Absorption, Distribution, Metabolism, Excretion, and Toxicity Evaluation in Drug Discovery. 14. Prediction of Human Pregnane X Receptor Activators by Using Naive Bayesian Classific...

ARTICLE *in* CHEMICAL RESEARCH IN TOXICOLOGY · DECEMBER 2014

Impact Factor: 3.53 · DOI: 10.1021/tx500389q · Source: PubMed

7 AUTHORS, INCLUDING:

Sheng Tian

Soochow University (PRC)

21 PUBLICATIONS **200** CITATIONS

SEE PROFILE

xuechu Zhen

Soochow University (PRC)

117 PUBLICATIONS **2,000** CITATIONS

SEE PROFILE

Tingjun Hou

Zhejiang University

241 PUBLICATIONS **4,626** CITATIONS

SEE PROFILE

Chemical Research in Toxicology

Article

pubs.acs.org/crt

Absorption, Distribution, Metabolism, Excretion, and Toxicity Evaluation in Drug Discovery. 14. Prediction of Human Pregnane X Receptor Activators by Using Naive Bayesian Classification Technique

Huali Shi,[†,⊥] Sheng Tian,[∥,⊥] Youyong Li,[†] Dan Li,[‡] Huidong Yu,*[,§] Xuechu Zhen,*[,∥] and Tingjun Hou*[,†,‡]

[†]Institute of Functional Nano and Soft Materials (FUNSOM), Soochow University, Suzhou, Jiangsu 215123, People's Republic of China

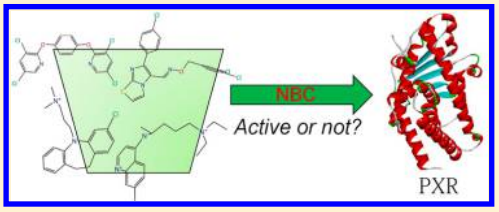
[‡]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, People's Republic of China

[§]Crystal Pharmatech Inc., 707 Alexander Road, Building 2, Suite 208, Princeton, New Jersey 08540, United States

[∥]College of Pharmaceutical Sciences, Soochow University, Suzhou, Jiangsu 215123, People's Republic of China

Supporting Information

ABSTRACT: The activation of pregnane X receptor (PXR), a member of the nuclear receptor (NR) superfamily, can mediate potential drug−drug interactions, and therefore, prediction of PXR activation is of great importance for evaluating drug metabolism and toxicity. In this study, based on 532 structurally diverse compounds, we present a comprehensive analysis with the aim to build accurate classification models for distinguishing PXR activators from nonactivators by using a naive Bayesian classification technique. First, the distributions of eight important molecular physicochemical properties of PXR activators versus nonactivators were compared, illustrating that the hydrophobicity-related molecular descriptors (Alog*P* and log *D*) show slightly better capability to discriminate PXR activators from nonactivators than the others. Then, based on molecular physicochemical properties, *VolSurf* descriptors, and molecular fingerprints, naive Bayesian classifiers were developed to separate PXR activators from nonactivators. The results demonstrate that the introduction of molecular fingerprints is quite essential to enhance the prediction accuracy of the classifiers. The best Bayesian classifier based on the 21 physicochemical properties, *VolSurf* descriptors, and LCFC_10 fingerprints descriptors yields a prediction accuracy of 92.7% for the training set based on leave-one-out (LOO) cross-validation and of 85.2% for the test set. Moreover, by exploring the important structural fragments derived from the best Bayesian classifier, we observed that flexibility is an important structural pattern for PXR activation. In addition, chemical compounds containing more halogen atoms, unsaturated alkanes chains relevant to $\pi-\pi$ stacking, and fewer nitrogen atoms tend to be PXR activators. We believe that the naive Bayesian classifier can be used as a reliable virtual screening tool to predict PXR activation in the drug design and discovery pipeline.



INTRODUCTION

The pregnane X receptor (PXR), which is also known as the steroid and xenobiotic sensing nuclear receptor (SXR), is a promiscuous protein encoded by the NR1I2 (nuclear receptor subfamily 1, group I, member 2) gene. It can be activated by a variety of structurally diverse endogenous and exogenous compounds, including bile acids, steroid hormones, prescription drugs, pesticides, and environmental contaminants.[1,2] PXR regulates the expression of many enzymes involved in metabolism and important drug transporters, such as cytochrome P450 enzymes, glutathione-S-transferases, P-glycoprotein, and multidrug resistance (MDR) proteins.[3−7] The metabolism of xenobiotics and drug−drug interactions are related to the activation of PXR, and drugs or drug candidates that can activate PXR may have substantial impact on drug metabolism, transportation, and drug−drug interactions. Therefore, identification of PXR activators will be helpful to analyze the pharmacokinetic profiles of drug candidates and to detect potential drug−drug interactions in the early stage of drug discovery.

Numerous cell-based assays and enzyme-based assays have been employed to measure the activation of PXR, such as luciferase-based reporter assay,[8] scintillation proximity competition binding assay,[4,9−11] chloramphenicol acetyltransferase (CAT) assay,[4,8−10] and bile acid hydroxylase assay.[12,13] Moreover, some high-throughput screening assays (transactivation and immortalized hepatocyte assays) were also used to identify the activators of PXR.[4,10,14,15] However, these experimental assays are expensive, resource-intensive, and time-consuming. Therefore, great effort has been dedicated to develop computational approaches or models to predict PXR

Received: September 24, 2014

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

activation.[11,16−22] Compared with experimental testing, computational models provide more rapid and efficient ways to recognize the activators of PXR.

