JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

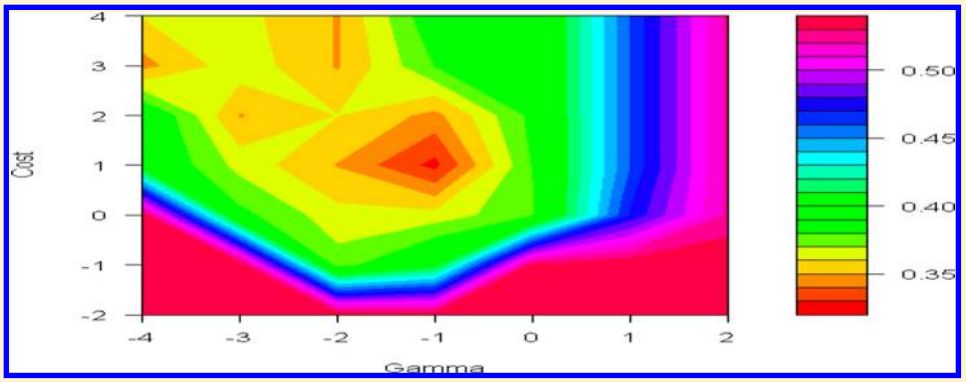
Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure–Activity Relationship and Machine Learning Methods

Qingda Zang,[†] Daniel M. Rotroff,[‡,§] and Richard S. Judson*[,‡]

[†]ORISE Postdoctoral Fellow and [‡]National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

[§]Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, United States

Supporting Information



ABSTRACT: There are thousands of environmental chemicals subject to regulatory decisions for endocrine disrupting potential. The ToxCast and Tox21 programs have tested ~8200 chemicals in a broad screening panel of in vitro high-throughput screening (HTS) assays for estrogen receptor (ER) agonist and antagonist activity. The present work uses this large data set to develop in silico quantitative structure–activity relationship (QSAR) models using machine learning (ML) methods and a novel approach to manage the imbalanced data distribution. Training compounds from the ToxCast project were categorized as active or inactive (binding or nonbinding) classes based on a composite ER Interaction Score derived from a collection of 13 ER in vitro assays. A total of 1537 chemicals from ToxCast were used to derive and optimize the binary classification models while 5073 additional chemicals from the Tox21 project, evaluated in 2 of the 13 in vitro assays, were used to externally validate the model performance. In order to handle the imbalanced distribution of active and inactive chemicals, we developed a cluster-selection strategy to minimize information loss and increase predictive performance and compared this strategy to three currently popular techniques: cost-sensitive learning, oversampling of the minority class, and undersampling of the majority class. QSAR classification models were built to relate the molecular structures of chemicals to their ER activities using linear discriminant analysis (LDA), classification and regression trees (CART), and support vector machines (SVM) with 51 molecular descriptors from QikProp and 4328 bits of structural fingerprints as explanatory variables. A random forest (RF) feature selection method was employed to extract the structural features most relevant to the ER activity. The best model was obtained using SVM in combination with a subset of descriptors identified from a large set via the RF algorithm, which recognized the active and inactive compounds at the accuracies of 76.1% and 82.8% with a total accuracy of 81.6% on the internal test set and 70.8% on the external test set. These results demonstrate that a combination of high-quality experimental data and ML methods can lead to robust models that achieve excellent predictive accuracy, which are potentially useful for facilitating the virtual screening of chemicals for environmental risk assessment.

INTRODUCTION

Humans and ecological species are exposed to tens of thousands of man-made chemicals in the environment. Some environmental chemicals can mimic natural hormones and disrupt normal functions of the endocrine system.[1−5] The U.S. Environmental Protection Agency (EPA) has identified ~30 000 chemicals as having the potential for human exposure, and the information about these chemicals is compiled and stored in the ACToR database.[6,7] These chemicals cover a wide variety of use classes, including consumer products, food additives, and human and veterinary drugs. Most current protocols for testing

Received: September 9, 2013
Published: November 26, 2013

of chemical compounds for their biological activity and toxicity are time-consuming and expensive. While it is feasible to accurately evaluate a small number of compounds using experimental in vivo methods, it becomes prohibitive in terms of both cost and time when dealing with tens of thousands of chemicals. Due to high resource costs, only a small fraction of these chemicals have been thoroughly tested and well assessed for their potential risks to both human health and the environment.[8−11] Thus, most environmental chemicals have little to no data regarding their toxicological potential, in particular, their ability to interact with the estrogen receptor (ER).[12,13]

To facilitate safety evaluations of the huge number of untested chemicals, combined with approximately 1500 new chemical substances submitted to EPA each year,[14,15] it is imperative that chemical producers and regulatory agencies have access to rapid and inexpensive approaches. High-throughput screening (HTS) assays have become an emerging and viable tool for chemical prioritization.[16−18] The U.S. Tox21 consortium has launched a screening campaign with the purpose of testing thousands of compounds in HTS in vitro assays for targets including the ER, and EPA's ToxCast project is currently screening thousands of chemicals for a wide array of molecular targets.[19−21] In particular, ToxCast/Tox21 have run ∼1800 chemicals in a battery of 18 in vitro ER assays and ∼8200 chemicals (including the 1800) in a subset of these assays.

The increasing availability of empirical data from in vitro methods for large number of chemicals enables the development of an in silico toolbox based on predictive models built from data mining and machine learning methods.[22,23] Quantitative structure−activity relationships (QSARs) can be statistically derived on the basis of structural features and physicochemical properties of chemicals.[24−28] QSAR models assume that structurally similar molecules exhibit similar activities that can be predicted by using a collection of structural and molecular features. QSAR analysis is one of the most widely used methods to link the structure of a chemical to its biological and chemical activity and is used to classify molecules as active or inactive against specific targets.[29−31] Molecular descriptors and structural fingerprints have been extensively applied in substructure/similarity searching, clustering, classification, and other statistical learning studies of toxicological and pharmacological properties of compounds.[32−38] Major classes of molecular descriptors include connectivity and shape of molecules, electro-topological state, quantum chemical and geometrical properties. There are thousands of theoretical molecular descriptors available in the literature, most of which can be calculated by available software packages. Since not all of the molecular descriptors are related to the activity of chemicals, it is important to choose appropriate ones for model building. Feature selection (FS) methods are able to optimally select a subset of molecular features that reduce noise and eliminate redundant descriptors which are not relevant to the activity prediction.[38−40] FS enables the development of more accurate and efficient computational tools and can help elucidate the relationship between the structure of a chemical and its biological activity.

A common problem in ML model building occurs when the training HTS data are highly imbalanced with only a small number of active chemicals compared to the number of inactive chemicals. The Tox21 data set, used in the present work, contains 8303 compounds, comprised of 383 active and 7920 inactive compounds. Machine learning based on such extremely skewed data results in failure of traditional classification algorithms because they aim to minimize the overall error rate, rather than pay special attention to the rare class. Most classifiers are designed to maximize the prediction accuracy, so they usually focus on the majority class to a greater extent than the minority class, resulting in poor performance for predicting the active compounds. Many methods have been proposed for dealing with the imbalanced classification issue, and they can be grouped into three categories: cost-sensitive learning, oversampling the minority class, and undersampling the majority class.[41−45] In the current research, we present a new strategy to appropriately tackle the problem of class imbalance using what we term a "target-independent" clustering method. For many targets such as ER, the activity of chemicals requires the presence of specific structural groups, and structural neighbors of an active compound are also at increased likelihood of being active. We observed that if chemicals are clustered based on structural similarity using a large unselected set of molecular descriptors, independently of any information on target activity, most clusters contain no ER active molecules, while a few clusters are highly enriched in actives. Given this observation, the basic approach is to cluster all chemicals, retain inactive chemicals that are located in the clusters containing one or more test-positive training set chemicals, and discard other chemicals in pure inactive-clusters, thus increasing the ratio of active to inactive compounds for the ML QSAR development stage.

The main goal of the present work is to apply ML approaches, such as linear discriminant analysis (LDA), classification and regression tree (CART), and support vector machines (SVM) as fast and low-cost tools for binary classification of active and inactive compounds based on HTS data from ToxCast and Tox21 for the specific molecular target, viz., ER, using the target-independent enrichment method.

MATERIALS AND METHODS

Data Sources. The present study was conducted using data from ToxCast and Tox21 chemical libraries, which consist of 1814 and 8303 unique chemicals, respectively, after removing the replicates and mixtures. Chemicals were selected for testing in the ToxCast and Tox21 programs (and hence for testing in ER assays) to represent broad use classes to which humans are exposed. These include pesticide active and inert ingredients; industrial chemicals such as solvents, surfactants, and plastics; cosmetics and personal care ingredients; food additives; and pharmaceuticals. Hence they cover very broad regions of chemical structure and property space, which makes for a challenging prediction problem. One particular issue is that, with this diverse chemical library, there are some chemicals with few close structural neighbors, making them hard to predict accurately. All high throughput screening (HTS) data are managed through the ToxCast database, one of several federated databases making up the ACToR system—a collection of publicly accessible data sets on environmental chemicals for data domains, containing such information as chemical identities and structures, hazard testing, exposure and occurrence data, chemical use categories, and regulatory status, etc. (http://actor.epa.gov). The initial information on these chemicals includes their names, CAS registry numbers (CASRN), simplified molecular input line entry systems (SMILES), and INCHI keys as a primary identifier.

In order to determine the estrogenic activity of the ToxCast chemicals, an aggregation of in vitro assay results was used to develop a composite ER Interaction Score for indicating the likelihood of a chemical interacting with ER. This Interaction

Score was derived from a collection of 13 in vitro assays that measured binding, receptor dimerization, gene transcription, and ER-dependent cell growth. The score is essentially a weight-of-evidence (WOE) approach to handling cases where not all assays agree. Details are given elsewhere.[46] Among the 1814 chemicals with Interaction Scores, 317 (17.5%) chemicals had ER Interaction Scores > 0. In this study, compounds with Interaction Scores > 0% were regarded as active, while those with Interaction Scores = 0% were regarded as inactive. This set of 1814 chemicals form the training and internal test sets for the ML QSAR model. The remainder of the 8303 Tox21 chemicals (tested in 2 of the 13 assays, both for ER agonist behavior) form the external test set.

Molecular Descriptors and Structural Fingerprints. The molecular structures of the 8303 Tox21 chemicals were represented in Daylight SMILES format, and their SMILES strings were converted into three-dimensional (3D) structures using the molecular modeling software Molecular Operating Environment (MOE, Chemical Computing Group, version 2012.10).[47] These compound molecules were processed using the MOE Wash function as follows: compounds with multiple components, such as mixtures and salts, were discarded; compounds complexed with metal ions and structurally ambiguous compounds were eliminated; all counterions and solvent molecules were removed; and explicit hydrogen atoms were added. All molecules were considered in their neutral form. Then the 3D structures were generated and their geometries were optimized using the MMF94 force field. The molecular descriptors were calculated using the QikProp software (Schrödinger version 3.2).[48] Of the 8303 chemicals, QikProp failed on 1693. For the remaining 6610 chemicals, a total of 51 descriptors were generated for each molecule as given in the Supporting Information (Table S1).

All chemicals with available structures were fingerprinted using publicly available SMARTS sets FP3, FP4, MACCS from OpenBabel,[49] PADEL,[50] and PubChem.[51] A total of 4328 bits of structural fingerprints were generated. For fingerprint systems, chemical structures are represented by binary strings, where each bit indicates the presence or absence of certain structural and/or physicochemical feature, such as a specific element, a particular structural fragment, or atom environment in the molecule.

Chemical Clustering. Using the above structure-feature fingerprints, a Tanimoto similarity matrix was calculated and chemicals were clustered using Euclidean distance and Ward's method. A total of ~4000 features showed up in at least one structure Tanimoto similarity for the entire collection of the 8126 unique Tox21 structures. Chemical clusters at different levels of the hierarchy were then employed as part of the chemical classification.

Data Sets. Of the 6610 chemicals with QikProp molecular descriptors, 1537 chemicals were from the ToxCast data set, consisting of 264 active and 1273 inactive compounds, respectively, on the basis of Interaction Score cutoff. These chemicals were randomly partitioned into two subsets: a training set and a test set (also called an internal test set as compared to the external test set). The random division was performed for each individual class separately, so that two-thirds of each class was in the training set (176 active and 849 inactive compounds, data set I) while one-third of each class in the test set (88 active and 424 inactive compounds, data set II).

