SVM and GLM modeling mentioned in the following were performed using the "e1071" package and the built-in GLM function, respectively, on the R language platform (v3.1.0).

Evaluation Statistics. Besides the regular modeling metrics for the regression (e.g., coefficient of determination (R^2) and mean absolute error (MAE) and classification models (e.g., specificity, sensitivity, and correct classification rate (CCR)), to facilitate the comparison of optimal descriptor sets selected by RF and LASSO, we defined the following metrics:

Suppose a subset of s optimal descriptors were selected from a total number of p descriptors, the transparency was defined as s/p.[27] Transparency represents the ability of a variable selection algorithm to extract the key ones from a pool of variables with noisy information. Usually, transparency was calculated for the variable set which maximizes the predictive performance of a model.

Highest Pairwise Correlation Rate (HPCR). Highly correlated descriptors might be generated by similar algorithm or represent similar properties. For a total number of p descriptors in one data set, each descriptor could find the one with the highest correlation and there should be p pair of them. We defined the HPCR as the p pair of highest correlated descriptors divided by the total number of p descriptors and the value will be 1 (p/p) for the whole data set. For a subset of s optimal descriptors selected from a total number of p descriptors, the highest pairwise correlation counterpart of each descriptor in the subset may or may not fall within the subset. Suppose there are the k pair of highest correlated descriptors in the subset, the HPCR value would be k/s. If one randomly selects s variables from p of them, that obtained value would be the background HPCR. To determine whether an algorithm tends to collect or exclude highly correlated variables, the real HPCR was divided by the background HPCR to render the level up ratio (eq 3). The accurate background HPCR

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

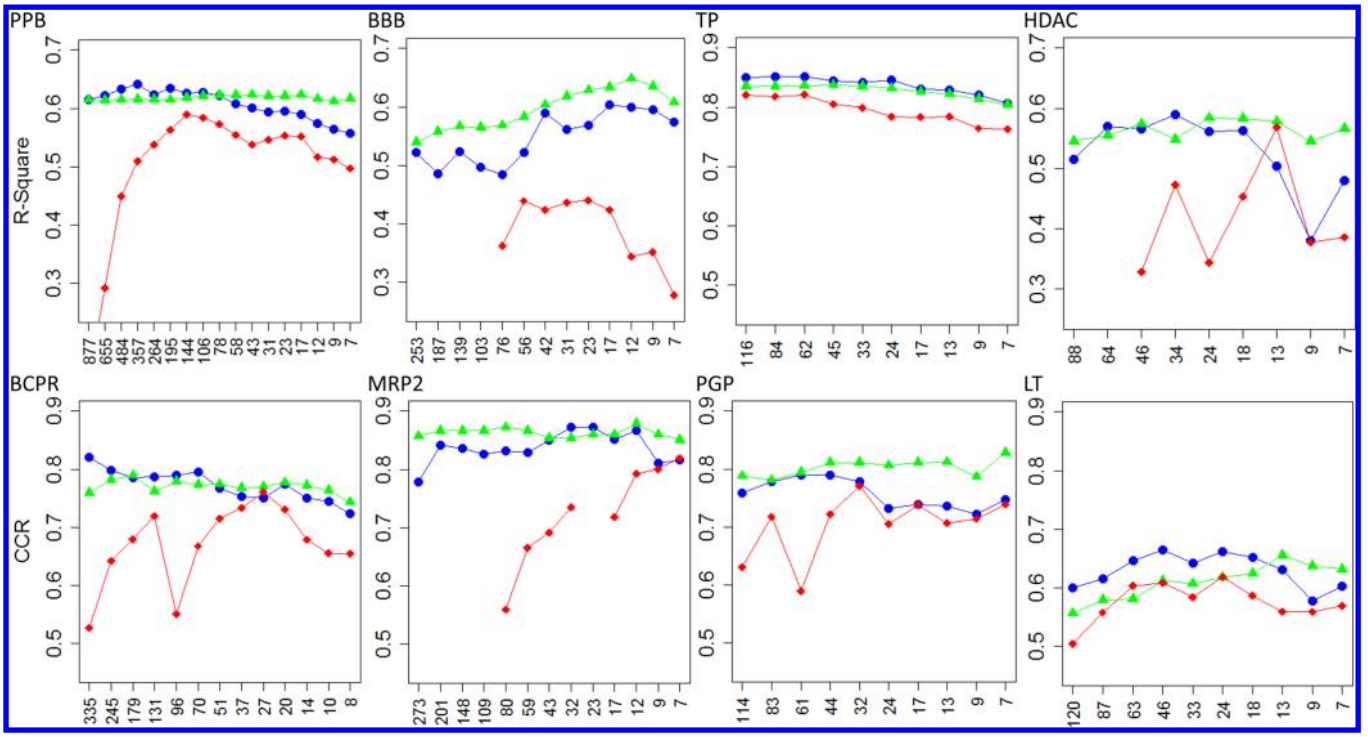


Figure 1. Model performance of three learning algorithms (RF, SVM, and GLM) based on the variable sets selected by recursive random forests. The determination of coefficient (R^2) and correct classification rate (CCR) were used as metrics to indicate model performance for regression and classification models, respectively. Green triangle: RF; blue dot: SVM; and red diamond: GLM.

needs to be calculated many times and be averaged through randomly sample s variables from p of them with replacement

$$\text{level up ratio} = \frac{k/s}{\sum_N \frac{k_b}{s_b}/N} = \frac{kN}{s\sum_N \frac{k_b}{s_b}} \quad (3)$$

where subscript b means the randomly sampled subset, and each data set was sampled for 10,000 times (N = 10,000) to render an accurate background HPCR.

Average Pairwise Pearson's Correlation Coefficient (APR). HPCR is an indicator of the ability of an algorithm to select highly correlated variables. APR is used to show the overall view of the correlated variables in the selected subset. The background APR can be calculated based on the total number of p descriptors in one data set.

Tanimoto coefficient (TC, eq 4) is used to evaluate the similarity between two descriptor sets selected by RF and LASSO separately. Suppose S_1 and S_2 are two subset variables selected by RF and LASSO, respectively. The similarity between S_1 and S_2 can be expressed as follows

$$\text{TC} = \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|} \quad (4)$$

where $|S_1|$ and $|S_2|$ and $|S_1 \cap S_2|$ mean the number of descriptors in subset S_1, S_2, and intersection of S_1 and S_2, respectively.

RESULTS

In this work, we compared the variable selection of random forests and LASSO on four continuous and four categorical data sets. A recursive variable selection procedure was adapted to random forests. For each data set, a series of variable sets selected by recursive RF modeling were used as input for GLM and SVM modeling to select the best performance models.