The computational approaches for predicting PXR activation activities can be roughly classified into two categories: ligand-based and structure-based. Based on the structurally diverse small chemical compounds with known PXR activation activities, pharmacophore modeling and quantitative structure−activity relationship (QSAR) techniques have been used to develop the ligand-based prediction models of PXR activation activities.[2,11,19,23−27] For example, in 2005, Schuster and Langer[26] identified structural features essential for PXR activation by employing ligand-based pharmacophore modeling. Their results revealed that most highly active PXR ligands share five hydrophobic features and two H-bond acceptors. The five hydrophobic features occupy large hydrophobic area of the binding pocket, and the two H-bond acceptors may form H-bonds with Gln285 and His407, which are quite necessary for PXR activation.[26] In addition, a number of QSAR models for predicting PXR activation have been developed based on sophisticated machine learning approaches, including recursive partitioning (RP), artificial neural network (ANN), support vector machine (SVM), and naive Bayesian classification (NBC).[2,19,23−25,27] For example, in 2007, Ung and co-workers[17] utilized three machine learning methods (SVM, k-nearest neighbor, and probabilistic ANN) to establish the classification models of PXR activation based on 205 diverse molecules, and their predictions indicated that the best classifier can correctly predict 85.0% of PXR activators and 73.6% of PXR nonactivators. In 2011, based on a training set of 177 PXR activators/nonactivators, Pan and co-workers built PXR activation classification models using the NBC technique.[21] For the test set consisting of 145 molecules, the specificity and global accuracy (GA) of the best naive Bayesian classifier can achieve 92.0% and 69.0%, respectively.[21] There are 13 crystal structures of human PXR available in the Protein Data Bank (PDB) database,[28] which provide good starting points to understand the ligand−PXR interactions by using structure-based modeling approaches, such as molecular docking. For example, in 2009, Ekins and co-workers[19] docked 30 activators and 89 nonactivators into six PXR crystal structures, but unfortunately, they found that the predictions from molecular docking could not discriminate PXR activators from non-activators effectively. The best docking prediction can only achieve a prediction accuracy of 66.7% for the activators and 55.0% for the nonactivators.[19] The failure of molecular docking to distinguish PXR activators from nonactivators may be explained by the intrinsic features of the PXR ligand-binding domains (LBD), which is quite large, flexible, and hydrophobic.[29] It seems that it is still difficult to use the PXR crystal structure for prospective molecular docking studies.

In summary, the studies for the *in silico* prediction of PXR activation are still limited.[5,8,11,17,19,30−34] Moreover, most PXR classification models were developed based on relatively small data sets, and the statistical reliabilities of these models may not be guaranteed, which may restrict their applications for much larger compound data sets. Therefore, in this study, we collected an extensive data set of 532 chemical compounds with known PXR activation activities. Subsequently, based on the extensive data set, the distributions of eight important physicochemical properties for the PXR activators and nonactivators were examined. Then, the NBC technique was employed to build the classification models based on 21 physicochemical properties, *VolSurf* descriptors and molecular fingerprints. Finally, the important structural features favorable or unfavorable for PXR activation were highlighted and discussed.

MATERIALS AND METHODS

Data Set of PXR Activators and Nonactivators. All of the chemical compounds with PXR activation activities were collected from the literature.[8,11,17,19,31−35] The first important source is the previously reported compilations with quantitative EC_{50} values, such as Kortagere's data set,[19] Ung's data set,[17] Khandelwal's data set,[18] Ekins' data set,[8,11,19,33] and so forth.[31,32,34] The data set compiled by Kortagere et al.[19] has 164 PXR activators and 134 nonactivators. In Kortagere's data set, the compounds with EC_{50} < 100 μM were classified as PXR activators and those with $EC_{50} \geq 100$ μM as nonactivators. The data set reported by Ung et al. has 98 activators and 77 nonactivators, and that reported by Khandelwal et al. has 82 activators and 63 nonactivators.[17,18] The classification thresholds for these two data sets are the same as that used by Kortagere et al.[19] Moreover, the same threshold (EC_{50} <100 μM or ≥100 μM) was also used by Ekins et al.[19] to classify 65 PXR activators and 58 nonactivators. Therefore, the classification threshold of EC_{50} <100 μM or ≥100 μM was employed here to assign each compound to be active or nonactive. Another important source is the data set with 350 compounds without EC_{50} values reported by Matter et al.[35] The classifications for those 350 molecules without EC_{50} values were determined by examining the experimental PXR activation, which is generally in close accord with the classification threshold of EC_{50} <100 μM or ≥100 μM. Finally, the consistency of each compound was checked. If the classifications for the same compound from different literature sources are not consistent, this compound was not included in our data set. The canonical SMILES string of each molecule was compared with those of the other molecules iteratively to eliminate all duplicates, and the nonduplicated entries were merged after careful analysis. Compared with the previous studies, we believe that the computational models for predicting PXR activities are more reasonable based on the compiled PXR data set with a unified threshold based on EC_{50} <100 μM or ≥100 μM.

The final data set used in this study contains 340 activators and 192 nonactivators with diverse structures. Each molecule in the data set was optimized by using molecular mechanics (MM) with the MMFF94 force field[36] in SYBYL-X simulation package (version 1.1).[37] The data set was split into a training set with 478 molecules for generating the naive Bayesian classifiers and a test set with 54 molecules, which combines a test set for the selection of variables and an external set for evaluating the classification performance used in Matter's study,[35] for estimating the classification performance of the models.

Calculation of Molecular Descriptors and Molecular Fingerprints. Here, 21 physicochemical properties widely used in ADME (absorption, distribution, metabolism, and excretion) predictions and QSAR modeling were chosen as molecular descriptors in our analysis (Table S1 in Supporting Information). These 21 descriptors include octanol−water partitioning coefficient (Alog*P*) based on Ghose and Crippen's method,[38] the apparent partition coefficient at pH = 7.4 (log $D_{7.4}$) based on Csizmadia's method, molecular solubility (log *S*) based on Tetko's multiple linear regression model,[39] molecular weight (MW), the number of hydrogen bond donors (N_{HBD}), the number of hydrogen bond acceptors (N_{HBA}), the number of rotatable bonds (N_{rot}), the polar surface area (PSA), the molecular fractional polar surface area (MFPSA), the molecular polar surface area (MPSA), the solvent accessible surface area (SASA), the sum of oxygen and nitrogen atoms (n_{O+N}), the number of aromatic bonds (n_{AB}), the number of aromatic rings (n_{AR}), the number of rings (n_R), the number of three rings (n_{R3}), the number of four rings (n_{R4}), the number of five rings (n_{R5}), the number of six rings (n_{R6}), the number of seven rings (n_{R7}), the number of eight rings (n_{R8}), and the number of nine rings and more (n_{R9+}). All these molecular descriptors were calculated by using the Discovery Studio (DS) molecular simulation package