The remainder of the Tox21 data served as an external test set with 171 active and 4902 inactive compounds (data set III) after excluding the chemicals from ToxCast. For the Tox21 data, we treated a compound as active if both of the following two assays were active. Otherwise it was inactive. The two assays were designated Tox21_ERa_BLA_Agonist_ratio (beta lactamase reporter gene assay run in Bla cells) and Tox21_ERa_LUC_BG1_Agonist (luciferase reporter gene assay run in BG1 cells).[52]

Table 1 summarizes all of the data sets, where the training set was used to build and optimize the binary classification models and the test sets were used to validate the models. Data were organized as a matrix, which is a rectangular array with chemicals as rows and molecular descriptors/structural fingerprints as columns. ER activity classification data were appended as an outcome column to each of the data sets, and the chemicals were labeled active or inactive based on the respective ER HTS assay data. Each chemical was represented as a 4379 dimensional feature vector with 1−51 dimensions as numeric molecular descriptors and 52−4379 as binary bits of fingerprints.

Table 1. Data Sets Used for Classification Study

data set	total chemicals	active chemicals	inactive chemicals	active/inactive
Tox21	6610	435	6175	1:14.2
ToxCast	1537	264	1273	1:4.82
training set (I)	1025	176	849	1:4.82
internal test set (II)	512	88	424	1:4.82
external test set (III)	5073	171	4902	1:28.7

When building a QSAR classification model, data diversity should always be addressed. For the QikProp descriptors, the chemical property space coverage was investigated, and the distribution of nine properties, i.e., molecular weight (mol_MW), solvent accessible surface area (SASA), solvent accessible volume (volume), predicted polarizability (QPpolrz), hexadecane/gas partition coefficient (QPlogPC16), octanol/gas partition coefficient (QPlogPoct), octanol/water partition coefficient (QPlogPo/w), predicted aqueous solubility (QPlogS), and prediction of binding to human serum albumin (QPlogKhsa) for data sets I−III was plotted in Figure 1. As shown in this figure, while data sets I and II present very similar distributions with respect to most of these properties, i.e., the internal test set shares a similar chemical space with the training set, data set III is more sparsely scattered than data sets I and II because it contains a significantly higher number of compounds and hence more diverse structures. In particular, the portion of the Tox21 chemical library that does not overlap with ToxCast contains a large number of pharmaceutical active ingredients.

Feature Selection. Random forest (RF), a popular machine learning approach, can effectively discover molecular information for distinguishing activities and properties of various compounds and assess the importance of descriptors to the model.[53−55] In RF classification, an ensemble of classification trees is grown from separate bootstrap samples of the training data using the decision tree algorithm, which is described later in this section. The bootstrap sampling is performed through a random selection with replacement from the chemicals in the training data during tree growth. The chemicals that are not employed for tree growth are called out-of-bag (OOB) samples. Each tree provides a prediction for its OOB chemicals, and the average of these results over all trees gives an overall OOB validation.

Descriptor importance measures the degree of association between a given descriptor and the prediction results of a classification model, and hence, descriptors with great

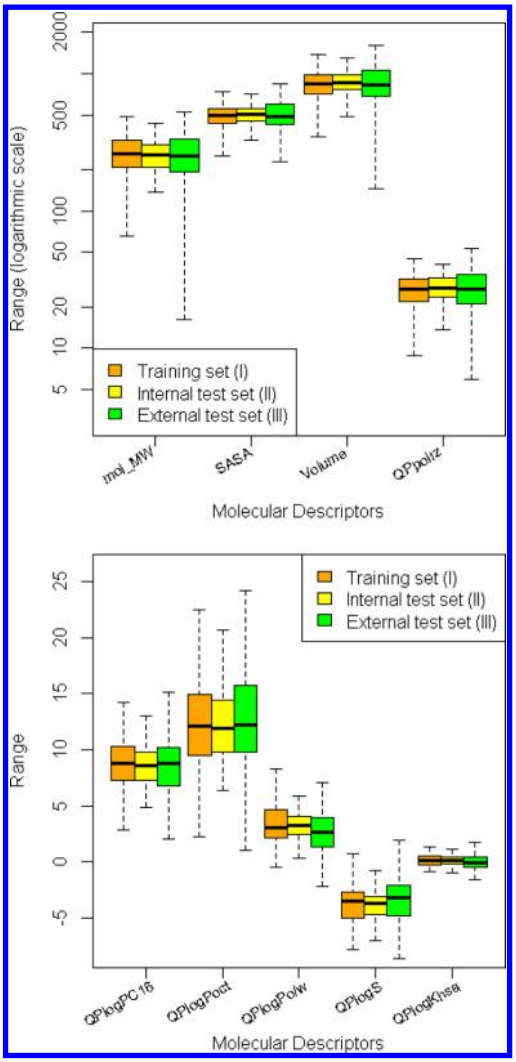


Figure 1. Data distribution of molecular descriptors for training and test sets. The box plots show the minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and maximum of each property.

importance have strong association with the prediction performance. The measure of descriptor importance from RF is based on the deterioration of classification performance after random permutation of these descriptors. To estimate the importance for a specific descriptor, the OOB chemical samples are passed down the tree and the prediction accuracy is calculated. The calculation is implemented tree by tree as the RF model is constructed, and the decrease in prediction accuracy of these permutations is then averaged over all the trees. When all the descriptors are successively permuted in the OOB chemical samples using random noise, the relative change of the prediction error is compared to that found for the original descriptor set and used to measure the importance of the descriptors. A substantial decrease in the prediction accuracy suggests that the descriptor has strong association with biological activity, and the most discriminative descriptors are the most important ones. After the importance of all the descriptors has been estimated, RF will return a ranked list of the descriptor importance. The descriptors can be sorted based on their relative importance and plotted against the importance value or weight score, and thus, an importance spectrum similar in appearance to a scree plot is obtained, which is useful for determining the number of relevant descriptors. It is worth noting that the exact importance order of descriptors may vary in different runs due to intercorrelations between descriptors, and therefore, the subset of descriptors selected by RF may not be the unique solution, but it is the one with the smallest OOB error rate.

Linear Discriminant Analysis. As a widely used linear classification technique, linear discriminant analysis (LDA) is a parametric method based on the assumption that the data in each class are normally distributed. Here, the aim of LDA is to qualitatively predict the class affiliation for unknown chemicals by establishing a linear discriminant function so that (1) between-class variance is maximized; (2) within-class variance is minimized; and (3) the optimal separation between the given classes is achieved. LDA provides a classification model by determining the linear combination of the molecular descriptors or bits of fingerprints that best predicts the category or group to which a given compound belongs. Classification of a chemical is performed by calculating the Mahalanobis distance of the chemical from the gravity center of each class.[56,57] The Mahalanobis distance between a chemical (x_i) and the data center ($\bar{x}$) is defined as

$$D(x_i) = [(x_i - \bar{x})^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}(x_i - \bar{x})]^{0.5} \quad (1)$$

where i denotes the index of a specific chemical and $(X^{\mathrm{T}}X)^{-1}$ is the covariance matrix. The center is estimated by the arithmetic mean vector. If the test chemical is located nearest to the gravity center of its actual class, then it is correctly classified. Otherwise it would be incorrectly assigned to the other class with the smallest Mahalanobis distance.

Decision Trees. The classification and regression tree (CART) is a nonparametric method which models the chemical data as a tree structure. The building of a CART model consists of three steps:[58,59] (1) an overlarge tree, also named the maximal tree, is constructed by recursive binary partitioning of the training chemical samples into regions that are increasingly homogeneous with respect to molecular descriptors; (2) the overgrown tree is pruned by cutting away some branches and a series of less complex trees are derived with improved predictive ability; and (3) the tree with the optimal size or node, which corresponds to the minimal classification error, is chosen by a cross-validation (CV) procedure.

The maximal tree fits the training chemicals almost perfectly but usually gives low predictive accuracy for test chemicals since the overlarge terminal nodes can yield an overfitted model. The optimal tree with less complexity but better predictive ability can be obtained by successively cutting the terminal branches from the maximal tree. The optimal tree node is determined by minimizing the cost-complexity parameter ($\mathrm{CP}_\alpha(T)$), consisting of the tree cost ($Q_l(T)$) and tree complexity ($|T|$):[58]

$$\text{minimize: } \mathrm{CP}_\alpha(T) = \sum_{l=1}^{|T|} n_l Q_l(T) + \alpha|T| \quad (2)$$

where $|T|$ and n_l denote the number of terminal nodes and the number of chemicals in node l, respectively; α, an adjustable tuning parameter between 0 and 1, represents the penalty for each additional terminal node and the best compromise between tree nodes and prediction error.

Support Vector Machines (SVM). This approach aims to find an optimal hyperplane as the decision boundary that can maximize the margin between the active and inactive classes of compounds. In the linearly separable case, the class boundary can be determined on the basis of the molecular features:[30−34]

$$w^{T}x + b = 0 \quad (3)$$

where b and w denote a threshold parameter and a weight vector normal to the hyperplane, respectively; $b/\|w\|$ represents the perpendicular distance to the origin with $\|w\|$ being the Euclidean norm. In case of nonlinearity, the complex class boundary can be modeled by kernel functions which map linearly inseparable input data into a higher dimensional feature space where the two classes can be linearly separated. For classification tasks, Gaussian radial basis function (RBF) is a commonly used kernel function due to its good generalization performance:[60]

$$K(x_i, x_j) = \exp(-\gamma \left\|x_i - x_j\right\|^2) \quad (4)$$

where x_i and x_j denote two independent molecular descriptors and γ is a tuning parameter controlling the amplitude of the kernel function and hence the generalization capability of the model. A set of non-negative variables ξ_i are introduced in order to allow violation of the margin constraints. An objective function is used to balance the model complexity and prediction errors:[32,33]

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (5)$$

The optimization requires simultaneously minimizing the sum of the training errors $\Sigma\xi_i$ and maximizing the margin $1/2\|w\|^2$ by applying a regularization parameter C.

Handling Data Imbalance. As shown in Table 1, the training set contains 176 active compounds and 849 inactive compounds, resulting in a ratio of active to inactive chemicals of 1:4.82, and hence less than 20% active samples are present in this data set. In addition, a much smaller number of active compounds (171) exist in the external test set with a total of 5073, leading to a ratio of 1:28.7 of the active to the inactive. This difference is that active fraction is intentional. The ~1800 chemicals tested in the larger battery of ToxCast ER assays were enriched with compounds in structural classes known to be estrogenic, such as steroids and phenols. A data set is considered to be class-imbalanced if one class contains significantly more samples than the other. In the case of machine learning (ML) from extremely imbalanced data, the interest usually leans toward correct classification of the minority or "rare" class. In our study, active compounds are designated as the positive class while inactive compounds are designated as the negative class. For the classification of highly imbalanced data sets, ML models are prone to assign most samples into the class in which the majority samples belong (i.e., inactive compounds), resulting in a large number of false negatives. There are two common approaches to cope with the problem of imbalanced data:[61,62] one is based on cost-sensitive learning and the other is to produce balanced data sets by applying sampling techniques. In order to reduce the bias, cost-sensitive learning imposes a larger penalty on the classification error for the minority class by incorporating class weights into the classifier and assigning more weight on the minority class. As an important tuning parameter for achieving desired performance, class weights are determined from the proportion of active to inactive chemicals in the training set. An alternate method for dealing with skewed data is to use sampling to make the class distribution more balanced by adjusting the proportion of active to inactive chemicals to approach 1:1. There are two basic sampling methods: oversampling and undersampling. That is, either the minority class is oversampled through the random replication of active chemicals or the majority class is undersampled through the random removal of inactive chemicals or some combination of these two approaches is deployed.

In the current study, we explored a new strategy for increasing the ratio of active to inactive compounds for managing the reasonably large and highly imbalanced chemical data sets. Unlike the traditional undersampling method which randomly balances the data class, the strategy of cluster-selection used here is to find the inactive chemical samples close to the boundary between active−inactive classes. As only inactive chemicals in clusters containing active chemicals are considered important for providing a significant contribution to the classification, removal of chemicals in clusters that contain only inactive samples does not substantially affect the model performance. Therefore, we retained all of the active compounds due to their rarity and extracted a subset of the inactive compounds from the clusters containing one or more active chemicals. The selected informative inactive compounds were combined with all of the active compounds to form a smaller training set for model building. For the training set, the extracted 456 inactive compounds were aggregated together with the 176 active compounds. This combination reduced the ratio of active-to-inactive compounds to 1:2.59 from an original ratio of 1:4.82.