Recursive Modeling. Recursive RF modeling was performed for all eight data sets. One fourth of the least important descriptors were removed at each step until there were seven left. The model performance and selected variables at each step were recorded. Depending on the total number of descriptor in one data set, the descriptor thinning steps range from 9 for the human histone deacetylase (HDAC) to 15 for the plasma protein binding (PPB) data set. Meanwhile, SVM and GLM modeling were performed using descriptor sets selected by RRF for each data set. For data sets with more descriptors than compounds (Table 1), general linear modeling will break down for the high dimension, low sample size data. Thus, the GLM modeling skipped $n < p$ cases for the first few variable sets until $n > p$ as the descriptor thinning (e.g., BBB, HDAC, and MRP2).

Figure 1 shows model performance of RF, SVM, and GLM using variables selected at each step of RRF for all eight data sets. Each symbol in Figure 1 (i.e., green triangle, blue dot, or red diamond) represents the performance of one integrated model based on a 5-fold cross-validation procedure. Totally, there are 274 integrated models in Figure 1. Clearly, the predictive performance of the RF model was enhanced as the removing of a number of unimportant descriptors. Its performance remains quite stable even though the descriptor thinning process exceeded the optimal point that produces maximized predictive performance. Employing the descriptor sets selected by RF, the predictive performance of SVM was also improved with the removing of noisy variables. However, the performance of the SVM model declined tremendously as the descriptor thinning process came to the last few steps (Figure 1). SVM showed superior predictive power over RF for six out of eight data sets (see Supplementary Tables S2 and S3). GLM was more fragile to noisy variables than the other two machine learning methods (Figure 1). Its predictive

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

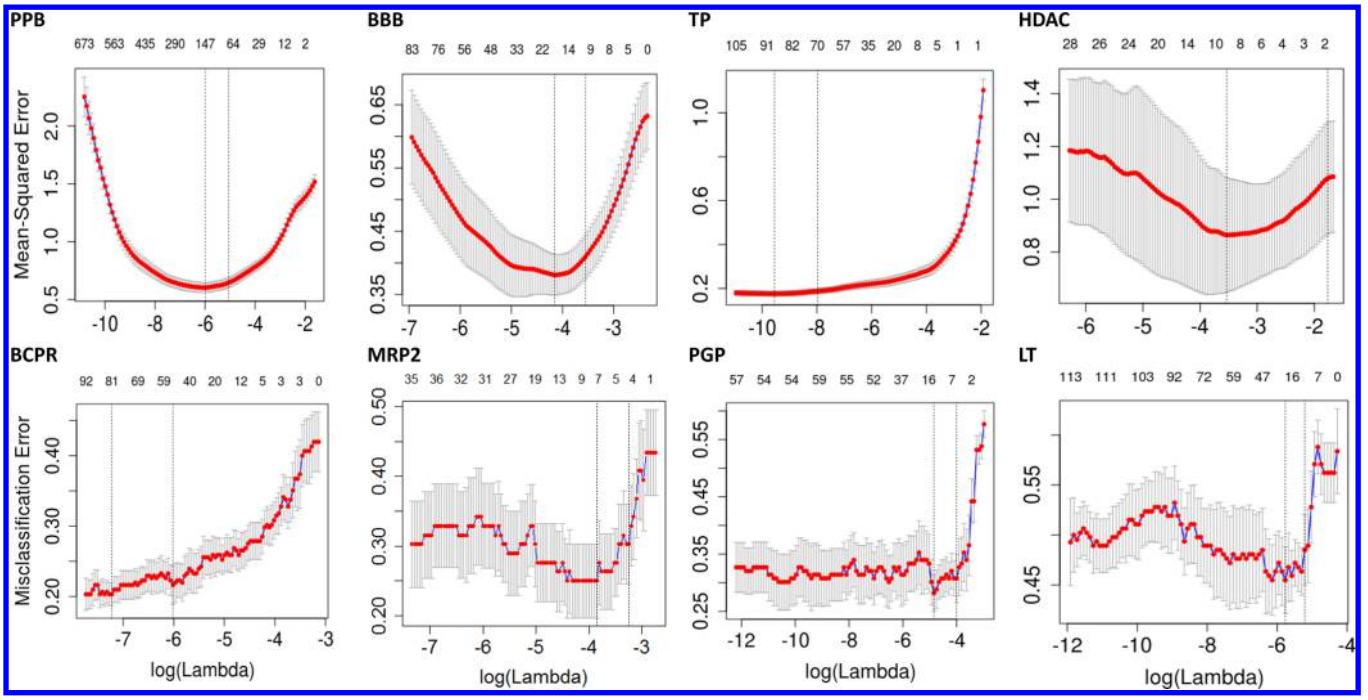


Figure 2. Mean cross-validated errors (cvm) (red dots) with standard errors (error bar) can be used to determine the optimum penalty lambda (λ). The vertical dotted lines at the left of each plot indicates the optimum λ and corresponding descriptor number. The vertical dotted line at the right of each plot indicates the largest value of λ such that error is within one standard error of the minimum.

Table 2. Parameter λ and Associated Mean Cross-Validated Error (CVM) of the Five LASSO Models (M1, M2, M3, M4, and M5) under the Context of 5-Fold Cross-Validation for All Eight Data Sets

	model									
	M1		M2		M3		M4		M5	
data	λ	CVM[a]	λ	CVM	λ	CVM	λ	CVM	λ	CVM
PPB	2.51×10^{-03}	0.60	2.13×10^{-03}	0.61	1.93×10^{-03}	0.63	1.51×10^{-03}	0.64	1.73×10^{-03}	0.61
BBB	1.57×10^{-02}	0.38	5.10×10^{-03}	0.27	1.80×10^{-02}	0.41	4.66×10^{-03}	0.35	9.76×10^{-03}	0.37
TP	7.10×10^{-05}	0.17	1.13×10^{-04}	0.19	1.15×10^{-04}	0.18	1.11×10^{-04}	0.20	7.06×10^{-05}	0.18
HDAC	1.88×10^{-03}	1.17	2.05×10^{-03}	0.63	2.47×10^{-03}	0.80	2.04×10^{-03}	0.84	2.08×10^{-03}	0.68
BCPR	7.27×10^{-04}	0.20	1.49×10^{-03}	0.20	1.44×10^{-03}	0.20	1.75×10^{-03}	0.21	2.18×10^{-03}	0.21
MRP2	2.12×10^{-02}	0.25	2.21×10^{-02}	0.19	1.50×10^{-02}	0.17	4.57×10^{-03}	0.18	1.43×10^{-03}	0.19
PGP	7.88×10^{-03}	0.28	1.46×10^{-04}	0.28	5.29×10^{-03}	0.30	4.00×10^{-03}	0.21	4.41×10^{-04}	0.23
LT	3.12×10^{-03}	0.45	1.13×10^{-03}	0.39	1.41×10^{-03}	0.41	1.02×10^{-03}	0.35	1.01×10^{-03}	0.37