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

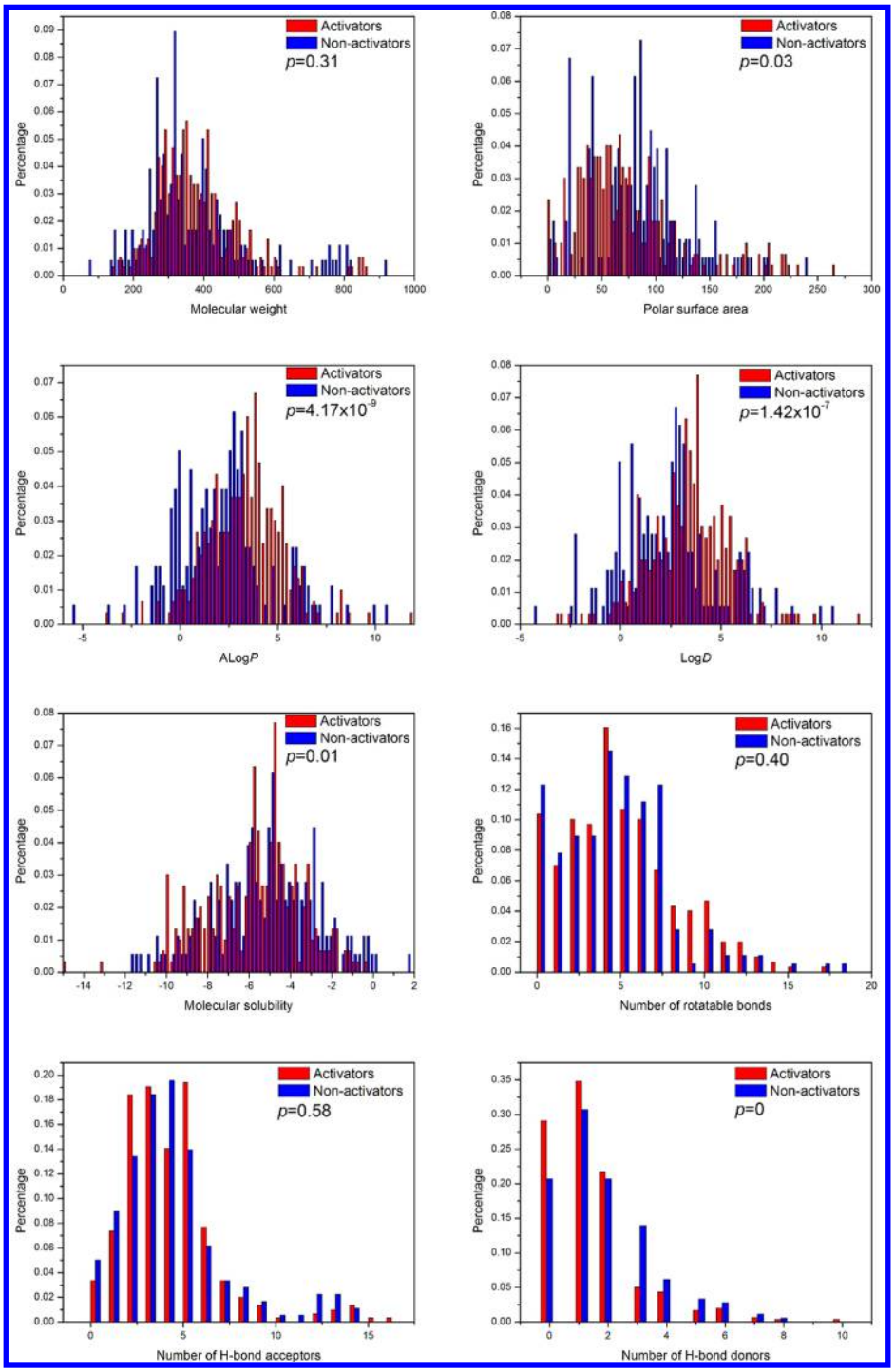


Figure 1. Distributions of eight molecular physicochemical properties, including AlogP, log *D*, log *S*, MW, N_{HBD}, N_{HBA}, N_{rot}, and PSA for the PXR activators and nonactivators.

(version 3.1).[40] In addition, a set of 76 *VolSurf* descriptors[41] calculated in the MOE simulation package were incorporated,[42] and they have been shown to be useful in the prediction of pharmacokinetic properties.[18,43]

In addition, the molecular fingerprints, which are used to characterize the substructural features of a molecule, were also used in this study. The SciTegic extended-connectivity fingerprints (FCFP, ECFP, and LCFP) and Daylight-style path-based fingerprints (FPFP, EPFP, and LPFP) were employed. All those fingerprints have been discussed in detail in the previous studies.[44,45] Each fingerprint class consists of four letters and one connected number, which is the maximum diameter (in bond lengths) of the largest structure represented by the fingerprint or the maximum length of the path. For each fingerprint class, four different maximum diameters (4, 6, 8, and 10) were considered in our analysis. All these molecular fingerprints were generated automatically using DS 3.1.[40]

Naive Bayesian Classifiers of PXR Activators and Non-activators. Here, the NBC technique was employed to build the classification models for distinguishing PXR activators from non-activators. Compared with other machine learning approaches, NBC can handle much data simultaneously, can learn fast, and is tolerant of random noise during the model building process. In addition, NBC just needs a small training set to evaluate the necessary parameters (means and variances of the variables) for classification. In this study,

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

Table 1. Performance of the Naive Bayesian Classifiers for the Training Set Based on Different Molecular Descriptors[a]

descriptors	TP	FN	FP	TN	SE	SP	Q_+	Q_-	GA	C	AUC
MP[b]	197	102	39	140	0.659	0.782	0.835	0.579	0.705	0.427	0.701
V[c]	196	103	33	146	0.656	0.816	0.856	0.586	0.715	0.456	0.645
MP+V	184	115	19	160	0.615	0.894	0.906	0.582	0.720	0.499	0.680
MP+V+ECFC_10	272	27	10	169	0.910	0.944	0.965	0.862	0.923	0.840	0.813
MP+V+ECFP_10	268	31	7	172	0.896	0.961	0.975	0.847	0.921	0.839	0.807
MP+V+EPFC_10	255	44	33	146	0.853	0.816	0.885	0.768	0.839	0.661	0.745
MP+V+EPFP_10	250	49	19	160	0.836	0.894	0.929	0.766	0.858	0.712	0.762
MP+V+FCFC_10	260	39	7	172	0.870	0.961	0.974	0.815	0.904	0.809	0.811
MP+V+FCFP_10	261	38	9	170	0.873	0.950	0.967	0.817	0.902	0.803	0.803
MP+V+FPFC_10	250	49	25	154	0.836	0.860	0.909	0.759	0.845	0.682	0.735
MP+V+FPFP_10	242	57	18	161	0.809	0.899	0.931	0.739	0.843	0.689	0.745
MP+V+LCFC_10	271	28	7	172	0.906	0.961	0.975	0.860	0.927	0.851	0.814
MP+V+LCFP_10	272	27	10	169	0.910	0.944	0.965	0.862	0.923	0.840	0.809
MP+V+LPFC_10	255	44	10	169	0.853	0.944	0.962	0.793	0.887	0.776	0.762
MP+V+LPFP_10	254	45	10	169	0.849	0.944	0.962	0.790	0.885	0.772	0.761