Evaluation of Prediction Performance. In machine learning, the commonly used metrics for measuring the performance of classification models are the error rate and the overall accuracy. However, when they are used for evaluating models built on extremely imbalanced data, these measures are not appropriate and can lead to misleading predictive results because they do not consider misclassification costs and, hence, are strongly biased to favor the majority class as the evaluation metrics are dependent on the data distribution. A very high accuracy can still be achieved when a classifier predicts every case as the majority class. For instance, for the external test set (data set III) with the active/inactive chemical ratio of 1:28.7, the overall accuracy can still reach 96.6% even if all chemicals are predicted to be inactive, but this model would fail on all of the active chemicals.

To evaluate the prediction ability of the classification models built on imbalanced data sets, alternative metrics that measure the classification performance on positive and negative classes separately are needed. In the present study, we adopted metrics such as sensitivity, specificity, and geometric mean. Sensitivity, also referred to as the true positive rate or positive class accuracy, is the percentage of active compounds correctly predicted. Specificity, also known as the true negative rate or negative class accuracy, is the percentage of inactive compounds correctly predicted. Simply, sensitivity and specificity are the prediction accuracy for active and inactive classes, respectively. On the other hand, the overall prediction accuracy is the total percentage correctly predicted and is used to measure the overall prediction performance of the classification model. These statistic metrics are calculated by the following formulas

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (7)$$

$$\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (8)$$

where TP and TN denote the number of true positives (active chemicals) and true negative (inactive chemicals), respectively, while FP (false positives) and FN (false negatives) represent the incorrect classification of inactive chemicals into the active class and active chemicals into the inactive class, respectively. An ideal classification model should maximize the accuracy on the both sides of active and inactive classes. Thus, G-mean, the geometric mean of sensitivity and specificity, has been proposed as a criterion to emphasize the joint performance of sensitivity and specificity, which is defined as follows:[43,63]

$$\text{G-mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (9)$$

The optimal model can be selected from the ensemble of models and evaluated based on the G-mean value. Moreover, measures can be combined to create a two-dimensional performance graph, such as a receiver operating characteristic (ROC) curve, which is a graphical representation of the trade-off between TP and FP, and can be used to visualize, organize, and select classifiers based on their performance.[64,65] In the ROC curve, the FP rate and TP rate are represented on the *X*-axis and *Y*-axis, respectively. The lower left point (0, 0) indicates that all chemicals are classified as negative or inactive class while the upper right point (1, 1) represents the case where all chemicals are assigned to positive or active class. In addition, the point (0, 1) demonstrates perfect discrimination between the two classes whereas the line $y = x$ suggests that the class affiliation is randomly guessed. For quantitatively evaluating the overall performance of a classifier, we can measure the area under the ROC curve (AUC), which varies between 0 and 1. An AUC value of 1 suggests perfect classification performance while a value of 0.5 indicates a model with no discrimination capability.

Statistical Analysis. All data processing, multivariate analysis, and model building were implemented using the R statistical analysis software for Windows (version 2.15.1).[66] The packages *pheatmap*, *varSelRF*, *MASS*, *rpart*, *e1071* as well as *ROCR* in R were used to perform hierarchical clustering, feature selection, linear discriminant analysis (LDA), classification and regression tree (CART), support vector machine (SVM), and the receiver operating characteristic (ROC) analysis, respectively. Classification algorithms were executed on the normalized data set.

RESULTS AND DISCUSSION

A series of binary QSAR models were developed to classify the chemicals as active vs inactive according to their estrogenic activity on the basis of the most significant molecular descriptors and structural fingerprints selected via the RF method using three machine learning approaches: linear discriminant analysis, classification and regression trees, and support vector machines (LDA, CART and SVM). The data set I (training set) was trained with 10-fold CV to yield optimal parameters for all the three classifiers. In order to adequately assess the reliability and prediction ability of the QSAR models, both the internal and external validations were performed, where the internal validation was based on the ToxCast data set (II) and the external validation was accomplished using the Tox21 data set (III). The influence of various methods for handling the imbalanced data such as cost-sensitive algorithm, cluster-selection strategy, as well as over- and undersampling on the accuracy of the predictions was investigated. Figure 2 depicts the workflow of the whole classification process.

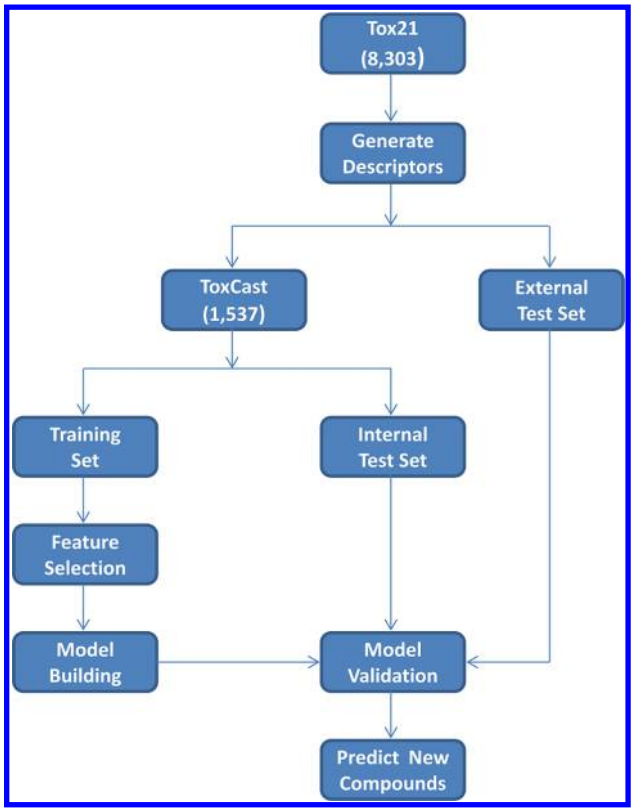


Figure 2. Flowchart for the whole classification process.

Feature Selection. To develop accurate and robust classification models, a sufficient number of chemical samples and appropriate number of molecular features are generally required. In our model building, a preliminary classification with all features would result in low prediction accuracy due to the existence of insignificant descriptors, as only a minority of the 51 molecular descriptors and 4328 bits of structural fingerprints will make significant contributions to the model. In order to extract the descriptors and fingerprints most relevant to the classification of active and inactive chemicals, feature selection (FS) was conducted to reduce the level of overfitting and the potential noise generated by irrelevant features.

During the FS using the random forest (RF) method, the weight score was calculated for each of the molecular descriptors and bits of structural fingerprints from the distribution of active and inactive chemicals in the training set, and then, these features were ranked by their importance in a descending order of their weight scores as illustrated in Figure 3. In principle, the larger the weight score is, the more discriminative the descriptor or fingerprint is, i.e., it makes a greater contribution to the correct classification of chemicals. All of the 51 molecular descriptors from QikProp and the top 51 features from 4328 bits of fingerprints are listed in Table 2. It can be observed from this table that the most significant molecular descriptors are SASA (total solvent accessible surface area) and volume (total solvent accessible volume). This is anticipated because a large polar surface area of a compound favors its binding to the receptor. In addition, among the higher ranked descriptors are QPlogKhsa (prediction of binding to human serum albumin), QPpolrz (predicted polarizability), QPlogPC16 (predicted hexadecane/gas partition coefficient), and CIQPlogS (conformation-independent predicted aqueous solubility). Likewise, QPlog-Po/w (predicted 1-octanol/water partition coefficient), an

dx.doi.org/10.1021/ci400527b | J. Chem. Inf. Model. 2013, 53, 3244−3261

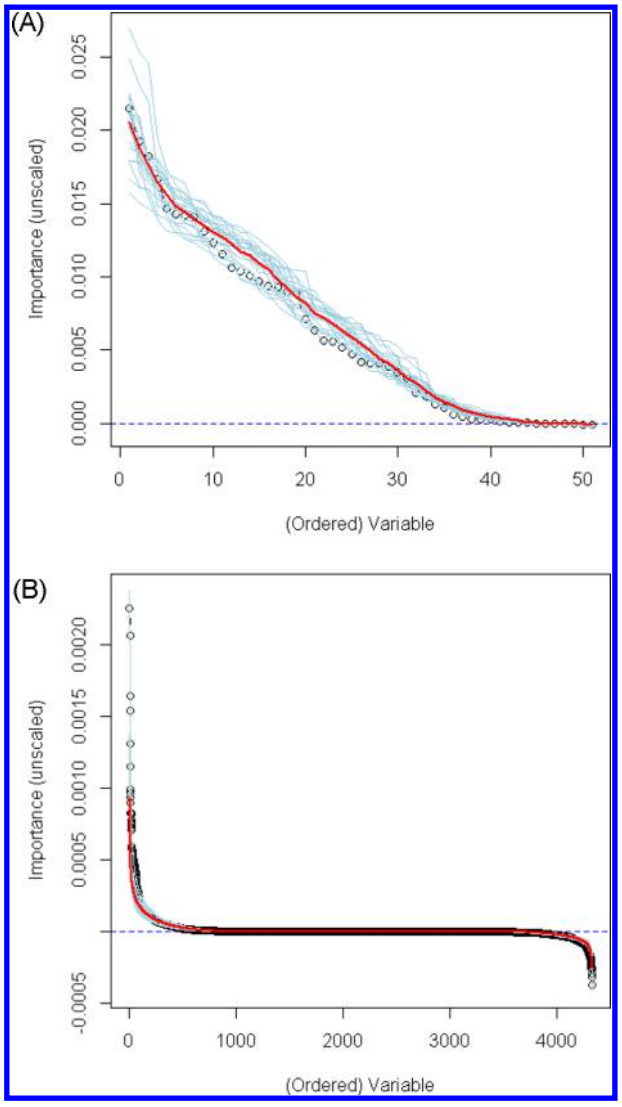


Figure 3. Feature importance computed from random permutations of random forest: (blue lines) feature importance from the permuted data sets; (red line) the average of the importance from the permuted data set. (A) 51 molecular descriptors from QikProp; (B) 4328 bits of structural fingerprints.

indicator of a chemical's ability to pass from aqueous environments through hydrophobic membrane barriers, was also found to be important in modeling discrimination between active and inactive chemicals. Furthermore, molecular globularity (glob) and van der Waals surface area of polar nitrogen and oxygen atoms (PSA) were found to be the key descriptors for constructing the model for predicting estrogenic activity. As shown in Table 2, when the fingerprint importance was assessed, it is observed that the fingerprints from the class of PUBCHEM took up a large portion within the top 51 most important features, followed by MACCS and then PADEL. When evaluating the combination of the descriptors and fingerprints, as one can see from the table, the top 37 selected features came from molecular descriptors, implying that molecular descriptors play a vital role and make more remarkable contributions while the structural fingerprints only play a marginal role in our model development.

The top 17 molecular descriptors and the top 19 bits of structural fingerprints have importance well above their counterparts with weight scores $>10.0 \times 10^{-3}$ and $>7.0 \times 10^{-4}$, respectively, and hence, they were retained for further model building of binary classification. The other features that are obviously redundant or unrelated to the prediction of the estrogenic activity were eliminated. The names, SMARTS, and substructures of the top 19 bits of fingerprints are listed in the Supporting Information (Table S2). The top 17 molecular descriptors can be divided into three categories, namely: (I) six descriptors in the category of simple molecular size and shape, such as molecular weight (mol_MW), surface area (SASA, PSA, and FISA (hydrophilic component of the SASA)), and volume (glob and volume); (II) four descriptors in the category of partition coefficients, such as QPlogPC16, QPlogPo/w, QPlogPoct (predicted octanol/gas partition coefficient), and QPlogPw (predicted water/gas partition coefficient); (III) seven descriptors in the category of predicted properties, including QPlogKhsa, QPpolrz, CIQPlogS, QPlogS (predicted aqueous solubility), QPPCaco (predicted apparent Caco-2 cell permeability), QPPMDCK (predicted apparent MDCK cell permeability), and QPlogKp (predicted skin permeability). The descriptors accptHB and donorHB, i.e., the estimated number of hydrogen bonds that would be accepted by the solute from water molecules or donated by the solute to water molecules, are ranked nos. 24 and 33, respectively, and hence, they are not key features for the model building.

Linear Discriminant Analysis (LDA). The percentages of correctly classified compounds in the training and test sets as well as on 10-fold CV, including sensitivity, specificity, overall accuracy, and G-mean, are presented in Table 3. As shown in the table, the prediction accuracies of the internal test set from different methods are comparable to each other when the 17 selected descriptors were employed. The internal test results are very close to those of the 10-fold CV, suggesting that the LDA models achieve good generalization. In addition, these models present a good balance between sensitivity and specificity with both ~70%, meaning that they have an equal ability to discriminate the active and inactive compounds. Receiver operating characteristic (ROC) curves for the three classifiers are shown in Figure 4, which compare the performance of the different methods for handling the imbalanced data sets. The solid and dotted lines represent the prediction of the training and internal test sets, respectively in Figure 4A–C. The area under the ROC curve (AUC) is proportional to the model performance. For LDA classification, oversampling method performed best in terms of AUC (0.745) in the validation of internal test set, followed by cost-sensitive (0.735), undersampling (0.723), and cluster-selection (0.708) methods. The use of cluster-selection did not improve the model performance compared to the other methods. One reason is that there may be some nonlinear relationship between the molecular structures and the estrogenic activity, but LDA is a linear classifier. When there is a nonlinear boundary between the active and inactive classes, removal of inactive samples far from the boundary does not change the nonlinear relationship. Rather, the prediction accuracy declines due to the loss of some discriminative inactive samples.