[a]: CVM would be squared-error for regression models (PPB, BBB, TP, and HDAC) and misclassification error for categorical models (BCPR, MRP2, PGP, and LT).

performance was enhanced tremendously for most data sets after recursively removing noisy variables but remained unstable. The instability of predictive performance of GLM was also conspicuous for data sets such as PPB, HDAC, BCPR, and PGP (Figure 1). For example, GLM model prediction for the PPB data set could be enhanced and reach the highest point (R^2 = 0.59) as the number of descriptor reduces to 144 from 877. Then the performance would decline steadily as the number of descriptor reduces from 144 to 7. For the BCPR data set, the correct classification rate reached 71.9% as the descriptor number reduced from 335 to 131. However, the CCR dropped to 55.0% after one more round of noisy elimination with 96 descriptors left. Then the CCR gradually increased as the descriptor thinning and reached the highest peak of 76.0% with descriptor number of 37. The instability was much sharper for the MRP2 data set. The GLM broke down when we tried to build a linear classifier using 23 descriptors (see MRP2 in Figure 1). Generally, the best GLM models performed worse than those of SVM and RF.

LASSO Modeling. Based on the whole descriptors, LASSO modeling was performed to select the best subset of important descriptors through the tuning of parameter λ (eq 2) and hence to develop predictive models. The optimum lambda (λ) was optimized by the inner 10-fold cross-validation. The mean square errors and misclassification errors of inner cross-validation were used as objective metrics to optimize λ for continuous and categorical models, respectively. Totally, five LASSO models were developed as the results of 5-fold outer cross-validation for one data set. A plot of λ versus the mean cross-validation errors (CVM) for each data set is shown in Figure 2. Generally large data sets (e.g., PPB with 1242 compounds) tend to show less variation in CVM than that of small data sets (e.g., HDAC with 59 compounds) for continuous models. The optimal λ and corresponding

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

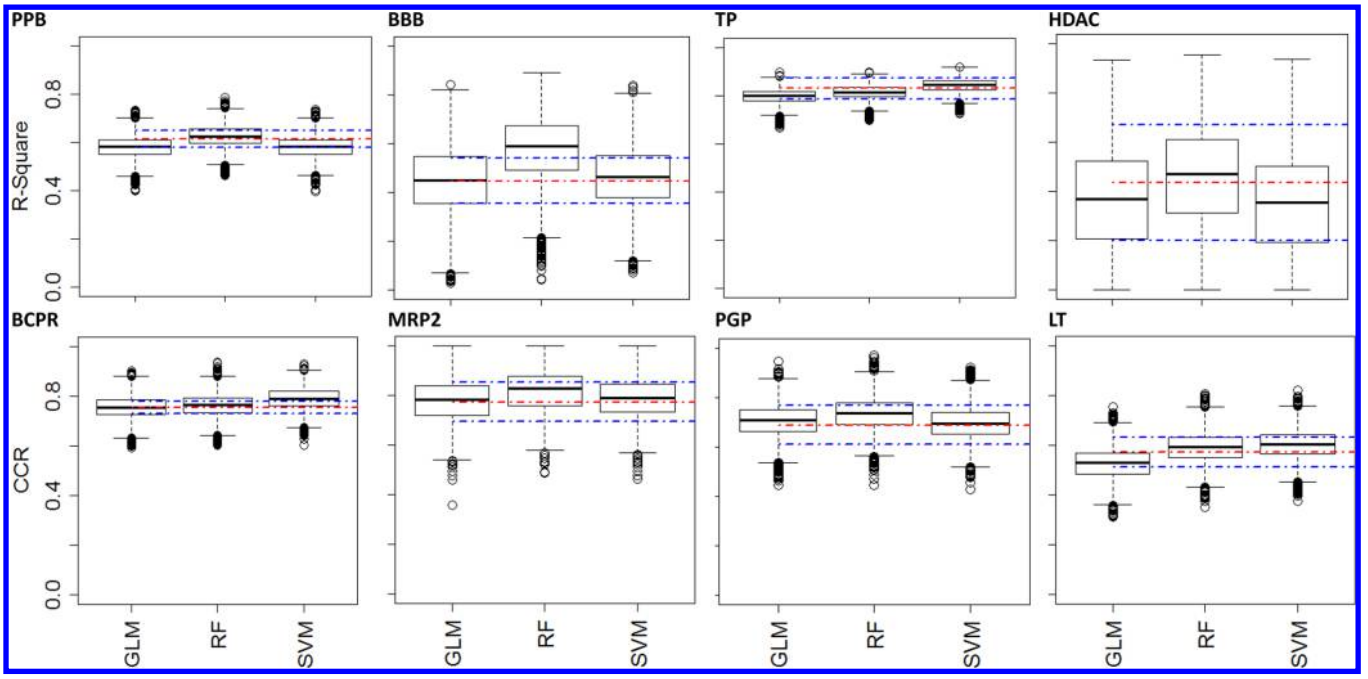


Figure 3. Predictive performance (mean ± standard error based on 20%-off bootstrapping, n = 10,000) of the GLM, RF, and SVM modeling based on descriptor groups selected by LASSO. The red dashed line and two blue dashed lines show the LASSO performance and the standard error based on 20%-off bootstrapping (n = 10,000), respectively.

descriptor number for each data set can be determined from Figure 2.

Table 2 listed the optimal parameter λ and associated mean cross-validated errors (CVM) for each of the five folds of the eight data sets. For a single data set, the parameter λ showed fluctuation among the five folds (e.g., the range of λ was at least 1 order of magnitude for MRP2 and PGP in Table 2), while their associated minimal CVM just showed slight differences. Even though the number of optimal descriptors varies slightly among each of the folds, most of them shared vast inconsistencies (see Supplementary Table S4). Take the HDAC data set as an example, five variable subsets were selected under the context of 5-fold cross-validation. The number of variables in the five subsets are 28, 31, 29, 29, and 28. The number of variables in the intersection set (common variables shared by all five subsets) is 4. The number of variables in the union set (unique variables of all five subsets) is 40. It should be pointed out that all five parts in the cross-validation were split randomly with stratified sampling to guarantee the representativeness of the subsets. Note that if the one data set is homogeneous, variable sets selected by linear LASSO models based on different parts of the data set should share high consistency. Thus, Supplementary Table S4 showed us the degree of diversity of data sets like BBB, HDAC, MRP2, PGP, and LT.