[a]Fingerprints based on the maximum diameter of 10. [b]MP represents 21 molecular physicochemical properties. [c]V represents *VolSurf* descriptors.

each compound in the PXR data set was categorized into the PXR active (+) or nonactive (−) class, and a vector $\mathbf{f} = \langle f_1, f_2, ..., f_n \rangle$ is the calculated values for the n feature variables F_1, F_1, ..., F_n (physicochemical properties, *VolSurf* descriptors, and molecular fingerprints). Then, based on Bayes's theorem, we get

$$p(C|F_1, F_2, ..., F_n) = \frac{p(C)p(F_1, ...F_n|C)}{p(F_1, ..., F_n)} \quad (1)$$

In eq 1, C refers to a compound's class, $p(C|F_1,F_2,\cdots,F_n)$ is the posterior probability of the compound class, $p(C)$ is the prior probability, a probability induced from the training set, $p(F_1,\cdots F_n|C)$ is the probability that a compound has certain descriptors given that it is a PXR activator or nonactivator, and $p(F_1,\cdots,F_n)$ is the marginal probability that the given molecular descriptors will occur in the data set. The three probabilities on the right side of eq 1 can be learned from a training set consisting of a number of PXR activators and nonactivators. The mathematical procedure to train a naive Bayesian classifier was described previously.[17,22,46−52] All of the naive Bayesian classifiers were developed in DS 3.1.[40]

Validation of the Prediction Accuracies of the Naive Bayesian Classifiers. Based on the predictions of each naive Bayesian classifier, the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were counted. The predictive performance of each classifier was assessed by sensitivity, SE = TP/(TP + FN), specificity, SP = TN/(TN + FP), prediction accuracy of PXR activators, Q_+ = TP/(TP + FP), prediction accuracy of PXR nonactivators, Q_- = TN/(TN + FN), global accuracy, GA = (TP + TN)/(TP + TN + FP + FN), and Matthews correlation coefficient,

$$C = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}}$$

Matthews correlation coefficient C ranges from 0 to 1 and a perfect classification gives a correlation coefficient of 1. In addition, the classification capacity was measured by the area under the receive operating characteristic (ROC) curve (AUC), which is a graphical plot to illustrate the performance of a binary classifier by changing its discrimination threshold.

RESULTS AND DISCUSSIONS

Property Distributions for PXR Activators and Nonactivators. To understand the relationships between essential molecular physicochemical properties and PXR activation, the distributions of eight molecular properties for the PXR active and nonactive classes are depicted in Figure 1. These eight molecular physicochemical properties, including AlogP, log $D_{7.4}$, log S, MW, N_{HBD}, N_{HBA}, N_{rot}, and PSA, have been widely used in ADME predictions.[53−56]

The significance of difference between the means of the PXR activators and nonactivators for each physicochemical property was evaluated by Student's t-test. As shown in Figure 1, it is obvious that any single physicochemical property cannot separate the PXR activators from nonactivators effectively, indicated by the quite large P values. For example, molecular weight is distributed between 140 and 860 with a mean of 320 for the activators and between 80 and 920 with a mean of 350 for the nonactivators, respectively. The P value associated with the difference between the means of the two MW distributions for the PXR activators and nonactivators is 0.31, indicating that molecular weight has no any prediction capacity to distinguish the PXR activators from nonactivators.

For the eight studied molecular properties, three of them are related to the hydrophobicity of a molecule: log P, log $D_{7.4}$ and log S. As can be seen from Figure 1, log P is distributed between −3.75 and 12.00 with a mean of 3.00 for the active class and between −5.60 and 10.60 with a mean of 1.95 for the nonactive class. For log $D_{7.4}$, the active class is distributed between −3.30 and 11.75 with a mean of 3.45, and the nonactive class is distributed from −4.20 to 10.50 with a mean of 2.55; log S is distributed from −15.00 to −0.30 with a mean of −6.05 for the active class and from −11.75 to 1.80 with a mean of −6.05 for the nonactive class. The mean values of log P for the 179 nonactivators and 299 activators are 1.95 and 3.00, respectively. This means that PXR activators tend to be more hydrophobic than nonactivators. The PXR activators with higher hydrophobicity may form stronger hydrophobic interactions with the hydrophobic binding region of PXR. The P values for log P, log $D_{7.4}$ and log S are 4.17×10^{-9}, 1.42×10^{-7}, and 0.01, respectively. Our findings are consistent with the previous studies that increased hydrophobicity is usually favorable for the binding of a small molecule to PXR.[8,11,17,18,26,33]

For the descriptor N_{rot} and the other three descriptors, PSA, N_{HBD}, and N_{HBA}, which represent the electrostatic or H-bond features, similar phenomena can also be observed. The P values for N_{rot}, PSA, N_{HBD}, and N_{HDA} are 0.40, 0.03, 0, and 0.58, respectively. Overall, because the LBD of PXR is quite flexible, numerous diverse chemical compounds can bind to PXR. We cannot get acceptable discriminating capacity for separating the

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

Table 2. Performance of the Naive Bayesian Classifiers for the Test Set Based on Different Molecular Descriptors[a]

descriptors	TP	FN	FP	TN	SE	SP	Q_+	Q_-	GA	C	AUC
MP[b]	29	12	5	8	0.707	0.615	0.853	0.400	0.685	0.286	0.702
V[c]	25	16	5	8	0.610	0.615	0.833	0.333	0.611	0.194	0.638
MP+V	28	13	7	6	0.683	0.462	0.800	0.316	0.630	0.129	0.662
MP+V+ECFC_10	37	4	4	9	0.902	0.692	0.902	0.692	0.852	0.595	0.876
MP+V+ECFP_10	37	4	4	9	0.902	0.692	0.902	0.692	0.852	0.595	0.861
MP+V+EPFC_10	35	6	5	8	0.854	0.615	0.875	0.571	0.796	0.458	0.762
MP+V+EPFP_10	38	3	5	8	0.927	0.615	0.884	0.727	0.852	0.576	0.728
MP+V+FCFC_10	37	4	4	9	0.902	0.692	0.902	0.692	0.852	0.595	0.882
MP+V+FCFP_10	36	5	4	9	0.878	0.692	0.900	0.643	0.833	0.556	0.865
MP+V+FPFC_10	35	6	5	8	0.854	0.615	0.875	0.571	0.796	0.458	0.735
MP+V+FPFP_10	36	5	5	8	0.878	0.615	0.878	0.615	0.815	0.493	0.720
MP+V+LCFC_10	37	4	4	9	0.902	0.692	0.902	0.692	0.852	0.595	0.882
MP+V+LCFP_10	36	5	4	9	0.878	0.692	0.900	0.643	0.833	0.556	0.865
MP+V+LPFC_10	36	5	4	9	0.878	0.692	0.900	0.643	0.833	0.556	0.797
MP+V+LPFP_10	35	6	4	9	0.854	0.692	0.897	0.600	0.815	0.521	0.794

[a]Fingerprints based on the maximum diameter of 10. [b]MP represents 21 molecular physicochemical properties. [c]V represents *VolSurf* descriptors.

activators from nonactivators by using any single physicochemical property.