When the 19 selected bits of fingerprints were used to build the model, it is observed that the predicted accuracy, sensitivity, specificity, and G-mean are all lower than those from the descriptor model as given in Table 3. A possible reason for this difference is that molecular descriptors contain more important information needed for classification than fingerprints and play a more important role in the development of successful classification models. When combining the 17 descriptors and

Table 2. Results of Feature Selection through Random Forest

rank	descriptor (desc)	score ($\times 10^4$)	fingerprint (FP)	score ($\times 10^4$)	desc + FP	score ($\times 10^4$)
1	SASA	218	PUBCHEM_710	22.1	volume	109
2	volume	195	PUBCHEM_573	20.6	SASA	108
3	QPlogKhsa	188	PUBCHEM_688	16.6	QPpolrz	106
4	QPpolrz	171	PADEL_KR3788	15.4	QPlogKhsa	99.6
5	QPlogPC16	158	PUBCHEM_564	13.9	QPlogPC16	93.9
6	CIQPlogS	150	MACCS_131	11.7	QPlogPo/w	92.7
7	QPlogPo/w	148	PUBCHEM_341	10.2	CIQPlogS	89.2
8	QPPCaco	145	PUBCHEM_698	9.98	mol_MW	88.1
9	FISA	141	PADEL_KR3025	9.73	QPlogS	87.2
10	mol_MW	140	PADEL_KR3692	9.41	QPPCaco	86.8
11	QPlogPoct	133	PUBCHEM_689	9.06	glob	82.8
12	glob	131	FP4_2	8.89	PSA	82.4
13	PSA	129	MACCS_128	8.74	FISA	81.3
14	QPlogS	126	PADEL_SubFP2	8.54	QPlogPoct	79.1
15	QPPMDCK	125	FP3_3	8.22	QPPMDCK	77.6
16	QPlogPw	112	MACCS_154	7.97	QPlogPw	69.3
17	QPlogKp	104	MACCS_152	7.55	QPlogKp	68.2
18	Jm	96.5	MACCS_157	7.21	#nonHatm	64.2
19	QPlogBB	92.1	MACCS_125	6.85	Jm	60.6
20	#nonHatm	87.2	MACCS_135	5.96	QPlogBB	60.6
21	QPlogHERG	81.5	PUBCHEM_464	5.95	FOSA	55.2
22	Absorption (%)	79.9	PUBCHEM_640	5.94	QPlogHERG	54.8
23	FOSA	74.4	MACCS_107	5.86	Absorption (%)	53.6
24	accptHB	73.7	PUBCHEM_351	5.76	PISA	48.7
25	ACxDN^{-5}/SA	67.7	PUBCHEM_664	5.74	accptHB	47.7
26	PISA	62.3	MACCS_127	5.66	ACxDN^{-5}/SA	46.8
27	#rotor	51.0	PADEL_KR4080	5.52	#rotor	35.4
28	#in56	44.2	PUBCHEM_592	5.43	#in56	30.2
29	#ringatoms	44.2	MACCS_114	5.42	#NandO	29.3
30	#NandO	43.7	MACCS_142	5.30	#ringatoms	29.1
31	#stars	38.2	PUBCHEM_356	5.29	#stars	26.0
32	WPSA	27.4	PUBCHEM_680	5.17	WPSA	18.7
33	donorHB	25.9	PUBCHEM_535	5.14	CNS	18.6
34	CNS	22.5	MACCS_126	5.11	donorHB	17.3
35	#metab	16.6	PUBCHEM_556	5.08	#metab	13.1
36	#noncon	9.77	MACCS_112	5.03	#noncon	8.01
37	OralAbsorption	9.67	MACCS_115	5.00	OralAbsorption	7.41
38	#rtvFG	7.68	PADEL_KR3640	4.88	PUBCHEM_710	7.39
39	RuleOfFive	6.46	PUBCHEM_405	4.78	PUBCHEM_573	7.05
40	RuleOfThree	6.33	PUBCHEM_619	4.69	PADEL_KR3788	7.01
41	SAFluorine	4.49	PUBCHEM_709	4.47	PUBCHEM_688	6.86
42	#acid	4.18	PUBCHEM_637	4.42	MACCS_131	6.66
43	#amine	2.98	PUBCHEM_696	4.32	PUBCHEM_564	6.65
44	SAamideO	1.73	PUBCHEM_490	4.21	PUBCHEM_689	6.62
45	#amide	0.87	MACCS_138	4.13	MACCS_128	6.59
46	#in34	0.17	PUBCHEM_340	4.10	PUBCHEM_698	6.54
47	dipole	0.01	PADEL_KR3706	4.08	PADEL_KR3692	6.49
48	dip^2/V	0	MACCS_153	4.07	FP3_3	6.48
49	IP(ev)	0	FP3_27	4.04	PUBCHEM_341	6.45
50	EA(ev)	0	PADEL_EstateFP16	4.03	MACCS_125	6.39
51	#amidine	0	MACCS_108	4.01	FP4_2	6.35

19 bits of fingerprints together, our initial assumption was that different molecular features could encode different aspects of information for the model building, so they may complement each other to yield better prediction performance. The classification results, however, did not support this expectation. The best model was obtained using the cost-sensitive method with 73.3% training chemicals being correctly predicted, which is comparable to that from the descriptor model (72.0%). For the internal test set, the combination model correctly classified the chemical samples with an overall accuracy of 69.7%, slightly lower than 70.5% from the corresponding descriptor model. These findings show that models built using the combined techniques do not lead to better prediction ability than those developed from only the descriptors. Rather, the performance of the descriptor model is slightly superior to the combination model with respect to classification of individual classes and

dx.doi.org/10.1021/ci400527b | J. Chem. Inf. Model. 2013, 53, 3244−3261

Table 3. Prediction Performance from Linear Discriminant Analysis (LDA)

model		sensitivity (%)	specificity (%)	accuracy (%)	G-mean (%)
17 descriptors					
cost-sensitive (active/inactive = 176/849)	training set	72.2	72.0	72.0	72.1
	10-fold CV	69.3	71.0	70.7	70.2
	internal test set	72.7	68.4	69.1	70.5
	external test set	70.2	63.0	63.3	66.5
undersampling (active/inactive = 176/176)	training set	72.2	73.8	73.0	73.0
	10-fold CV	68.8	70.5	69.6	69.6
	internal test set	70.5	68.2	68.6	69.3
	external test set	69.6	61.1	61.4	65.2
oversampling (active/inactive = 849/849)	training set	72.8	71.9	72.3	72.3
	10-fold CV	72.8	73.0	72.9	72.9
	internal test set	72.7	70.1	70.5	71.4
	external test set	70.2	64.2	64.4	67.1
cluster-selection (active/inactive = 176/456)	training set	67.6	65.4	66.0	66.5
	10-fold CV	65.3	65.6	65.5	65.4
	internal test set	65.9	69.1	68.6	67.5
	external test set	64.9	60.5	60.7	62.7
19 bits of fingerprints					
cost-sensitive (active/inactive = 176/849)	training set	68.8	67.4	67.6	68.1
	10-fold CV	66.5	66.8	66.7	66.7
	internal test set	68.2	62.5	63.5	65.3
	external test set	56.1	60.2	60.0	58.1
undersampling (active/inactive = 176/176)	training set	73.3	67.1	70.2	70.2
	10-fold CV	68.2	63.6	65.9	65.9
	internal test set	71.6	58.7	60.9	64.8
	external test set	55.6	56.8	56.7	56.2
19 bits of fingerprints					
oversampling (active/inactive = 849/849)	training set	68.9	66.2	67.6	67.6
	10-fold CV	68.6	64.1	66.3	66.3
	internal test set	68.2	60.8	62.1	64.4
	external test set	55.6	58.8	58.7	57.2
cluster-selection (active/inactive = 176/456)	training set	65.9	59.4	61.2	62.6
	10-fold CV	62.5	57.5	58.9	60.7
	internal test set	68.2	60.8	62.1	64.4
	external test set	53.2	60.6	60.3	56.8
17 descriptors + 19 bits of fingerprints					
cost-sensitive (active/inactive = 176/849)	training set	70.5	73.9	73.3	72.2
	10-fold CV	68.2	72.2	71.5	70.2
	internal test set	67.1	70.3	69.7	68.7
	external test set	66.7	63.8	63.9	65.2
undersampling (active/inactive = 176/176)	training set	73.9	73.3	73.6	73.6
	10-fold CV	67.1	66.5	66.8	66.8
	internal test set	65.9	65.8	65.8	65.9
	external test set	65.5	60.5	60.7	63.0
oversampling (active/inactive = 849/849)	training set	72.2	72.8	72.5	72.5
	10-fold CV	70.9	71.3	71.1	71.1
	internal test set	67.1	68.6	68.4	67.9
	external test set	66.1	62.6	62.8	64.3
cluster-selection (active/inactive = 176/456)	training set	68.8	69.3	69.1	69.0
	10-fold CV	63.6	67.5	66.5	65.5
	internal test set	64.8	65.6	65.4	65.2
	external test set	56.7	64.4	64.2	60.4

overall success rates. This is because given a fixed number of chemicals, an excessive number of molecular features may cause the model to be overfitted.

The LDA classification models can indicate the specific molecular features responsible for the discrimination between active and inactive classes. To evaluate the importance of the descriptors and fingerprints to the classification model, the group mean of each feature is displayed in Figure 5. Features with large positive or negative values contributed more to the model than those with small ones. For the descriptor set, QPpolrz, QPlogPC16, QPlogPo/w, QPlogS, CIQPlogS, and QPlogKhsa are the descriptors most contributing to the classification (Figure 5A), and the most discriminative ones are PUBCHEM_573, PUBCHEM_688, PUBCHEM_564, and PUBCHEM_710 from the fingerprint set (Figure 5B). Besides these four bits of PUBCHEM fingerprints, PADEL_KR3788, MACCS_131, PADEL_KR3025, and MACCS_125 also contributed significantly to the binary model. The descriptor model considered QPlogKhsa as most contributing to the classification model, whereas the fingerprint model identified PUBCHEM_710 as the most important feature. These observations agree well with the results in the RF selection process (Table 2). The descriptors that are activity-enhancing include QPpolrz, QPlogPC16, QPlogPo/w, and QPlogKhsa which are related to a compound's hydrophobicity, whereas those that are activity-decreasing are glob, QPlogS, and CIQPlogS. Seven dominant bits of fingerprints were identified, which are all the activity-enhancing