Comparison of Predictive Performance. To show the impact of variable selection strategy on the model performance, we compared LASSO models with those of RF, SVM, and GLM based on the following: (1) their associated optimal variable subsets selected by RRF and (2) variable sets selected by LASSO.

A detailed comparison of the best performance LASSO, RF, SVM, and GLM models for all eight data sets was illustrated in Figure S1. Generally, machine learning models (RF and SVM) outperformed GLM-RF (represents the GLM model based on the descriptor set selected by RF and so forth) and LASSO models for all eight data sets. For the linear learning methods, the performance of LASSO was on par with that of GLM-RF for most of these data sets (e.g., PPB, BBB, TP, BCPR, and MRP2 in Figure S1).

For the purpose of investigating the quality of LASSO variables, we performed RF, SVM, and GLM modeling based on the variables selected by LASSO. As we can see in Figure 3, the advantages of machine learning methods disappeared comparing with LASSO. The performance of RF-LASSO was only apparently superior to LASSO for one data set (BBB). SVM-LASSO was even worse than LASSO for PPB and HDAC data sets. Apparently, the predictive performance of RF-LASSO and SVM-LASSO was inferior to corresponding optimal RF and SVM models, respectively. Supplementary Tables S5 and S6 listed the performance of RF-LASSO models for all eight data sets.

Referring to the total number of descriptors in Table 1, penalty λ successfully shrunk most of the regression coefficient to zero even though it was quite small in value (Table 2). As mentioned before, the regression coefficients in LASSO modeling (eq 2) would also shrink to OLS if penalty $\lambda = 0$ (eq 1). Thus, we see the opportunity to compare the two linear modeling methods (GLM and LASSO) using the same LASSO variables to evaluate how λ could affect their model performances. LASSO just showed slightly better than GLM-LASSO for PPB, TP, HDAC, and LT data sets and showed no improvement for the other four data sets (Figure 3). This would indicate that the impact of λ was mainly on the variable selection rather than the model performance. Overall, the performance (R^2 and CCR for continuous and categorical data sets, respectively) of GLM-RF are 0.589, 0.440, 0.821, 0.569, 0.760, 0.819, 0.771, and 0.617 for PPB, BBB, TP, HDAC, BCPR, MRP2, PGP, and LT, respectively, while the performance of GLM-LASSO for those data sets are 0.579, 0.428, 0.799, 0.378, 0.751, 0.766, 0.697, 0.521. The difference in

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

Table 3. The Number of Optimal Descriptors and Corresponding Transparency in Combination of 32 Best Models (8 Data Sets × 4 Modeling Methods)[a]

model	PPB	BBB	TP	HDAC	BCPR	MRP2	PGP	LT
RF	43(0.05)	***12(0.05)***	45(0.39)	24(0.27)	179(0.53)	12(0.04)	***7(0.06)***	13(0.11)
SVM-RF	***357(0.40)***	17(0.07)	***84(0.72)***	***34(0.39)***	***335(1.00)***	***23(0.08)***	44(0.39)	***46(0.38)***
LASSO	183(0.21)	29(0.11)	87(0.75)	18(0.20)	65(0.19)	14(0.05)	33(0.29)	44(0.37)
GLM-RF	144(0.16)	23(0.09)	62(0.53)	13(0.15)	27(0.08)	7(0.03)	32(0.28)	24(0.20)

[a]For one specific data set, the italic bold font style represents variable number and associated transparency with the best performance among the four modeling methods.

Table 4. The Level Up Ratio of the Highest and Second Highest Pairwise Correlated Variables (H and SH) and Their Corresponding Averages for the Optimal RF and LASSO Descriptors

model		PPB	BBB	TP	HDAC	BCPR	MRP2	PGP	LT
RF	H	10.41	11.63	1.68	1.70	1.45	6.10	2.69	0.00
	SH	6.25	7.67	1.63	1.62	1.51	4.00	5.00	5.40
	average	8.33	9.65	1.66	1.66	1.48	5.05	3.85	2.70
LASSO	H	1.33	1.27	0.88	0.00	0.47	0.00	0.86	0.83
	SH	1.33	0.91	0.88	0.30	1.37	1.46	0.75	0.75
	average	1.33	1.09	0.88	0.15	0.92	0.73	0.80	0.79