Naive Bayesian Classifiers for PXR Activators and Nonactivators. As shown in the preceding section, any single molecular property cannot be used as a reliable descriptor to separate the PXR activators from nonactivators. Thus, NBC, a more robust machine learning approach, was employed to develop the classification models to distinguish the PXR activators from nonactivators, based on the 21 molecular physicochemical properties (Table S1 in Supporting Information), *VolSurf* descriptors,[18,43] and molecular fingerprints.

In total, 51 Bayesian classifiers were generated based on different combinations of molecular descriptors. The statistical results for these classifiers, which were obtained using the leave-one-out (LOO) cross-validation procedure, are summarized in Table 1 and Table S2 in Supporting Information. Apparently, only based on the 21 molecular properties, *VolSurf* descriptors, or the combination of both, the Bayesian classifiers cannot achieve satisfactory discrimination capacities, indicated by the quite low *C* values (~0.50). Then, by combination of any class of molecular fingerprints with the molecular properties and *VolSurf* descriptors, 48 Bayesian classifiers were generated. As shown in Table 1 and Table S2 in Supporting Information, the addition of the molecular fingerprints can improve the prediction accuracies significantly.

The best classifier for the training set based on the molecular properties, *VolSurf* descriptors, and LCFC_10 fingerprints has a sensitivity of 0.906, a specificity of 0.961, a prediction accuracy of 0.975 for the active class, a prediction accuracy of 0.860 for the nonactive class, and a global accuracy of 0.927. Then, the true prediction powers of the Bayesian classifiers were validated by the external test set (Table 2 and Table S3 in Supporting Information). Compared with the classifiers solely based on the molecular properties, *VolSurf* descriptors, or combination of both, the classifiers based on the molecular properties, *VolSurf* descriptors, and each class of fingerprints separate the two classes substantially better. The test set predicted by using the best Bayesian classifier based on the 21 physicochemical properties, *VolSurf* descriptors, and LCFC_10 fingerprints can achieve a sensitivity of 0.902, a specificity of 0.692, a prediction accuracy of 0.902 for the activators, a prediction accuracy of 0.692 for the nonactivators, and a global accuracy of 0.852. The quality of the best Bayesian classifier was further assessed by a ROC curve, and the AUCs for the training set and test set are 0.814 and 0.882, respectively, indicating that the prediction capacity of the best Bayesian classifier is quite reliable (Figure 2).

Furthermore, we also investigated the influences of the unbalanced nature of the PXR activators versus nonactivators in the training set. A total of 179 diverse PXR activators were chosen using the *Find Diverse Molecules* protocol in DS 3.1 based on the structural similarities of the FCFP_4 fingerprints.[40] Then, the classifiers were developed based on the balanced training set with 179 PXR activators and 179 nonactivators. As shown in Tables S4 and S5 in Supporting Information, the best classifier was developed based on the combination of the physicochemical properties, *VolSurf* descriptors, and LCFC_10 fingerprints, indicating that the performance of the Bayesian classifiers is not very sensitive to the unbalanced nature of the training set. Similar finding was also observed in our previous study.[51]

The two bimodal histograms of the training and test sets of the Bayesian scores given by the best Bayesian classifier are shown in Figure 3. Both of these histograms of the training and test sets illustrate that the nonactivators tend to have more negative Bayesian scores, while the activators have more positive values. Compared with the classification capabilities of the eight molecular properties shown in Figure 1, the best Bayesian classifier separates the two classes much better. The overlapped area between the Bayesian scores of the PXR activators and nonactivators is much smaller than those shown in Figure 1. For the training set, the activators and nonactivators have overlaps between −40 and 10, which can be named as the "uncertain area". The predictions in the uncertain area may be not very reliable.

Analysis of Important Fragments Given by Naive Bayesian Classifier. According to the predictions of the best Bayesian classifier, the relatively important structural features characterized by the fingerprints can be identified. The top 20 good and 20 bad fragments ranked by the Bayesian scores are shown in Figure 4. These fragments may provide useful information to experimental scientists for demystifying PXR activation.

As shown in Figure 4a, these 20 structural fragments are quite favorable for the likelihood of a PXR activator. It seems that almost all of these favorable fragments are essentially

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

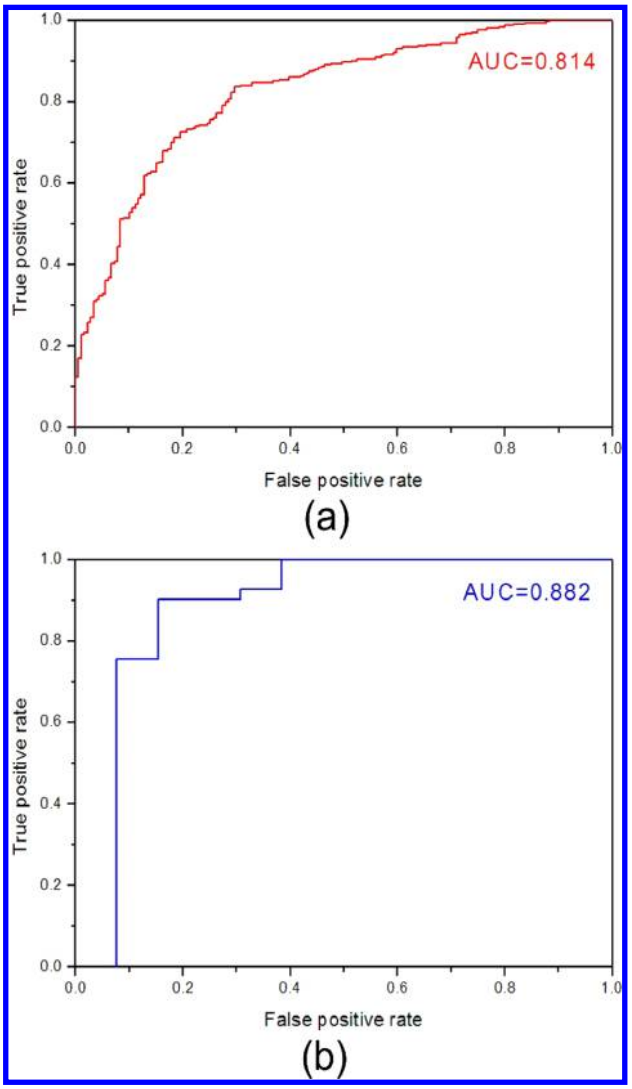


Figure 2. ROC curves of the best Bayesian classifier for (a) training and (b) test sets.