dx.doi.org/10.1021/ci400527b | J. Chem. Inf. Model. 2013, 53, 3244−3261

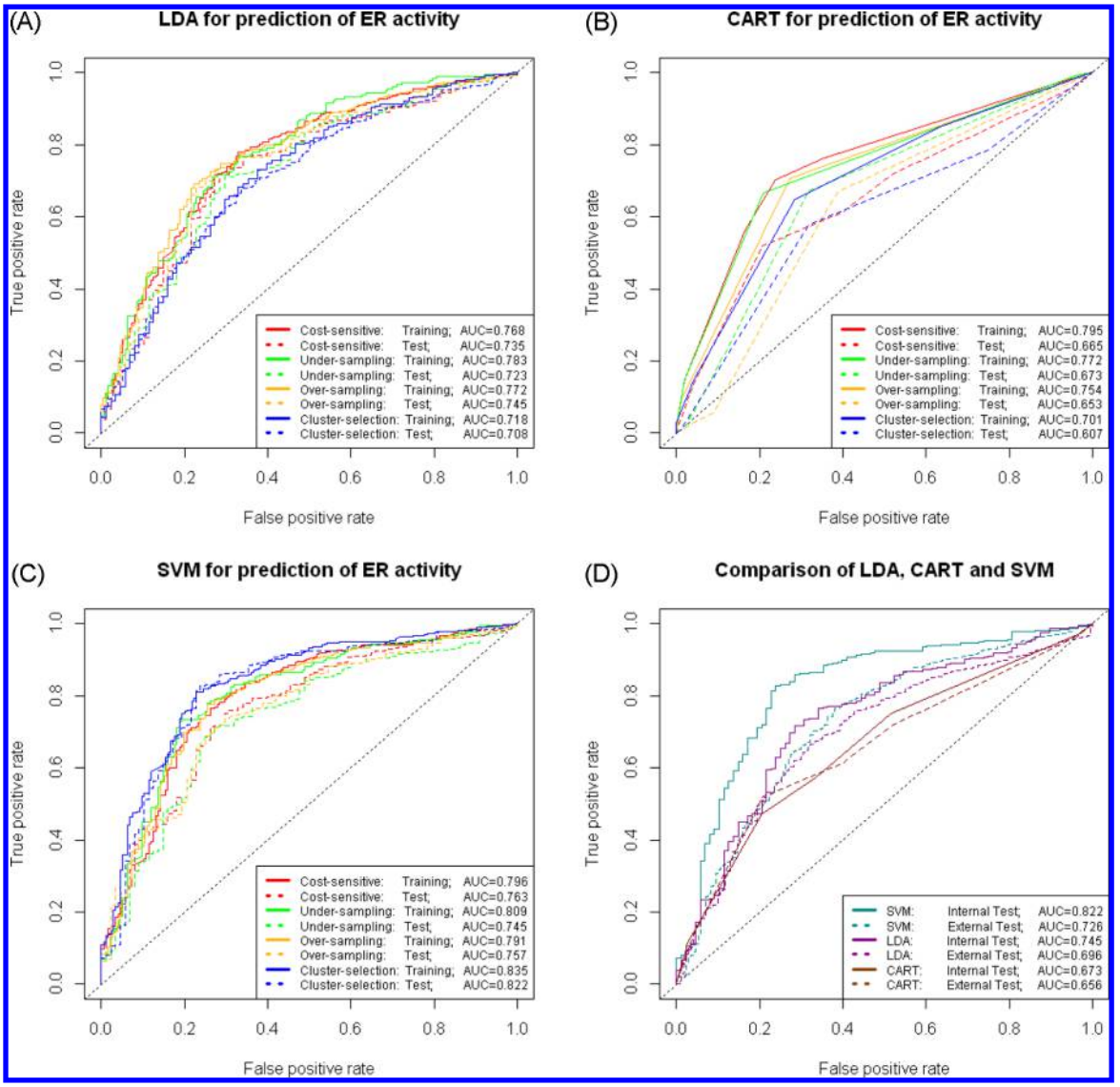


Figure 4. (ROC) curves of the three classifiers for the descriptor models: (A) LDA; (B) CART; (C) SVM. (D) ComReceiver operating characteristicparison of the classification performance from the best models of LDA (oversampling), CART (undersampling), and SVM (cluster-selection).

fingerprints from PUBCHEM, while the activity-decreasing fingerprints came from MACCS and PADEL.

Classification and Regression Trees (CART). Classification trees were built using three sets of molecular features corresponding to 17 descriptors, 19 bits of fingerprints, and their combination, respectively. The two designated active and inactive classes corresponded to the response variables. The optimal tree size (i.e., the number of nodes) was determined using 10-fold CV. The classification tree with seven nodes for the descriptor model is presented in Figure 6A. The first split at QPlogKhsa = −0.1655 divides the chemical samples into two groups. On the left side, QPpolrz splits the chemicals into active and (active + inactive) groups, and CIQPlogS further divides the (active + inactive) group into active and inactive classes. On the right side, the chemicals are split into (active + inactive) and inactive groups by SASA, and the (active + inactive) group is divided into active and inactive classes by QPlogPo/w and followed by the splitting of volume. The distribution of the training chemical samples over different terminal nodes is also shown in Figure 6A. The 10-fold CV relative error (RE) versus the cost-complexity parameter (CP) and the tree node is displayed in Figure 6C, where the lowest error occurred at tree node = 7, corresponding to the classification tree in Figure 6A. When modeling the data set for fingerprints, the optimal classification tree consists of six terminal nodes (Figure 6B). PUBCHEM_710 splits off the chemicals into (active + inactive) and inactive groups, and then PADEL_KR3788, PUBCHEM_688, MACCS_131, and PUBCHEM_573 sequentially divide the chemicals on the left-hand side into two separate classes: active and inactive. Figure 6D shows the RE versus the CP and tree node, where the error decreased as the number of terminal nodes increased and attained a minimal at tree node = 6.

Table 4 gives the comparison of the prediction accuracy from different methods with the RF-selected features by using various evaluation sets. For the descriptor model, the cost-sensitive and under-sampling methods yielded models of similar quality and outperformed oversampling and cluster-selection methods. Although the cluster-selection method shares the same prediction ability for active compounds as the undersampling method, for inactive chemicals the prediction accuracy decreases

dx.doi.org/10.1021/ci400527b | J. Chem. Inf. Model. 2013, 53, 3244−3261

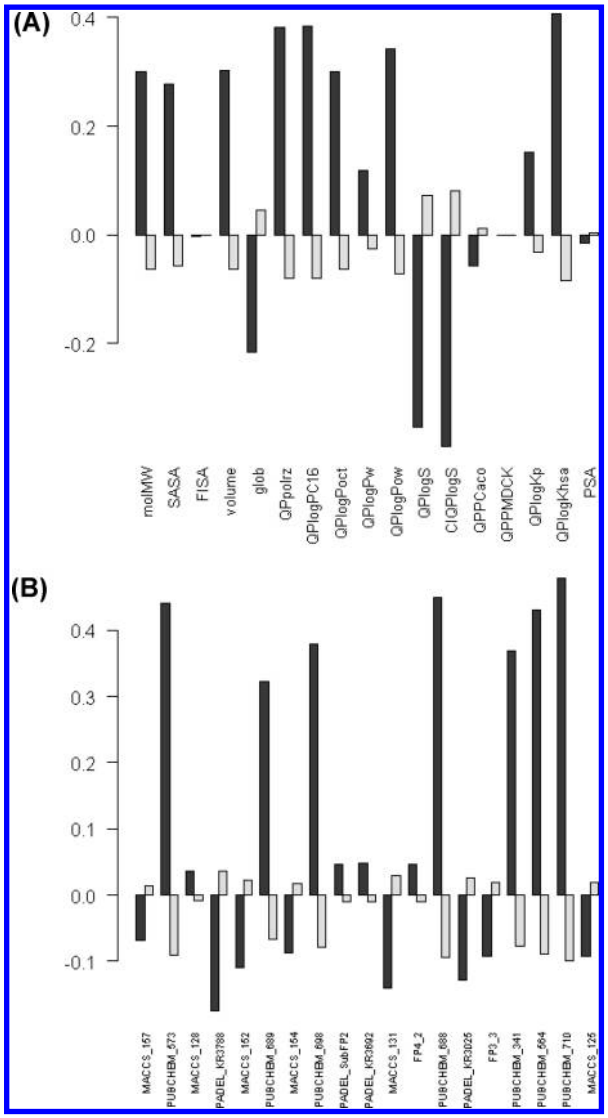


Figure 5. Group mean (scaled to 0 mean and unit variance) of molecular descriptors (A) and structural fingerprints (B): (black color) active class; (gray color) inactive class.

by 10.0%, leading to a dramatic drop of both overall accuracy and G-mean due to the removal of some discriminative inactive chemicals in the cluster selection process. The oversampling method has comparable performance to cost-sensitive and undersampling methods based on specificity, but it presents a relatively low sensitivity with a reduction of 10−15% compared to that of the other two methods, indicating that the model has a poor ability to discriminate the active compounds. From this point of view, one can consider that the undersampling and cost-sensitive models are more favorable than the other, as illustrated in the ROC curves in Figure 4B.

The descriptor model built from the oversampling method achieved balanced prediction accuracies for the training set with a sensitivity of 74.9% and a specificity of 75.4%. However, this method gave a model with a significant difference in predictive power between the training and test sets for the active class. While the model was very good at predicting inactive compounds on the test sets, it performed poorly when classifying active compounds with low sensitivity, leading to a great disparity between sensitivity and specificity. The model suffers from overfitting because it exhibits much higher prediction accuracy for the training set than that of the independent sets, and hence, it loses general prediction ability. Oversampling is prone to increase the likelihood of overfitting, since it makes exact copies of the active chemical samples to raise the weight of this class by replication but does not increase further useful information. The similar phenomenon occurred in our SVM model.

When the 19 selected bits of fingerprints were employed to derive the classification model, the prediction ability of the four methods is comparable to each other. Overall, the performance of the fingerprint model is inferior to that of the descriptor model based on all metrics for both internal and external test sets. Finally, it should be noted that the performance of the classification tree constructed using the combined features (17 descriptors + 19 bits of fingerprints) is equivalent to the tree constructed using only the molecular descriptors. Although the combination of descriptors and fingerprints contains more features and hence more structural information, the CART model selected only molecular descriptors QPlogKhsa, QPpolrz, QPlogPo/w, CIQPlogS, SASA, and volume, etc. as the splitting variables and ignored the fingerprints entirely. The tree-building algorithm chose the best feature at every split, and features that are less relevant to the estrogenic activity were omitted. Thus, for CART modeling, dimension reduction or feature selection is not essential. This observation highlights the singular importance played by descriptors in discriminating the active class from the inactive class in the CART model, which is consistent with the results from RF feature selection, where the top 37 features came from the molecular descriptors (Table 2).

Support Vector Machines (SVM). For comparison with the LDA and CART models, SVM classification was also performed to develop nonlinear models based on the same subsets of molecular descriptors and structural fingerprints. The SVM models with RBF kernel were trained and the parameters C and γ were optimized. A wide range of C and γ values were tuned simultaneously from a coarse grid ($C \in [10^{-2}, 10^{4}]$ and $\gamma \in [10^{-4}, 10^{2}]$, both in log 10 units) with a grid spacing of log 10^{n}. A contour plot was generated in logarithmic scales after all the combinations of C and γ had been searched. The optimization grids, plotted as the 10-fold CV error rate for the descriptor and fingerprint models, are shown in Figure 7. The small deep red "island" in Figure 7A represents an optimal setting of the C and γ values which yield the lowest prediction error for the descriptor model. A subregion ($C \in [1, 10^{2}]$ and $\gamma \in [10^{-2}, 1]$) showing relatively better performance was identified. The range of C and γ values was narrowed over a finer grid to increase the resolution and the final optimal parameters with $C = 8.5$ and $\gamma = 7.6 \times 10^{-2}$ were achieved. In the case of the fingerprint model, the optimal combination of C and γ is located in a single deep red triangle on a red band (Figure 7B). The values of $C = 1.8$ and $\gamma = 2.7 \times 10^{-3}$ produced the highest 10-fold CV accuracy.

The results of the classification models based on three different feature sets using the optimal paired values of C and γ are presented in Table 5. The ability of an SVM model to correctly classify active and inactive compounds in the training and test sets varies considerably with the molecular feature sets. The model employing fingerprints as predictive variables gave a moderate accuracy. Better results were obtained when the 17 molecular descriptors were employed, and the model consistently outperformed the one employing fingerprints. For the combination set, the resultant model's ability to correctly predict both active and inactive compounds slightly declined compared to the descriptor model.