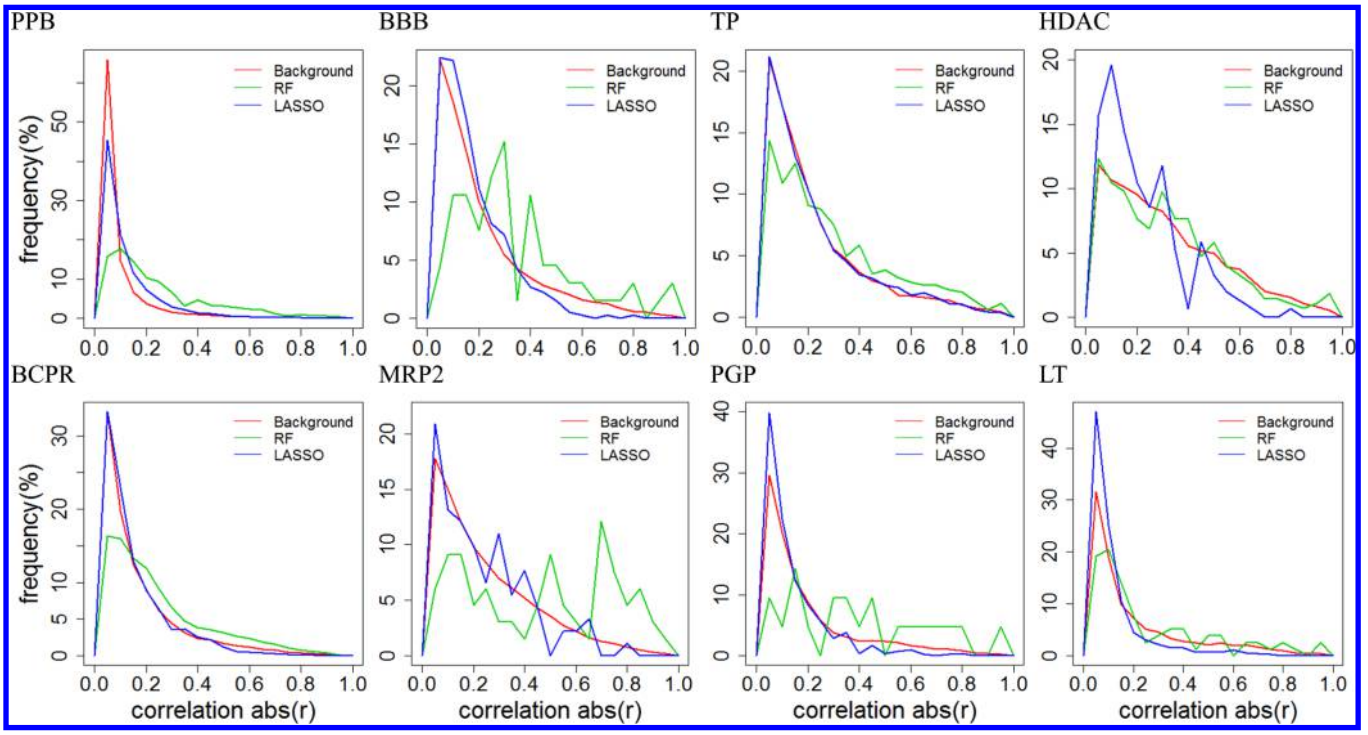


Figure 4. Probability distribution of the absolute pairwise correlation coefficients (Pearson) of the whole data set and optimal variables selected by RF and LASSO.

performance can attribute into the difference in the selected variables.

Model Transparency. Totally, there were 32 combinations (8 data sets × 4 modeling methods) of optimal models selected from 282 models (i.e., 274 models represented as different symbols in Figure 1 plus eight LASSO models). Transparency was calculated for the variable set which maximizes the predictive performance of one combination. Note the inconsistent number of optimal descriptors among the five LASSO models under the context of 5-fold cross-validation for the same data set (Supplementary Table S4). We averaged these numbers to calculate the transparency. For example, the number of optimal variables of five LASSO models for PPB are 147, 177, 170, 198, and 191 (Supplementary Table S4). Transparency was calculated using their average (177) divided by the total number of descriptors (877). Table 3 showed the number of optimal descriptors and corresponding transparencies for all 32 combinations.

Previous studies show that RF and LASSO can drastically reduce the number of variables in the data set without a significant deterioration of model prediction performance.[27] Our study showed that every combination model can be improved with a tremendous reduction of the number of descriptors. For example, the transparency of optimal RF models for the eight data sets ranges from 0.04 (MRP2) to 0.53 (BCPR), which means that as much as 96% of variables could

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

Table 5. Pearson's Correlation Coefficient (Mean ± Standard Error) for the Total Descriptors, RF Selected Descriptors, and LASSO Selected Descriptors of All Eight Data Sets

name	PPB	BBB	TP	HDAC	BCPR	MRP2	PGP	LT
background	0.07 ± 0.12	0.19 ± 0.19	0.21 ± 0.20	0.29 ± 0.22	0.15 ± 0.17	0.23 ± 0.19	0.18 ± 0.19	0.19 ± 0.21
RF	0.22 ± 0.20	0.32 ± 0.22	0.28 ± 0.23	0.30 ± 0.22	0.22 ± 0.19	0.44 ± 0.28	0.38 ± 0.27	0.24 ± 0.24
LASSO	0.10 ± 0.12	0.15 ± 0.12	0.21 ± 0.20	0.19 ± 0.15	0.13 ± 0.14	0.21 ± 0.17	0.11 ± 0.12	0.099 ± 0.13

be ruled out. Overall, the transparency of RF models was better than that of LASSO models which ranges from 0.07 (MRP2) to 0.75 (TP) (Table 3). Basically, models with less variables are preferred for the sake of simplicity and interpretation. The GLM-RF and SVM-RF also showed a great reduction in the number of descriptors and achieved better performance for most of those data sets (Table 3).

The level up ratio for highly correlated variables is a metric to indicate the selection of highly correlated variables. The highest and second highest pairwise correlation rate (HPCR1 and HPCR2) of the optimal variables sets selected by RF and LASSO for each data set were shown in Supplementary Table S7. The background HPCR1 and HPCR2 based on bootstrapping (n = 10,000) for RF and LASSO were listed in Supplementary Table S8. The level up ratio (eq 3) was calculated based on Supplementary Tables S7 and S8. The level up ratio of 1.0 means the modeling method shows no bias in the selection of strongly or weakly correlated variables. The level up ratio greater than 1.0 means the positive bias in selecting strongly correlated variables and vice versa. Table 4 showed the level up ratio for HPCR1, HPCR2, and their corresponding averages for RF and LASSO. Clearly, RF shows a stronger ability to select highly correlated variables than LASSO ranging from at least 1.61 times for BCPR to as high as 11.07 times for HDAC. The ability of LASSO to select highly correlated variables for most data sets (six out of eight) is negative (Table 4).

Furthermore, the level up ratio can be demonstrated through exploring the pairwise Pearson's correlation coefficient in the optimal variable sets. Probability distribution of the absolute pairwise correlation coefficients (Pearson) of the whole data set and the optimal variables selected by RF and LASSO is shown in Figure 4. The correlation coefficients of RF variables showed lower frequency than that of background and LASSO variables at the left part of axis of correlation coefficients. However, at the right part of axis of correlation coefficients the correlation coefficients of RF variables showed much higher frequency than that of background and LASSO variables. That was especially conspicuous for BBB, MRP2, and PGP data sets.

Table 5 showed the averaged pairwise Pearson correlation coefficient for the whole data set and optimal variables selected by RF and LASSO for all eight data sets. On the one hand, the APRs of RF variables for most data sets were much higher than these of the total descriptors which were referred to as background (Table 5). On the other hand, the APRs of LASSO descriptors were much lower than associated background. Overall, Table 4 and Table 5 are a two-step verification of the distinct differences in the strategies of variable selection for RF and LASSO. As illustrated early in this paper, high correlated variables in models are useful in detecting chemical or biological profiles. It would be of vital importance in identifying relevant and correlated bioassays to optimizing in vitro assay panels for the toxicity screening of drugs and drug candidates.

The similarity between the RF and LASSO variables, represented using Tanimoto coefficient (eq 4), is lower than 0.20 for seven out of eight data sets besides the TP of 0.54 (Figure 5). MRP2 is the data set with the lowest TC value

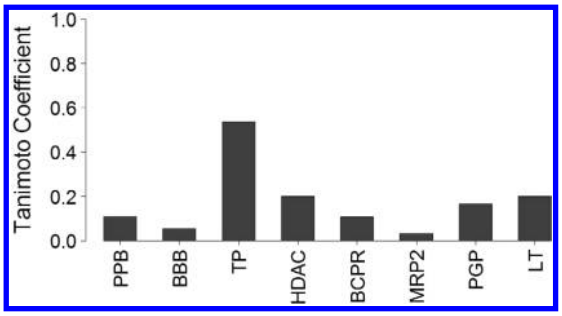


Figure 5. Tanimoto coefficient (similarity) between the LASSO variables and the optimal RF variables for all eight data sets.