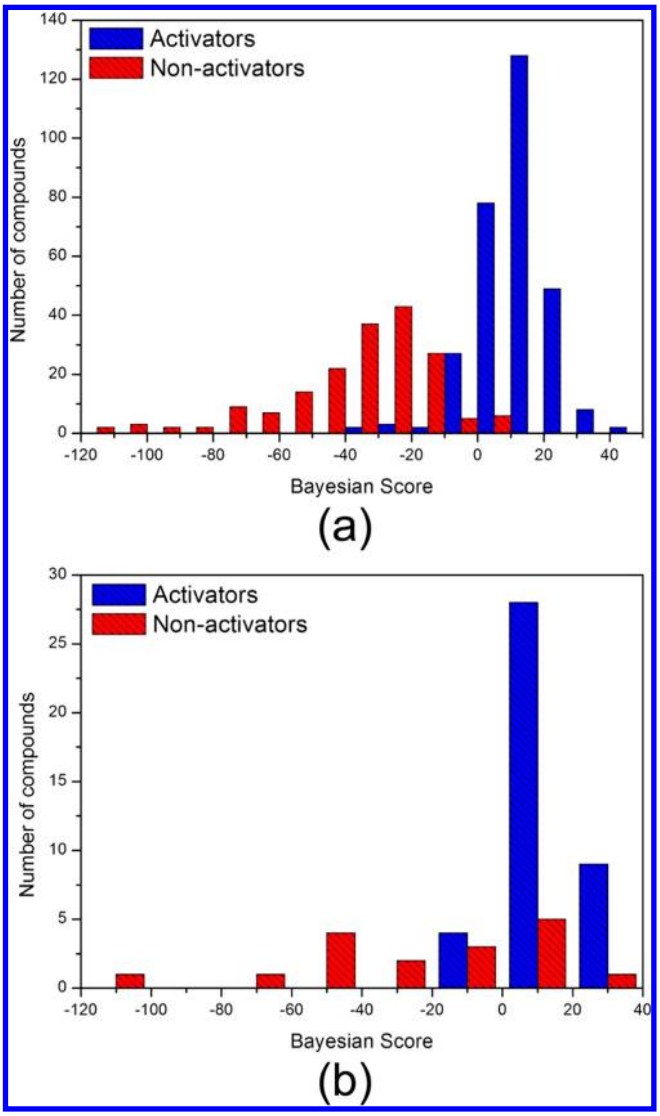


Figure 3. Distributions of the Bayesian scores predicted by the best Bayesian classifier based on the 21 molecular properties, *VolSurf* descriptors, and LCFC_10 fingerprints for the active and nonactive classes of the training set and test set. The Bayesian scores for the training set were obtained by using the leave-one-out (LOO) cross-validation procedure.

hydrophobic. It has been proven that the LBD of PXR contains extensive hydrophobic regions.[11,17] Thus, possession of these hydrophobic features appears to be favorable for the binding of a molecule to PXR. Moreover, it can be observed that there are more halogen atoms, especially chlorine atoms, in the favorable fragments than in the unfavorable fragments, as shown in Figure 4a (fragments 1, 16, and 20) and Figure 4b (fragment 14). Moreover, there are less nitrogen atoms in the favorable fragments than in the unfavorable fragments, as shown in Figure 4a (fragments 7, 8, 12, and 18) and Figure 4b (fragments 2, 3, 6, 9, 10, 16, 17, and 19), which has been reported in the previous study.[17] In addition, five unsaturated alkane chains are relevant to $\pi-\pi$ stacking, including fragments 4, 7, 14, 17, and 19 (Figure 4a), indicating that the existence of these fragments is favorable for PXR activation.[17]

By analyzing the structural fragments with negative contributions to PXR activation (Figure 4b), we can find that the unfavorable structural fragments seem to be more polar, that is, containing hydroxyl (fragment 15) or other polar groups. More polar substructures may be unfavorable to PXR activation. Then, we also observed that there are more amine fragments in the unfavorable fragments. These nitrogen atoms attached to unsaturated carbons or aromatic rings may limit the flexibility of compounds. Too rigid chemical compounds have difficulty passing the long and narrow cave of the LBD of PXR to form a stable interaction with PXR. Other important fragments with the negative contributions to PXR activation are aromatic rings, such as fragments 2, 6, 7, 11, 12, 14, and 16−20, and they may also limit the flexibility of chemical compounds.

Analysis of the Misclassified Compounds. As mentioned above, the best Bayesian classifier has good prediction capability for the test set. However, eight molecules in the test set still cannot be predicted correctly. Four out of the 41 PXR activators in the test set were predicted as the false negatives, and 4 out of the 13 PXR nonactivators in the test set as the false positives. Here, we listed four possible reasons for the misclassifications. First, the intrinsic limitation of the Bayesian classifiers based on the fingerprints may lead to these misclassifications. By analyzing the structures of these four misclassified nonactivators (Gao004, Gao010, Gao009, and Bra007g in Figure 5a), we observed that all of them contain one