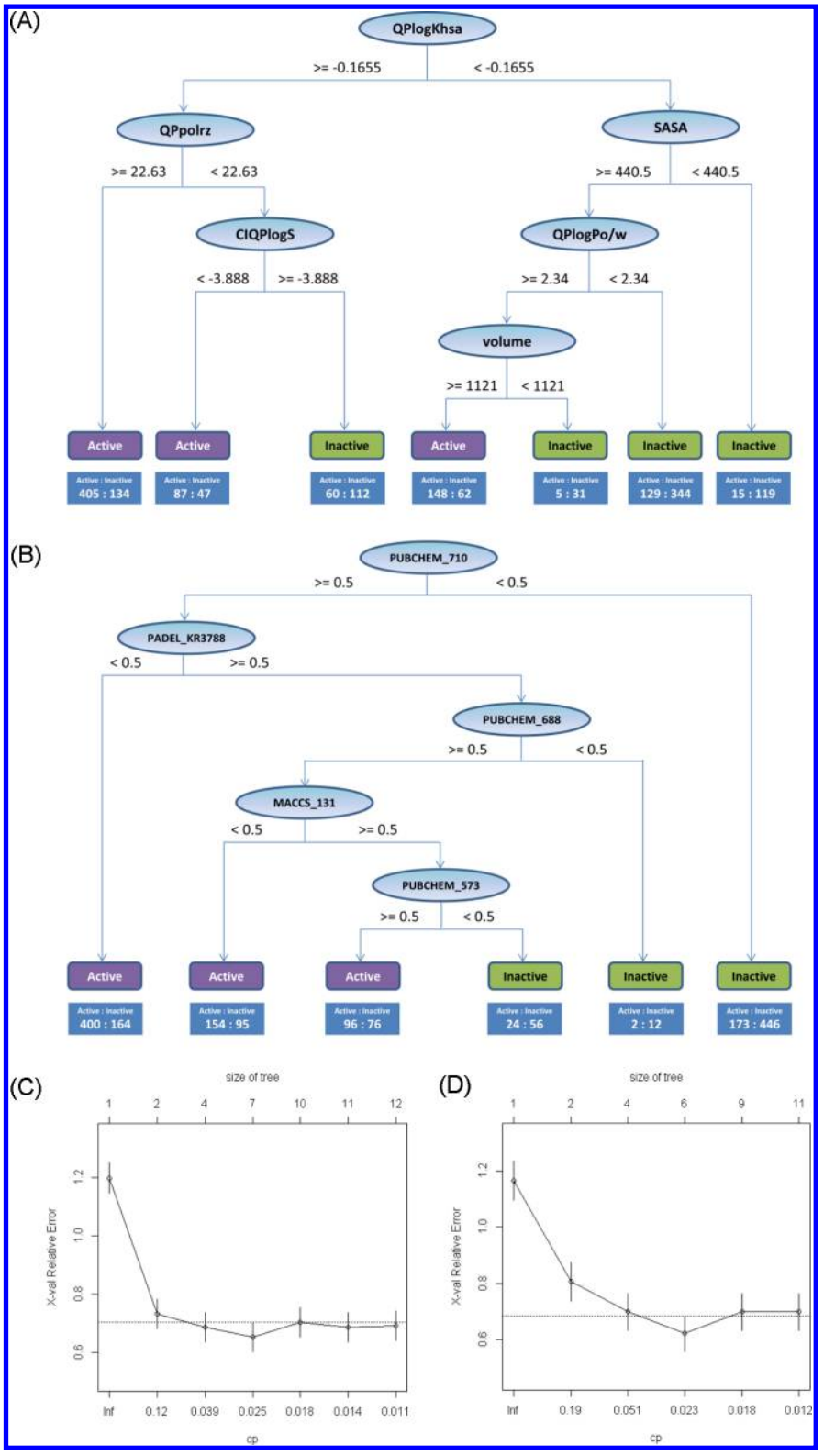


Figure 6. Classification trees (A and B) and corresponding plots of the 10-fold cross-validated relative error versus the complexity parameter CP (bottom axis) and tree size (top axis) (C and D) using cost-sensitive algorithm. The vertical bar for each point in parts C and D represents the standard error. (A and C) Model of molecular descriptors. (B and D) Model of structural fingerprints.

The results show that class discrimination improved significantly by the cluster-selection strategy with the SVM model built using the 17 descriptors. The model achieved overall prediction accuracies of 80.1% and 79.4% for the training set and 10-fold CV with G-mean values of 79.3% and 78.8%, respectively. It successfully recognized 67 out of 88 active compounds with a sensitivity of 76.1%, and also correctly categorized 351 out of the 424 inactive compounds giving a specificity of 82.8% along with an overall accuracy of 81.6% by the internal test set. As shown in Table 5, the prediction accuracies of the model based on the 10-

dx.doi.org/10.1021/ci400527b | J. Chem. Inf. Model. 2013, 53, 3244−3261

Table 4. Prediction Performance from Classification and Regression Tree (CART)

model		sensitivity (%)	specificity (%)	accuracy (%)	G-mean (%)
		17 descriptors			
cost-sensitive	training set	76.1	72.3	73.0	74.2
(active/inactive = 176/849)	10-fold CV	65.3	70.3	69.5	67.8
	internal test set	65.9	70.8	69.9	68.3
	external test set	66.7	66.4	66.4	66.6
undersampling	training set	73.9	73.3	73.6	73.6
(active/inactive = 176/176)	10-fold CV	71.0	63.6	67.3	67.2
	internal test set	70.5	67.7	68.2	69.1
	external test set	62.0	66.6	66.5	64.3
oversampling	training set	74.9	75.4	75.2	75.2
(active/inactive = 849/849)	10-fold CV	73.1	74.2	73.7	73.7
	internal test set	55.7	68.9	66.6	62.0
	external test set	52.6	65.3	64.9	58.6
cluster-selection	training set	71.6	64.9	66.8	68.2
(active/inactive = 176/456)	10-fold CV	67.6	56.8	59.8	62.0
	internal test set	68.2	57.8	59.6	62.8
	external test set	65.5	58.7	58.9	62.0
		19 bits of fingerprints			
cost-sensitive	training set	76.7	60.5	63.3	68.1
(active/inactive = 176/849)	10-fold CV	72.2	60.7	62.6	66.2
	internal test set	69.3	53.1	55.9	60.7
	external test set	69.0	52.6	53.1	60.3
undersampling	training set	70.5	68.7	69.6	69.6
(active/inactive = 176/176)	10-fold CV	65.3	60.8	63.1	63.0
	internal test set	61.4	59.0	59.4	60.2
	external test set	56.1	57.3	57.3	56.7
oversampling	training set	71.7	63.5	67.6	67.5
(active/inactive = 849/849)	10-fold CV	70.3	64.0	67.1	67.1
	internal test set	60.2	60.6	60.5	60.4
	external test set	59.1	58.5	58.5	58.8
cluster-selection	training set	66.5	63.2	64.1	64.8
(active/inactive = 176/456)	10-fold CV	52.3	62.3	59.5	57.1
	internal test set	56.8	59.4	59.0	58.1
	external test set	53.8	59.1	58.9	56.4

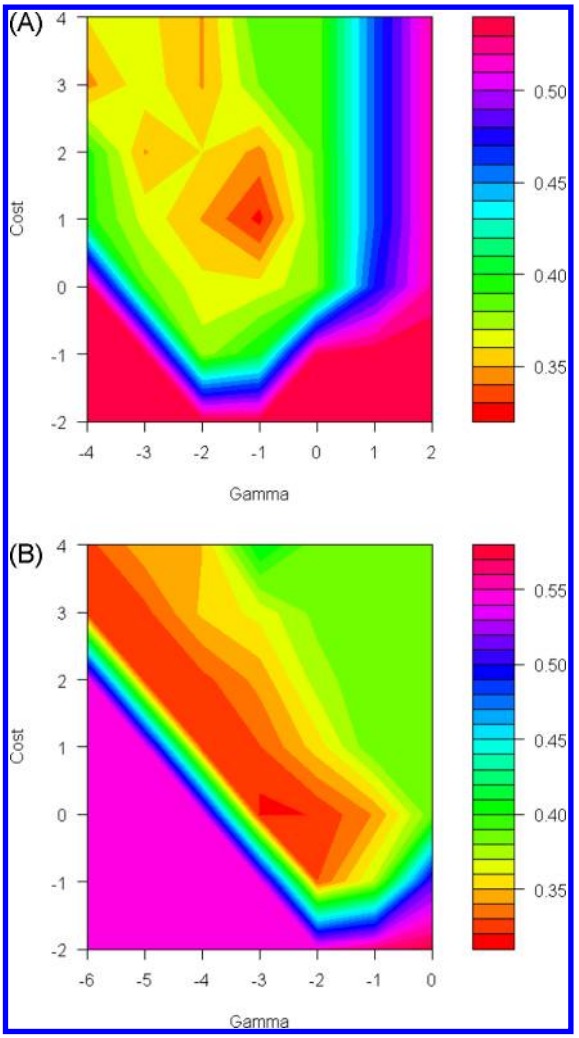


Figure 7. Contour plots in decimal logarithmic scales obtained from 7 × 7 grid search of the optimal values of γ (gamma) and C (cost) for the SVM model using cost-sensitive algorithm. (A) Model of molecular descriptors. (B) Model of structural fingerprints.

fold CV and those based on independent validation sets are very similar, and hence, the model is unlikely to be overfitted. In addition, the results obtained on the internal test set are relatively consistent with those obtained in training set, and both are substantially better than the corresponding values derived from other methods as observed in the ROC curves in Figure 4C. These findings demonstrate that the best SVM model built from the cluster-selection method is reliable and robust for identifying the activity of compounds. When applied to the external test set, the strategy yields a more balanced model with respect to the prediction of active and inactive classes with an overall prediction accuracy of 70.8%. This level of performance is satisfactory, considering the large size and extreme imbalance of the external test set.

In a highly imbalanced situation where the inactive compounds exceed the active compounds by a significant amount, an SVM classifier is very sensitive to the high class imbalance because the hyperplane of SVM is pushed toward the minority class, i.e., the active class. As a result, the prediction ability of such a model can be poor. When a fraction of the inactive samples is removed and the data become moderately imbalanced, the SVM model tends to be more accurate. The reason is that only support vectors (SVs) are used for model

Table 5. Prediction Performance from Support Vector Machine (SVM)

model		sensitivity (%)	specificity (%)	accuracy (%)	G-mean (%)
		17 descriptors			
cost-sensitive	training set	73.9	74.6	74.4	74.3
(active/inactive = 176/849)	10-fold CV	72.7	74.1	73.9	73.4
	internal test set	70.5	74.8	74.0	72.6
	external test set	67.3	68.4	68.3	67.8
undersampling	training set	75.6	76.1	75.9	75.9
(active/inactive = 176 176)	10-fold CV	71.6	72.2	71.9	71.9
	internal test set	72.7	71.2	71.5	72.0
	external test set	67.3	70.1	70.0	68.7
oversampling	training set	74.8	76.4	75.6	75.6
(active/inactive = 849/849)	10-fold CV	74.3	75.9	75.1	75.1
	internal test set	67.1	76.2	74.6	71.5
	external test set	64.9	71.5	71.3	68.1
cluster-selection	training set	77.8	80.9	80.1	79.3
(active/inactive = 176/456)	10-fold CV	77.3	80.3	79.4	78.8
	internal test set	76.1	82.8	81.6	79.4
	external test set	68.4	70.9	70.8	69.6
		19 bits of fingerprints			
cost-sensitive	training set	82.4	67.1	69.8	74.4
(active/inactive = 176/849)	10-fold CV	72.2	65.7	66.8	68.9
	internal test set	67.0	61.3	62.3	64.1
	external test set	59.7	59.6	59.6	59.7
undersampling	training set	79.0	76.1	77.6	77.5
(active/inactive = 176/176)	10-fold CV	70.5	65.9	68.2	68.2
	internal test set	65.9	59.2	60.4	62.5
	external test set	60.8	54.0	54.3	57.3
		19 bits of fingerprints			
oversampling	training set	78.6	71.0	74.8	74.7
(active/inactive = 849/849)	10-fold CV	74.6	68.1	71.3	71.3
	internal test set	61.4	65.1	64.5	63.2
	external test set	55.0	63.7	63.4	59.2
cluster-selection	training set	75.6	70.0	71.5	72.7
(active/inactive = 176/456)	10-fold CV	75.0	69.1	70.7	72.0
	internal test set	64.8	75.0	73.2	69.7
	external test set	59.1	57.3	57.4	58.2
		17 descriptors + 19 bits of fingerprints			
cost-sensitive	training set	84.1	76.8	78.1	80.4
(active/inactive = 176/849)	10-fold CV	73.9	74.6	74.4	74.3
	internal test set	65.9	75.2	73.6	70.4
	external test set	59.1	71.1	70.7	64.8
undersampling	training set	77.3	76.7	77.0	77.0
(active/inactive = 176/176)	10-fold CV	67.6	66.5	67.1	67.1
	internal test set	69.3	65.1	65.8	67.2
	external test set	60.2	62.8	62.7	61.5
oversampling	training set	78.2	76.9	77.6	77.6
(active/inactive = 849/849)	10-fold CV	76.4	73.3	74.9	74.8
	internal test set	65.9	73.1	71.9	69.4
	external test set	60.8	68.1	67.8	64.3
cluster-selection	training set	90.9	87.9	88.8	89.4
(active/inactive = 176/456)	10-fold CV	72.7	85.1	81.7	78.7
	internal test set	67.0	85.4	82.2	75.6
	external test set	64.3	70.5	70.3	67.3

building and many majority samples (inactive chemicals) far from the decision boundary can be disregarded without affecting classification. In the meanwhile the margin between the closest data points of active and inactive classes of compounds is increased, which is favorable for the model performance. The cluster-selection strategy shows great improvement over the other three methods based on all evaluation metrics, demonstrating the usefulness of our strategy for selecting a suitable set of inactive compounds to enhance the prediction ability. Retaining all of the active compounds and discarding samples of clusters containing no active chemicals resulted in the increased classification accuracy.