(0.03) between the RF and LASSO variable sets with the variable number of 12 and 14, respectively, which means that only one descriptor is shared in common between those two sets. Even the optimal variable sets selected by RF and LASSO cover a wide spectrum of chemical structural space, and they shared very few common variables (e.g., the number of optimal variables selected by RF and LASSO for the BCPR data set are 179 and 65, respectively. However, they only shared 13 descriptors in common with a TC of 0.11.).

DISCUSSION

Machine learning methods like Random forests were more efficient in capturing the nonlinearity of data compared with LASSO.[19] This can explain why the RF and SVM outperformed the linear modeling methods (i.e., LASSO and GLM) (Supplementary Tables S2 and S3). However, the performance of RF-LASSO and SVM-LASSO were not necessarily better than the LASSO model with the same descriptors (Figure 3). Those results indicated that the impact of the learning algorithm was actually quiet limited on the enhancement of model performance. A previous study showed that chemical descriptors were more important than learning algorithms for modeling.[39] This study indicated that the proper variable selection strategies are also more important than learning algorithms for modeling.

LASSO tends to arbitrarily select only one variable from a group of highly correlated ones.[28,29] The mutual exclusion of highly correlated variables during variable selection undermined the performance (Supplementary Figure S1 and Tables S2 and S3) as well as the interpretation of the LASSO model. On the one hand, RRF outperformed LASSO for all eight data sets. The statistical change that the superior performance of the RRF approach simply by chance is 7.77×10^{-05} according to the Fisher's exact probability test. On the other hand, we used two data sets (HDAC and BBB) as examples to show the difference in model explanation of RF and LASSO.

For the HDAC data set, the number of optimal variables selected by RF and LASSO were 18 and 24, respectively (Table

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

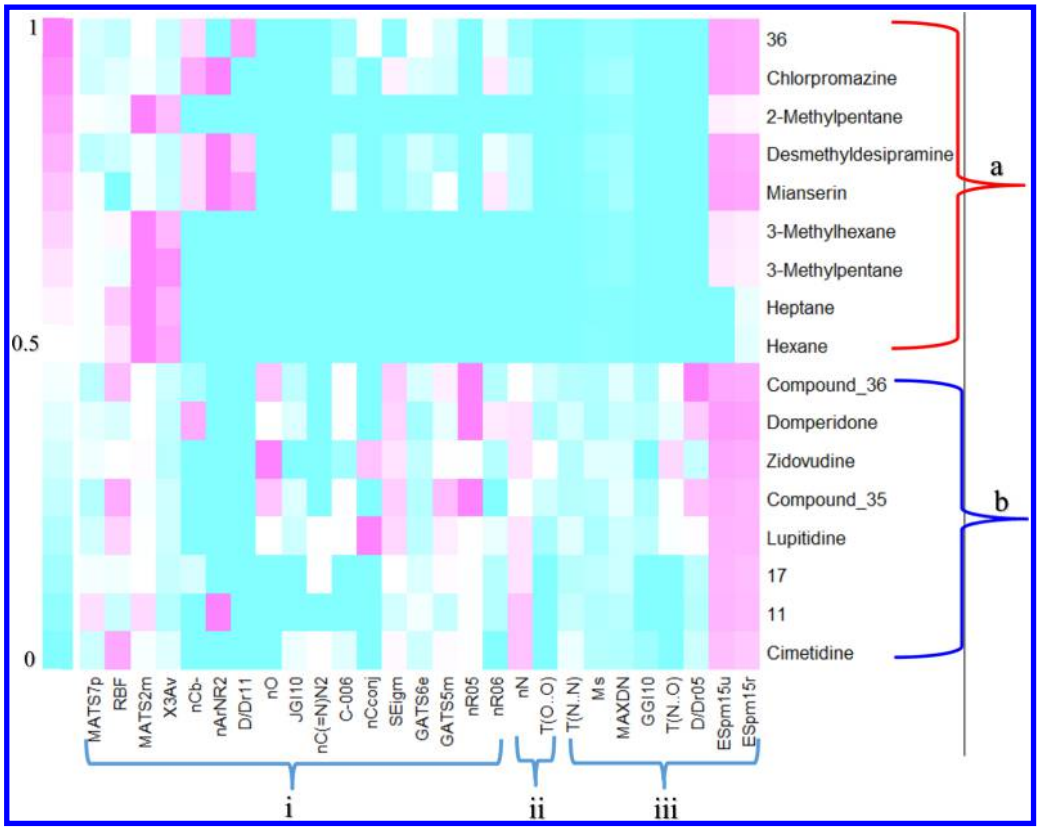


Figure 6. Descriptor profiles of nine high blood brain barrier (BBB) permeability compounds (a) and eight low BBB permeability compounds (b). The structure and BBB permeability of those 17 compounds can be found in the Supplementary Table S9. The meaning of descriptors selected by LASSO (i), recursive random forests (iii), and both of them (ii) was shown in Supplementary Table S10. Ten LASSO descriptors and two RF descriptors that are constant or near constant for those 17 compounds were not shown in the heatmap.

3), and they share 6 variables in common. The logarithm of the octanol/water partition coefficient (SlogP) was found to be of key importance to inhibit the human histone deacetylase by RRF. Four other descriptors showing high coefficient correlation (Pearson's $R > 0.85$) with SlogP were also selected as key variables. These descriptors are logP(o/w): another type of the logarithm of the octanol/water partition coefficient; logS: the logarithm of aqueous solubility; SMR_VSA5: subdivided surface areas are descriptors based on an approximate accessible van der Waals surface area; and vdw_vol: van der Waals volume. Based on this group of highly correlated descriptors, a conclusion can be easily drawn that chemicals' solubility or polarity play key roles in the inhibition of human histone deacetylase. This conclusion is consistent with the fact that six confirmed screening hits of HDAC inhibitors contain at least one hydrophilic group (−C(═O)−NH−OH or −NH−C(═O)−C−OH).[35] Of the 24 selected variables by RRF, over 46% corresponding HPCR1 and 42% HPCR2 fall within the same variable group (Supplementary Figure S1 and Table S7). The average HPCR1 and HPCR2 was 0.15 (Table 4) for LASSO variables which is the lowest in the eight data sets. We can see that LASSO was extremely strict on the mutual exclusion of highly correlated variables for the HDAC data set. Although other descriptors (SMR_VSA6, BCUT_SLOGP_3, and SlogP_VSA0) were slightly related to polarity, their pairwise correlation efficient were all less than 0.50. A solid interpretation like that of the RF model can hardly be reached for the LASSO model.