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

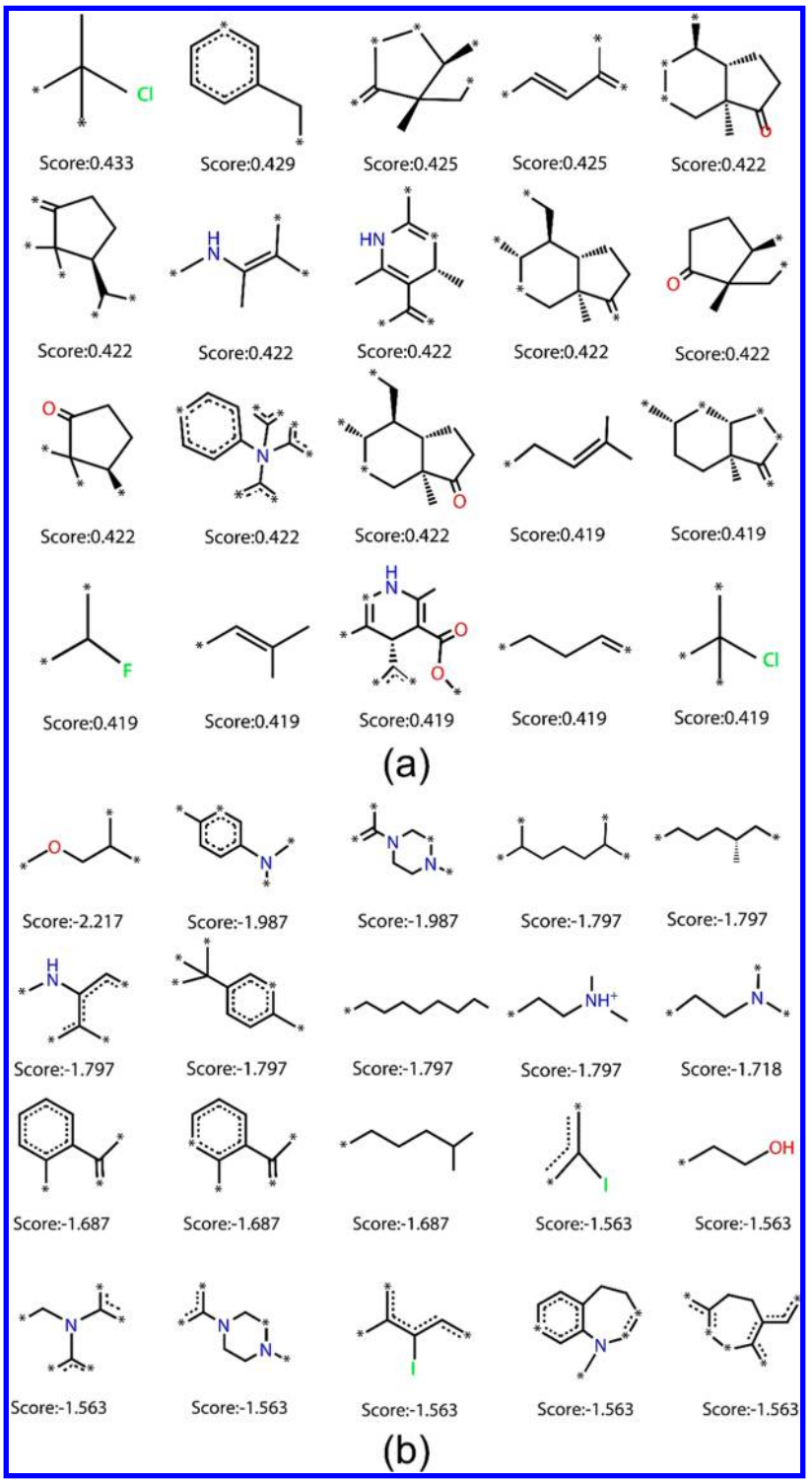


Figure 4. (a) The 20 good and (b) 20 bad structural fragments identified by the best naive Bayesian classifier.

or more favorable fragments (e.g., fragments 2, 4, 16, and 17). The four misclassified nonactivators comprising the favorable fragments tend to be predicted as activators. Similarly, the PXR activators with unfavorable fragments also tend to be incorrectly predicted as nonactivators (Figure 5b). We should know that the molecules with favorable fragments are more likely to be the activators of PXR and *vice versa*. However, the existence of these good or bad fragments in a molecule does not mean that it must be an activator or a nonactivator of PXR. Second, the structural information derived from molecular fingerprints may not cover sufficient structural patterns necessary for the binding of molecules to PXR. It is quite possible that some important fragments essential for PXR activation are not covered by the compounds in the training set. Thus, the compounds comprising these unexplored fragments in the test set cannot be predicted correctly. Although in this study a much larger PXR data set was employed for model development, more efforts are still needed to collect larger data sets of PXR activators and nonactivators. Third, as shown in Figure 3a, the Bayesian scores of the activators and non-