We also investigated a balanced data set where the inactive chemicals were selected on the basis of their pairwise Tanimoto coefficients representing molecular similarity and diversity. In order to achieve a ratio of 1:1 for these two classes, inactive chemicals, which share a Tanimoto coefficient of $\geq$0.7865 to one of the active chemicals, were collected. The SVM classifier gave a sensitivity of 71.0% and a specificity of 78.4% for the internal test set, which are not better than those from the cluster-selection method. The prediction ability of the classification model is highly influenced by the molecular similarity between the active and inactive chemicals. The chemical molecules that share close similarity to each other but possess opposite class signs can decrease the prediction performance because the model is unable to discriminate these chemicals very well.

The results given in Tables 3−5 show that SVM outperformed the other classifiers and produced the best prediction accuracies

dx.doi.org/10.1021/ci400527b | J. Chem. Inf. Model. 2013, 53, 3244−3261

in terms of internal and external prediction ability for all the data sets. The prediction accuracy of >70% was achieved in all cases considered for both the training and test sets. The order in overall prediction performance is SVM > LDA > CART by a comparison of the three models obtained in this study. SVM, a nonlinear algorithm for classification, is superior to that of LDA, which implies that using the same set of molecular features, the SVM approach is capable of recognizing highly nonlinear or overlapping boundary between active and inactive compounds. In contrast, the LDA approach can only capture linear relationships between the two classes. ROC curves show the best performance from the SVM model in both internal and external test sets (Figure 4D). It should be noted that the results from the internal test are always better than those from the external test. The difference in the performance could be explained by a different diversity of molecules in the two test sets. As one can see from Figure 1, data set II (internal test set) shares a similar chemical space with data set I (training set), which may account for the reasonably good prediction of our model on the internal test set. On the other hand, data set III (external test set) is composed of a wider range of chemical structures and physicochemical characteristics, implying that it is a more challenging test set for our model. This is in accordance with the relatively lower prediction performance for this data set.

When examining the misclassifications from the optimal SVM model with 17 molecular descriptors and cluster-selection method, it is observed that the 39 false negatives out of 176 active chemicals in the training set have a low ER Interaction Score near the borderline between the two classes. The same situation occurs on the internal test set, where all of the 21 false negatives involve active chemical samples with ER Interaction Scores of less than 15% while most of the 67 true positives have strong estrogenic activity with ER Interaction Scores of greater than 20%, as illustrated in the boxplots (Figure 8).

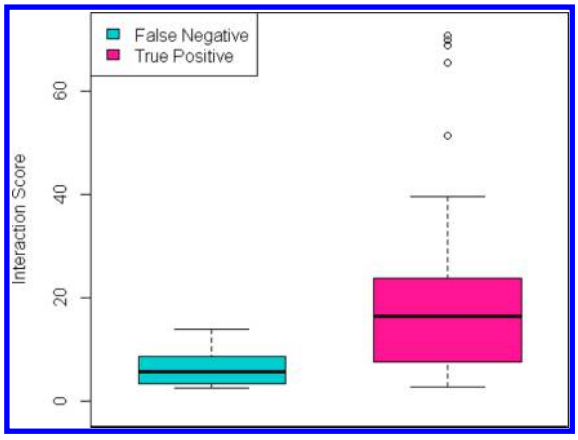


Figure 8. Boxplots of Interaction Scores from 88 active chemicals of the internal test set. Respectively, 67 and 21 chemicals were predicted as positive and negative classes by the SVM model using 17 descriptors and cluster-selection method.

It is useful to compare our results with those of some other recently published ER models. Hong et al.[67] reported a study on the prediction of estrogen receptor binding for a set of environmental chemicals consisting of 232 training and 463 test samples. The tree-based models achieved an overall accuracy of 82.5% on the test set. This is comparable to our best result using the SVM plus cluster selection method, which result in 81.6% accuracy (G-mean =79.4%) (Table 5). Li and Gramatica[68] derived QSAR classification models using 3 mehtods: *k*-nearest neighbors (kNN), the local lazy method, and random forests (RF). The models were built on 645 training chemicals and validated on 193 chemicals, yielding a maximum sensitivity and specificity of 82.9% and 83.8%, respectively, for a G-mean of 83.4% (all with the RF method). Usually, the quality of the classification models depends on many factors, such as the size of the data sets, the diversity of the chemical structures, the accuracy of the bioassays, the criteria for defining the cutoff between active and inactive classes, the descriptor sets and feature selection algorithms, as well as the machine learning methods, etc. Another useful comparison is with work by authors such as Chang et al.[43] examining the difficulty of achieving high prediction accuracy with imbalanced data sets such as ours. They built a series of models (not for ER) against two in vitro data sets and found test G-mean values in the 50−60% range, despite training G-means above 80%.

CONCLUSIONS

In the present study, we exploited the HTS data acquired from ToxCast and Tox21 programs for 8000 environmental chemicals and developed binary QSAR classification models that related the chemical structures to estrogenic activity by the application of three machine learning methods, LDA, CART, and SVM. These in silico models were trained with the smaller ToxCast for differentiating active and inactive chemicals and validated with the larger Tox21 data set. Combinations of various molecular descriptors and classifiers were performed to determine the optimal features and methods for generating the most predictive models. The limitations of the model's predictive capability outside the training set were also examined.

Not all of these molecular descriptors and structural fingerprints were essential for the prediction of the estrogenic activity of chemicals, and irrelevant features decreased the prediction accuracy. Feature selection by using random forest, which ranks the importance of each descriptor in the classification process, was useful in eliminating the unrelated and redundant descriptors to ER activities and improved the model's prediction performance. The molecular descriptors captured important information and were more discriminative than fingerprints in our binary classification. The models employing descriptors presented significantly superior results than those employing fingerprints.

To assess the performance of these machine learning methods, it is useful to examine whether their prediction ability is at a similar level in terms of overall accuracy, sensitivity, specificity, and G-mean. The best model was derived from SVM with the optimal settings of the RBF kernel function and the set of descriptors selected by RF method. When compared with LDA and CART, the SVM classification model presented better statistics and produced improved results, not only in cross validation, but also in the prediction of two independent test sets, giving the highest sensitivity of 76.1% and specificity of 82.8% for the internal validation set.

Although SVM performed very well for discrimination between active and inactive chemicals, the problem is that SVM is a black box technique. Hence, it is difficult to interpret the important role of functional groups and physicochemical properties to the model and explicitly provide information about the relationship between molecular features and biological activity. In contrast, LDA and CART models offer structural insights into the classification of chemicals, and reveal some key descriptors governing the estrogenic activity. The hydro-

phobicity is the most vital molecular property for increasing the estrogenic activity of a compound. The most significant descriptors usually include QPlogKhsa and QPlogPo/w. Likewise, QPpolrz was also found to be important in modeling. SASA and volume were also identified as positive features that increased potency. Only two descriptors were found to be significant in decreasing chemical activity: QPlogS and CIQPlogS. Fingerprint PUBCHEM_710 was always selected as the first feature to obtain a predictive model using CART. Although CART achieved lower classification accuracy than SVM and LDA, it is the simplest model with the best interpretability and no feature selection is necessary. Therefore, different classifiers can be employed to satisfy different information needs for our purposes.

In order to deal HTS data sets with highly skewed class distribution, four methods were investigated and compared. One is based on cost-sensitive learning algorithm while the others apply oversampling or undersampling to make the training data more balanced. Of these variations considered in the current research, cluster-selection method was found to be very effective in overcoming class imbalance, because it can minimize the effect of information loss for inactive class while maximize the effect of active class. This strategy provides a means to extract important inactive chemicals from the majority class and to eliminate the unimportant ones in order to rebalance the data set.

These models were developed by using ToxCast data set with about 1500 known active and inactive chemicals, which is substantially less in number and diversity than the Tox21 data set with 8000 chemicals and, thereby, covers a small portion of the chemical space of the Tox21. The reliability of the prediction model was strongly dependent on the structural similarity between training compounds and test compounds. Although satisfactory results were achieved in both cross validation and internal tests, the predictive power of the built model for a test set that is beyond the training chemical space, such as the Tox21 data set, should not be anticipated without caution. The performance of our classification model could be further improved with the availability of additional experimental activity data representative of a more diverse set of compounds.

Overall, the present research work suggests that the binary QSAR classification models are useful for in silico prediction of the estrogenic activity and for characterizing the molecular features of environmental chemicals, and highly accurate predictive models can be built based on chemical descriptors. These models can be applied in virtual screening of large databases for identifying compounds with potential risk at a reduced cost.

■ ASSOCIATED CONTENT

Supporting Information

Table S1: 51 molecular descriptors and properties generated from QikProp. Table S2: Top 19 bits of structural fingerprints selected from random forest. Tables S3–S5: True positive (TP), false negative (FN), true negative (TN), and false positive (FP) derived from LDA, CART, and SVM models. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Mailing address: 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA. Phone: (919) 541-3085. Fax: (919) 541-1194. E-mail: Judson.Richard@epa.gov.

Notes

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views of policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project was supported in part by an appointment to the Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA. We would like to express our appreciation to the anonymous reviewers for their constructive comments and suggestions.

■ REFERENCES

(1) Rotroff, D. M.; Dix, D. J.; Houck, K. A.; Knudsen, T. B.; Martin, M. T.; McLaurin, K. W.; Reif, D. M.; Singh, A. V.; Crofton, K. M.; Xia, M.; Huang, R.; Judson, R. S. Using *in vitro* high-throughput screening assays to identify potential endocrine disrupting chemicals. *Environ. Health Perspect.* **2013**, *121* (1), 7–14.

(2) Reif, D. M.; Martin, M. T.; Tan, S.; Houck, K. A.; Judson, R. S.; Richard, A. M.; Knudsen, T. B.; Dix, D. J.; Kavlock, R. J. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. *Environ. Health Perspect.* **2010**, *118* (12), 1714–1720.

(3) Soto, A. M.; Sonnenschein, C. Environmental causes of cancer: endocrine disruptors as carcinogens. *Nat. Rev. Endocrinol.* **2010**, *6* (7), 363–370.

(4) Mahoney, M. M.; Padmanabhan, V. Developmental programming: impact of fetal exposure to endocrine-disrupting chemicals on gonadotropin-releasing hormone and estrogen receptor RNA in sheep hypothalamus. *Toxicol. Appl. Pharmacol.* **2010**, *247* (2), 98–104.

(5) Birnbaum, L. S.; Fenton, S. E. Cancer and developmental exposure to endocrine disruptors. *Environ. Health Perspect.* **2003**, *111* (4), 389–394.

(6) Judson, R. S.; Martin, M. T.; Egeghy, P. P.; Gangwal, S.; Reif, D. M.; Kothiya, P.; Wolf, M. A.; Cathey, T.; Transue, T. R.; Smith, D.; Vail, J.; Frame, A.; Mosher, S.; Cohen-Hubal, E. A.; Richard, A. M. Aggregating data for computational toxicology applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) system. *Int. J. Mol. Sci.* **2012**, *13* (2), 1805–1831.

(7) Judson, R. S.; Richard, A. M.; Dix, D. J.; Houck, K. A.; Elloumi, F.; Martin, M. T.; Cathey, T.; Transue, T. R.; Spencer, R.; Wolf, M. A. ACToR – Aggregated Computational Toxicology Resource. *Toxicol. Appl. Pharmacol.* **2008**, *233* (1), 7–13.

(8) Knudsen, T. B.; Houck, K. A.; Sipes, N.; Singh, A. V.; Judson, R. S.; Martin, M. T.; Weissman, A.; Kleinstreuer, N.; Mortensen, H. M.; Reif, D. M.; Rabinowitz, J. R.; Setzer, W.; Richard, A. M.; Dix, D. J.; Kavlock, R. J. Activity profiles of 309 ToxCast™ chemicals evaluated across 292 biochemical targets. *Toxicology* **2011**, *282* (1–2), 1–15.

(9) Cohen-Hubal, E. A.; Richard, A. M.; Aylward, L.; Edwards, S. W.; Gallagher, J.; Goldsmith, J. M.; Isukapalli, S.; Tornero-Velez, R.; Weber, E. J.; Kavlock, R. J. Advancing exposure characterization for chemical evaluation and risk assessment. *J. Toxicol. Environ. Health B. Crit. Rev.* **2010**, *13* (2), 299–313.

(10) National Research Council. *Toxicity testing: strategies to determine needs and priorities*; The National Academies Press: Washington, DC, 1984.