For the BBB data set, the number of optimal variables selected by RF and LASSO were 29 and 12, respectively (Table 3), and they share 2 variables in common. The average HPCR1 and HPCR2 was 9.65 for RF variables which is the highest in the eight data sets (Table 4). We can see that RF was extremely efficient in gathering highly correlated variables for the BBB data set. Figure 6 showed the descriptor profiles of nine compounds showed high blood brain barrier permeability and eight compounds showed low BBB permeability (their structures can be found in Supplementary Table S9). A clear pattern of RF variables in Figure 6 can be found to differentiate those high and low BBB permeability compounds based on most of the RF selected variables (e.g., T(N···N), Ms, MAXDN, D/Dr05, and T(N···O)). However, the patterns of most LASSO variables were ambiguous between the high and low BBB permeability compounds. For the two variables shared by RF and LASSO, only nN (number of nitrogen atoms, see Supplementary Table S10) can distinguish compounds with different BBB permeability. Based on the chemical structures in Supplementary Table S9, we can see clearly the difference between the two groups of chemicals with high and low permeability. Compounds with high permeability tend to be more hydrophobic and contain less nitrogen (nitroaromatic, amide, imido, and nitro groups) than those with low permeability. Descriptors like T(N···N) (sum of topological distances between N···N) and T(N···O) (sum of topological distances between N···O) were another two descriptors selected by RF except the nN; they ranked the most and sixth most important variables in RF models. nC(═N)N2 (number of guanidine derivatives) and nArNR2 (number of tertiary amines (aromatic)) were another two descriptors selected by LASSO related to nitrogen atoms. However, they

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

just showed a mediocre ability to distinguish compounds with different permeability (Figure 6).

Model interpretation is a state-of-the art work since there are too many limitations such as chance correlation and the lack of interpretability of many descriptors.[31] The REACH regulation also placed no mandatory requirement on the mechanistic interpretation of models.[45] A previous study warned chance correlation between the response variables and predictor variables for small data sets.[46] For a data size of 59, as in the HDAC set, the likelihood of a random correlation is as high as 10.7% for absolute values of r larger than 0.30. This also holds true for the random correlation of two predictor variables. As shown in Figure 4 for HDAC and other sets, most of the absolute pairwise r for LASSO variables distributed in a narrow range of 0−0.30, while the distribution of absolute pairwise r for RF variables showed different views. For a specific data set, the distribution curve would follow an exponential decay from low to high absolute pairwise r.[46] LASSO enhanced the possibility to select random variables with low absolute pairwise r into the models, while RF would reduce the risk of selected random variables since it tends to selected highly correlated ones.

The predictive accuracy of random forests will be undermined when there are many irrelevant variables and the number of predictors exceeds the number of observations.[25] Recursive random forests modeling is a commonly used strategy to overcome that problem.[47,48] In this work, one-fourth of the least important variables were removed from the next round of modeling. Refined models wound be developed if the removal ratio of the least important variables were smaller than 25%. Consistent with previous work,[49] LASSO modeling in this study performed well in coping with multicollinearity while affording the risks of missing important variables. However, its predictive accuracy will be compromised and unstable in the existence of highly correlated predictors.[30] The objectives of variable selection are to identify (1) important variables highly related to the response variable for interpretation purposes and (2) a small number of variables sufficient for a good prediction of the response variable.[24] From this point of view, LASSO is not recommended to be the only modeling method for a new data set.

As shown in this study, the optimal variables selected by RF and LASSO showed low similarities for most of the cases. If more learning algorithms with variable selection (e.g., k nearest neighbors, genetic algorithm, and all subset models) were used for modeling, several models with different optimal variable sets will be developed. We have a number of metrics to evaluate the model performances and use the consensus prediction to achieve better accuracy. However, we still lack the standard procedure to evaluate different optimal variable sets. In this study, we proposed HPCR, APR, and Tanimoto coefficient to quantify the selected variables from different aspects. We hope those parameters could be used or be refined in the following studies to deal with variables selected by different methods. A standard procedure could be developed to select the true important variable for model interpretation and for further use to facilitate drug discovery and environmental toxicity assessment.

CONCLUSION

In this study, we applied four modeling methods (RF, SVM, LASSO, and GLM) in the data mining of eight data sets. Diverse response endpoints (continuous and categorical) and four types of descriptors were employed to guarantee the generality of our conclusion. The results demonstrated the following: (1) Recursive random forests as well as LASSO could allow a tremendous reduction of noisy descriptors, as high as 96% in this study, and enhance a model's predictive performance as well. (2) The mutual exclusion of highly correlated variables of LASSO tends to skip highly related variables that are important to predict response endpoints, which is completely contrary to random forests. These differences in variable selection resulted in the interior predictive performance of LASSO models and even RF, SVM, and GLM models based on variables selected by LASSO. (3) In parallel with the standard procedures to deal with predictions from multiple models, a new standard procedure to deal with the multiple variable sets selected by different algorithms are also desperately needed. Statistical metrics like model transparency, highest pairwise correlation rate, average pairwise person's correlation coefficient, and Tanimoto coefficient performed well in the quantification of optimal variables selected by RF and LASSO. Overall, recursive random forests outperforms LASSO in both identifying primary controlling variables and developing predictive models in this study. Our study shows the right choice of variables is more important than the learning algorithm for modeling. Our study improves the understanding of the difference in the variable selection between RF and LASSO in a quantitative way. The statistical metrics proposed in this study are hoped to be further used in the selection of key important features for drug discovery and environmental toxicity assessment.

ASSOCIATED CONTENT

Supporting Information

Tables S1−S10 and Figure S1. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: 86 0532-88030461. E-mail: xwzhunc@gmail.com. Corresponding author address: Room 424, Informatics Building, Qingdao Agriculture University, Qingdao, 266109 China.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financed by the National Natural Science Foundation of China (No. 21407087) and the Startup Foundation for Advanced Talents (No. 6631113336) of Qingdao Agricultural University.

REFERENCES

(1) Hansch, C.; Steward, A. R. The Use of Substituent Constants in the Analysis of the Structure-Activity Relationship in Penicillin Derivatives. *J. Med. Chem.* **1964**, *7*, 691−694.

(2) González, M. P.; Terán, C.; Saíz-Urra, L.; Teijeira, M. Variable Selection Methods in QSAR: An Overview. *Curr. Top. Med. Chem.* **2008**, *8*, 1606−1627.

(3) Xue, L.; Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363−372.

(4) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2003**, *44*, 1−12.