(11) Pease, W. *Toxic ignorance: the continuing absence of basic health testing for top-selling chemicals in the United States*; Environmental Defense Fund: Washington, DC, 1997.

(12) Judson, R. S.; Richard, A. M.; Dix, D. J.; Houck, K. A.; Martin, M. T.; Kavlock, R. J.; Dellarco, V.; Henry, T.; Holderman, T.; Sayre, P.; Tan, S.; Carpenter, T.; Smith, E. The toxicity data landscape for

environmental chemicals. *Environ. Health Perspect.* **2009**, *117* (5), 685−695.
(13) Egeghy, P. P.; Judson, R. S.; Gangwal, S.; Mosher, S.; Smith, D.; Vail, J.; Cohen-Hubal, E. A. The exposure data landscape for manufactured chemicals. *Sci. Total Environ.* **2012**, *414* (1), 159−166.
(14) U.S. EPA, Office of Pollution Prevention and Toxics (OPPT) chemical reviews and tools case study. http://www.who.int/ifcs/documents/forums/forum5/precaution/epa_en.pdf (accessed September 4, 2013).
(15) Overview: Office of Pollution Prevention and Toxics laws and programs. http://www.epa.gov/opptintr/pubs/oppt101-032008.pdf (accessed September 4, 2013).
(16) Kavlock, R. J.; Dix, D. J. Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. *J. Toxicol. Environ. Health B. Crit. Rev.* **2010**, *13* (2−4), 197−217.
(17) Wetmore, B. A.; Wambaugh, J. F.; Ferguson, S. S.; Sochaski, M. A.; Rotroff, D. M.; Freeman, K.; Clewell, H. J., III; Dix, D. J.; Andersen, M. E.; Houck, K. A.; Allen, B.; Judson, R. S.; Singh, R.; Kavlock, R. J.; Richard, A. M.; Thomas, R. S. Integration of dosimetry, exposure and high-throughput screening data in chemical toxicity assessment. *Toxicol. Sci.* **2012**, *125* (1), 157−174.
(18) Judson, R. S.; Kavlock, R. J.; Setzer, R. W.; Cohen-Hubal, E. A.; Martin, M. T.; Knudsen, T. B.; Houck, K. A.; Thomas, R. S.; Wetmore, B. A.; Dix, D. J. Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. *Chem. Res. Toxicol.* **2011**, *24* (4), 451−462.
(19) Martin, M. T.; Dix, D. J.; Judson, R. S.; Kavlock, R. J.; Reif, D. M.; Richard, A. M.; Rotroff, D. M.; Romanov, S.; Medvedev, A.; Poltoratskaya, N.; Gambarian, M.; Moeser, M.; Makarov, S. S.; Houck, K. A. Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. *Chem. Res. Toxicol.* **2010**, *23* (3), 578−590.
(20) Judson, R. S.; Houck, K. A.; Kavlock, R. J.; Knudsen, T. B.; Martin, M. T.; Mortensen, H. M.; Reif, D. M.; Rotroff, D. M.; Shah, I. A.; Richard, A. M.; Dix, D. J. *In vitro* screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ. Health Perspect.* **2010**, *118* (4), 485−492.
(21) 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* (1), 5−12.
(22) Judson, R. S.; Elloumi, F.; Setzer, R. W.; Li, Z.; Shah, I. A Comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinf.* **2008**, *9*, 241.
(23) DiMaggio, P. A.; Subramani, A.; Judson, R. S.; Floudas, C. A. A novel framework for predicting *in vivo* toxicities from *in vitro* data using optimal methods for dense and sparse matrix reordering and logistic regression. *Toxicol. Sci.* **2010**, *118* (1), 251−265.
(24) Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **2010**, *29*, 476−488.
(25) Zhang, L.; Sedykh, A.; Tripathi, A.; Zhu, H.; Afantitis, A.; Mouchlis, V. D.; Melagraki, G.; Rusyn, I.; Tropsha, A. Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol. Appl. Pharmacol.* **2013**, *272* (1), 67−76.
(26) Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A. M.; Rusyn, I.; Tropsha, A. Use of *in vitro* HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of *in vivo* toxicity. *Environ. Health Perspect.* **2011**, *119* (3), 364−370.
(27) Zhu, H.; Ye, L.; Richard, A. M.; Golbraikh, A.; Wright, F. A.; Rusyn, I.; Tropsha, A. A novel two-step hierarchical quantitative structure-activity relationship modeling workflow for predicting acute toxicity of chemicals in rodents. *Environ. Health Perspect.* **2009**, *117* (8), 1257−1264.
(28) Tseng, Y. J.; Hopfinger, A. J.; Esposito, E. X. The great descriptor melting pot: mixing descriptors for the common good of QSAR models. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 39−43.
(29) Seal, A.; Passi, A.; Jaleel, U. C. A.; Wild, D. J. In-silico predictive mutagenicity model generation using supervised learning approaches. *J. Cheminf.* **2012**, *4* (1), 10.
(30) Su, B. H.; Shen, M. Y.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. *In silico* binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. *J. Chem. Inf. Model.* **2010**, *50* (7), 1304−1318.
(31) Shen, M. Y.; Su, B. H.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. A comprehensive support vector machine binary hERG classification model based on extensive but biased end point hERG data sets. *Chem. Res. Toxicol.* **2011**, *24* (6), 934−949.
(32) Li, H.; Ung, C. Y.; Yap, C. W.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Chen, Y. Z. Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem. Res. Toxicol.* **2005**, *18* (6), 1071−1080.
(33) Xue, Y.; Li, H.; Ung, C. Y.; Yap, C. W.; Chen, Y. Z. Classification of a diverse set of tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chem. Res. Toxicol.* **2006**, *19* (8), 1030−1039.
(34) Yang, X. G.; Chen, D.; Wang, M.; Xue, Y.; Chen, Y. Z. Prediction of antibacterial compounds by machine learning approaches. *J. Comput. Chem.* **2009**, *30* (8), 1202−1211.
(35) Li, H.; Ung, C. Y.; Yap, C. W.; Xue, Y.; Li, Z. R.; Chen, Y. Z. Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. *J. Mol. Graph. Model.* **2006**, *25* (3), 313−323.
(36) Dejaegher, B.; Dhooghe, L.; Goodarzi, M.; Apers, S.; Pieters, L.; Vander Heyden, Y. Classification models for neocryptolepine derivatives as inhibitors of the β-haematin formation. *Anal. Chim. Acta* **2011**, *705* (1−2), 98−110.
(37) Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P.; Olsen, L.; Jørgensen, F. S. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* **2009**, *37* (3), 658−664.
(38) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection. *J. Chem. Inf. Model.* **2011**, *51* (2), 229−236.
(39) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1630−1638.
(40) Carbon-Mangels, M.; Hutter, M. C. Selecting relevant descriptors for classification by Bayesian estimates: a comparison with decision trees and support vector machines approaches for disparate data sets. *Mol. Inf.* **2011**, *30*, 885−895.
(41) Li, Q.; Wang, Y.; Bryant, S. H. A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics* **2009**, *25* (24), 3310−3316.
(42) Tang, Y.; Zhang, Y. Q.; Chawla, N. V.; Krasser, S. SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst. Man. Cybern. B. Cybern.* **2009**, *39* (1), 281−288.
(43) Chang, C. Y.; Hsu, M. T.; Esposito, E. X.; Tseng, Y. J. Oversampling to overcome overfitting: exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods. *J. Chem. Inf. Model.* **2013**, *53* (4), 958−971.
(44) Chen, J.; Tang, Y. Y.; Fang, B.; Guo, C. In silico prediction of toxic action mechanisms of phenols for imbalanced data with random forest. *J. Mol. Graph. Model.* **2012**, *35*, 21−27.
(45) Khalilia, M.; Chakraborty, S.; Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inform. Decis. Mak.* **2011**, *11*, 51.
(46) Rotroff, D. M.; Martin, M. T.; Dix, D. J.; Houck, K. A.; Knudsen, T. B.; Sipes, N. S.; Reif, D. M.; Xia, M.; Huang, R.; Judson, R. S. Interaction score endocrine testing in the 21st century: using in vitro assays to predict estrogen receptor signaling responses, in preparation.
(47) *MOE (Molecular Operating Environment)*; Chemical Computing Group: Montreal, Quebec, Canada, 2012.
(48) *QikProp*, version 3.2; Schrödinger: New York, USA, 2011.

(49) O'boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(50) Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−1474.
(51) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed August 8, 2012).
(52) Huang, R.; Sakamuru, S.; Martin, M.; Reif, D.; Judson, R.; Houck, K.; Shockley, K.; Fostel, J.; Witt, K.; Tong, W.; Rotroff, D.; Zhao, T.; Shinn, P.; Dix, D.; Kavlock, R.; Tice, R. R.; Simeonov, A.; Austin, C. P.; Xia, M. Profiling of the Tox21 10K compound library for environmental agonists and antagonists of the estrogen receptor signaling pathway, in preparation.
(53) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47* (1), 150−158.
(54) Diaz-Uriarte, R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinf.* **2007**, *8*, 328.
(55) Hao, M.; Li, Y.; Wang, Y.; Zhang, S. A classification study of respiratory syncytial virus (RSV) inhibitors by variable selection with random forest. *Int. J. Mol. Sci.* **2011**, *12* (2), 1259−1280.
(56) Luan, F.; Zhang, R.; Zhao, C.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. Classification of the carcinogenicity of N-nitroso compounds based on support vector machines and linear discriminant analysis. *Chem. Res. Toxicol.* **2005**, *18* (2), 198−203.
(57) Zang, Q.; Keire, D. A.; Wood, R. D.; Buhse, L. F.; Moore, C. M. V.; Nasr, M.; Al-Hakim, A.; Trehy, M. L.; Welsh, W. J. Combining ^{1}H NMR spectroscopy and chemometrics to identify heparin samples that may possess DS impurity or OSCS contaminant. *J. Pharm. Biomed. Anal.* **2011**, *54*, 1020−1029.
(58) Varmuza, K.; Filzmoser, P. *Introduction to multivariate statistical analysis in chemometrics*; CRC Press: Boca Raton, FL, USA, 2009.
(59) Zang, Q.; Keire, D. A.; Buhse, L. F.; Wood, R. D.; Mital, D. P.; Haque, S.; Srinivasan, S.; Moore, C. M. V.; Nasr, M.; Al-Hakim, A.; Trehy, M. L.; Welsh, W. J. Identification of heparin samples that contain impurities or contaminants by chemometric pattern recognition analysis of proton NMR spectral data. *Anal. Bioanal. Chem.* **2011**, *401*, 939−955.
(60) Zang, Q.; Keire, D. A.; Wood, R. D.; Buhse, L. F.; Moore, C. M. V.; Nasr, M.; Al-Hakim, A.; Trehy, M. L.; Welsh, W. J. Determination of galactosamine impurities in heparin samples by multivariate regression analysis of their ^{1}H NMR spectra. *Anal. Bioanal. Chem.* **2011**, *399*, 635−649.
(61) Eitrich, T.; Kless, A.; Druska, C.; Meyer, W.; Grotendorst, J. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J. Chem. Inf. Model.* **2007**, *47*, 92−103.
(62) Zhang, L.; Fourches, D.; Sedykh, A.; Zhu, H.; Golbraikh, A.; Ekins, S.; Clark, J.; Connelly, M. C.; Sigal, M.; Hodges, D.; Guiguemde, A.; Guy, R. K.; Tropsha, A. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *J. Chem. Inf. Model.* **2013**, *53*, 475−492.
(63) Zang, Q.; Keire, D. A.; Wood, R. D.; Buhse, L. F.; Moore, C. M. V.; Nasr, M.; Al-Hakim, A.; Trehy, M. L.; Welsh, W. J. Class modeling analysis of heparin ^{1}H NMR spectral data using the soft independent modeling of class analogy and unequal class modeling techniques. *Anal. Chem.* **2011**, *83*, 1030−1039.
(64) Hao, M.; Li, Y.; Wang, Y.; Zhang, S. A classification study of human β_3-adrenergic receptor agonists using BCUT descriptors. *Mol. Divers.* **2011**, *15*, 877−887.
(65) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21* (20), 3940−3941.
(66) R Development Core Team *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2011; http://www.R-project.org/ (accessed September 4, 2013).
(67) Hong, H.; Tong, W.; Fang, H.; Shi, L.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J. D.; Branham, W.; Sheehan, D. M. Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environ. Health Perspect.* **2002**, *110* (1), 29−36.
(68) Li, J.; Gramatica, P. QSAR classification of estrogen receptor binders and pre-screening of potential pleiotropic EDCs. *SAR QSAR Environ. Res.* **2010**, *21* (7−8), 657−669.