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746

(5) Kiralj, R.; Ferreira, M. M. C. Is Your QSAR/QSPR Descriptor Real or Trash? *J. Chemom.* **2010**, *24*, 681−693.
(6) Kubinyi, H. Variable Selection in QSAR Studies.II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.
(7) Kubiny, H.; Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.
(8) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.
(9) Baumann, K.; Albert, H.; von Korff, M. A Systematic Evaluation of the Benefits and Hazards of Variable Selection in Latent Variable Regression. Part I. Search Algorithm, Theory and Simulations. *J. Chemom.* **2002**, *16*, 339−350.
(10) Zheng, W. F.; Tropsha, A. Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the K-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(11) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol. Sci.* **2007**, *95*, 5−12.
(12) Boverhof, D. R.; Zacharewski, T. R. Toxicogenomics in Risk Assessment: Applications and Needs. *Toxicol. Sci.* **2006**, *89*, 352−360.
(13) Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants. *Chem. Res. Toxicol.* **2014**, *27*, 1643−1651.
(14) Bisgin, H.; Chen, M.; Wang, Y.; Kelly, R.; Fang, H.; Xu, X.; Tong, W. A Systems Approach for Analysis of High Content Screening Assay Data with Topic Modeling. *BMC Bioinf.* **2013**, *14*, S11.
(15) Davis, A. P.; Murphy, C. G.; Johnson, R.; Lay, J. M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B. L.; Rosenstein, M. C.; Wiegers, T. C.; Mattingly, C. J. The Comparative Toxicogenomics Database: Update 2013. *Nucleic Acids Res.* **2013**, *41*, D1104−D1114.
(16) Zhu, H. From QSAR to QSIIR: Searching for Enhanced Computational Toxicology Models. In *Methods in molecular biology*; Springer: Clifton, NJ, 2013; Vol. *930*, pp 53−65.
(17) Tropsha, A. Potential of Short-Term Biological Assays to Quantitatively Predict Chronic Toxicity. *Toxicol. Lett.* **2013**, *221* (Suppl), S52−S53.
(18) Tibshirani, R. The Lasso Method for Variable Selection in the Cox Model. *Stat. Med.* **1997**, *16*, 385−395.
(19) Hassan, S.; Farhan, M.; Mangayil, R.; Huttunen, H.; Aho, T. Bioprocess Data Mining Using Regularized Regression and Random Forests. *BMC Syst. Biol.* **2013**, *7* (Suppl 1), S5.
(20) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(21) Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable Selection Using Random Forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225−2236.
(22) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.
(23) Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinf.* **2008**, *9*, 307.
(24) Diaz-Uriarte, R.; de Andres, S. Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinf.* **2006**, *7*, 3.
(25) Chang, Y. Variable Selection via Regression Trees in the Presence of Irrelevant Variables. *Commun. Stat. Comput.* **2013**, *42*, 1703−1726.
(26) Neves, H. H. R.; Carvalheiro, R.; Queiroz, S. A. A Comparison of Statistical Methods for Genomic Selection in a Mice Population. *BMC Genet.* **2012**, 13.
(27) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Benchmarking Variable Selection in QSAR. *Mol. Inf.* **2012**, *31*, 173−179.
(28) Bondell, H. D.; Reich, B. J. Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics* **2008**, *64*, 115−123.
(29) Lu, F.; Petkova, E. A Comparative Study of Variable Selection Methods in the Context of Developing Psychiatric Screening Instruments. *Stat. Med.* **2014**, *33*, 401−421.
(30) Savin, I. A Comparative Study of the Lasso-Type and Heuristic Model Selection Methods. *J. Econ. Stat. (Jahrbuecher Natl. Stat.)* **2013**, *233*, 526−549.
(31) Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2007**, *48*, 25−26.
(32) Zhu, X.-W.; Sedykh, A.; Zhu, H.; Liu, S.-S.; Tropsha, A. The Use of Pseudo-Equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding. *Pharm. Res.* **2013**, *30*, 1790−1798.
(33) Zhang, L. Y.; Zhu, H.; Oprea, T. I.; Golbraikh, A.; Tropsha, A. QSAR Modeling of the Blood-Brain Barrier Permeability for Diverse Organic Compounds. *Pharm. Res.* **2008**, *25*, 1902−1914.
(34) 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.
(35) Tang, H.; Wang, X. S.; Huang, X. P.; Roth, B. L.; Butler, K. V.; Kozikowski, A. P.; Jung, M.; Tropsha, A. Novel Inhibitors of Human Histone Deacetylase (HDAC) Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation. *J. Chem. Inf. Model.* **2009**, *49*, 461−476.
(36) Sedykh, A.; Fourches, D.; Duan, J.; Hucke, O.; Garneau, M.; Zhu, H.; Bonneau, P.; Tropsha, A. Human Intestinal Transporter Database: QSAR Modeling and Virtual Profiling of Drug Uptake, Efflux and Interactions. *Pharm. Res.* **2012**, *30*, 996−1007.
(37) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A Computational Ensemble Pharmacophore Model for Identifying Substrates of P-Glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.
(38) Zhu, X.-W.; Sedykh, A.; Liu, S.-S. Hybrid in Silico Models for Drug-Induced Liver Injury Using Chemical Descriptors and in Vitro Cell-Imaging Information. *J. Appl. Toxicol.* **2014**, *34*, 281−288.
(39) Young, S. S.; Yuan, F.; Zhu, M. Chemical Descriptors Are More Important than Learning Algorithms for Modelling. *Mol. Inf.* **2012**, *31*, 707−710.
(40) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.
(41) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(42) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.
(43) Friedman, J. H.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1−22.
(44) Yuan, Z.; Mattick, J. S.; Teasdale, R. D. SVMtm: Support Vector Machines to Predict Transmembrane Segments. *J. Comput. Chem.* **2004**, *25*, 632−636.
(45) OECD. *Guidance Document on the Validation of (quantitative) structure-Activity Relationships [(q)sar] Models*; Paris, 2007.
(46) Hutter, M. C. Determining the Degree of Randomness of Descriptors in Linear Regression Equations with Respect to the Data Size. *J. Chem. Inf. Model.* **2011**, *51*, 3099−3104.
(47) Granitto, P. M.; Furlanello, C.; Biasioli, F.; Gasperi, F. Recursive Feature Elimination with Random Forest for PTR-MS Analysis of Agroindustrial Products. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 83−90.
(48) Wu, X.; Wu, Z.; Li, K. Identification of Differential Gene Expression for Microarray Data Using Recursive Random Forest. *Chin. Med. J. (Engl).* **2008**, *121*, 2492−2496.
(49) Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *Am. Stat.* **2009**, *63*, 308−319.

DOI: 10.1021/ci500715e
J. Chem. Inf. Model. 2015, 55, 736−746