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

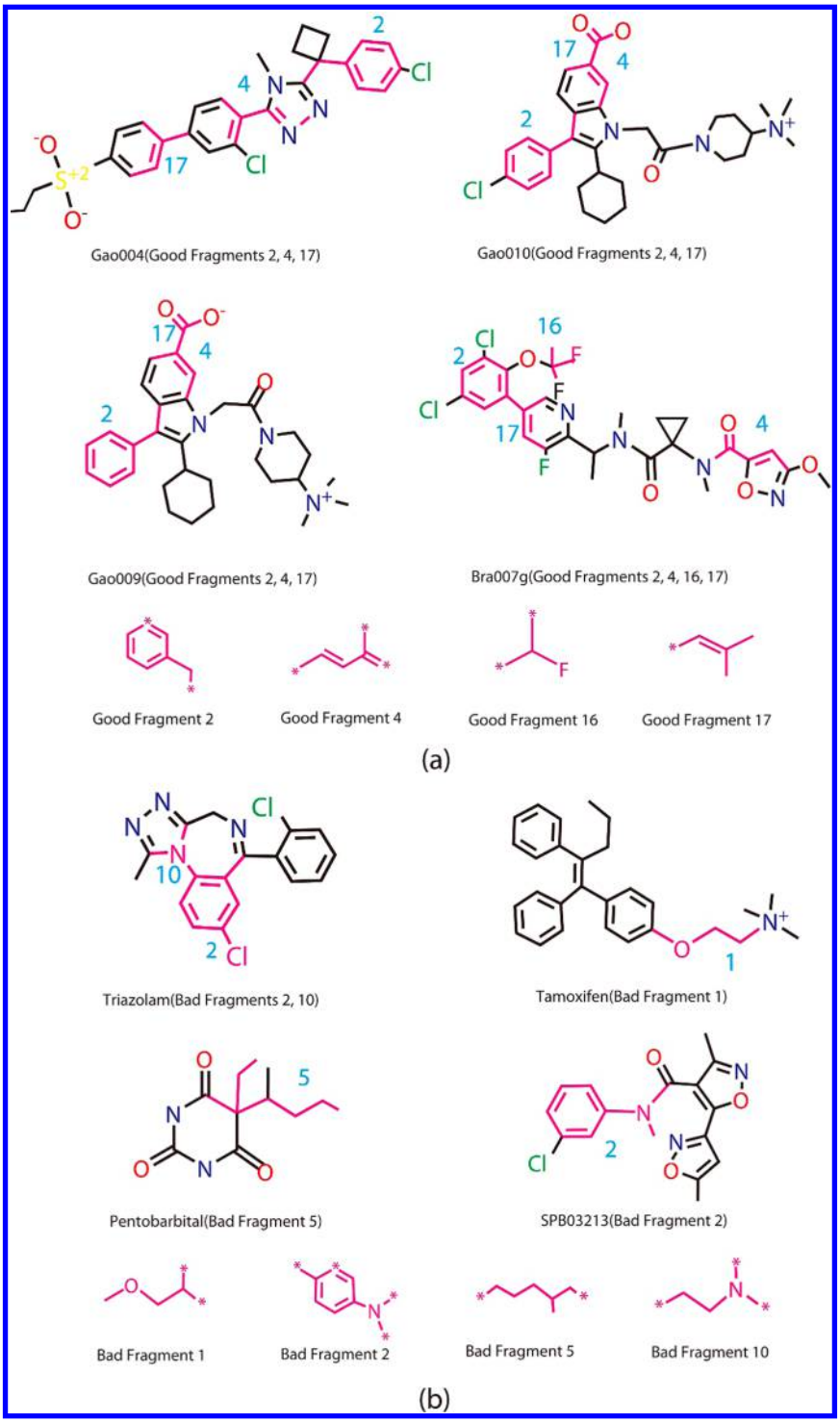


Figure 5. (a) The four misclassified PXR nonactivators and (b) the four misclassified PXR activators by using the best Bayesian classifier.

activators have overlaps between −40 and 10, which is the "uncertain area". The Bayesian scores of seven compounds (Table S6 in Supporting Information) fall into this region. Thus, the predictions given by the best Bayesian classifier for these compounds are certainly not reliable. At last, we used the threshold of 100 μM EC_{50} for classifying the 54 molecules in the test set into the activators and nonactivators. The cutoff of EC_{50} for defining a molecule to be active or nonactive may be arbitrary.

CONCLUSIONS

Here, we reported an extensive data set of 532 structurally diverse PXR activators and nonactivators. Subsequently, the distributions of eight important molecular physicochemical properties for the PXR activators and nonactivators were examined systematically. We found that any single molecular property could not distinguish the PXR activators from nonactivators effectively. Then, a robust machine learning approach, NBC, was employed to develop the classifiers for predicting PXR activation. The best Bayesian classifier based on

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

the 21 molecular properties, *VolSurf* descriptors, and LCFC_10 fingerprints achieves AUC values of 0.814 for the training set and 0.882 for the test set. Finally, the 20 important structural fragments favorable or unfavorable for PXR activation were highlighted, and they provide a deeper insight on designing improved drug candidates for avoiding PXR activation activities in the early stage of drug discovery. The well-validated Bayesian classifier can be served as a powerful tool to filter out potential PXR activators.

■ ASSOCIATED CONTENT

Supporting Information

Descriptions of the 21 molecular physicochemical descriptors used for PXR activation predictions, performance of the naive Bayesian classifiers for the training and test sets based on different molecular descriptors (fingerprints based on three different maximum diameters, 4, 6, or 8), performance of the naive Bayesian classifiers for the balanced training set (179 PXR activators and 179 PXR nonactivators) and for the test set (41 PXR activators and 13 PXR nonactivators) based on different molecular descriptors (fingerprints based on the maximum diameter of 10), and Bayesian scores predicted by the best Bayesian classifier for the eight misclassified molecules in the test set. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Tingjun Hou. E-mail: tingjunhou@zju.edu.cn or tingjunhou@hotmail.com.

*Huidong Yu. E-mail: huidongyu@yahoo.com.

*Xuechu Zhen. E-mail: zhenxuechu@suda.edu.cn.

Author Contributions

⊥H.S. and S.T. contributed equally.

Funding

This study was supported by the National Science Foundation of China (Grant 21173156), the National Basic Research Program of China (973 program, 2012CB932600), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory for Carbon-Based Functional Materials and Devices, and Collaborative Innovation Center of Suzhou Nano Science and Technology.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS:

PXR, pregnane X receptor; NR, nuclear receptor; SXR, xenobiotic sensing nuclear receptor; CAT, chloramphenicol acetyltransferase; LBD, ligand-binding domain; MDR, multidrug resistance; LOO, leave-one-out; QSAR, quantitative structure−activity relationship; ADME, absorption, distribution, metabolism, and excretion; RP, recursive partitioning; ANN, artificial neural network; SVM, support vector machine; NBC, naive Bayesian classification; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives; SE, sensitivity; SP, specificity; Q_+, prediction accuracy of PXR activators; Q_-, prediction accuracy of PXR nonactivators; GA, global accuracy; *C*, Matthews correlation coefficient; ROC, receive operating characteristic; AUC, the area under the ROC curve

■ REFERENCES

(1) Chawla, A., Repa, J. J., Evans, R. M., and Mangelsdorf, D. J. (2001) Nuclear receptors and lipid physiology: Opening the X-files. *Science 294*, 1866−1870.

(2) Dybdahl, M., Nikolov, N. G., Wedebye, E. B., Jónsdóttir, S. Ó., and Niemelä, J. R. (2012) QSAR model for human pregnane X receptor (PXR) binding: Screening of environmental chemicals and correlations with genotoxicity, endocrine disruption and teratogenicity. *Toxicol. Appl. Pharmacol. 262*, 301−309.

(3) Lehmann, J. M., McKee, D. D., Watson, M. A., Willson, T. M., Moore, J. T., and Kliewer, S. A. (1998) The human orphan nuclear receptor PXR is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions. *J. Clin. Invest. 102*, 1016.

(4) Jones, S. A., Moore, L. B., Shenk, J. L., Wisely, G. B., Hamilton, G. A., McKee, D. D., Tomkinson, N. C., LeCluyse, E. L., Lambert, M. H., and Willson, T. M. (2000) The pregnane X receptor: A promiscuous xenobiotic receptor that has diverged during evolution. *Mol. Endocrinol. 14*, 27−39.

(5) Kliewer, S. A., Goodwin, B., and Willson, T. M. (2002) The nuclear pregnane X receptor: A key regulator of xenobiotic metabolism. *Endocr. Rev. 23*, 687−702.

(6) Ekins, S. (2004) Predicting undesirable drug interactions with promiscuous proteins *in silico*. *Drug Discovery Today 9*, 276−285.

(7) Xie, W., Uppal, H., Saini, S. P., Mu, Y., Little, J. M., Radominska-Pandya, A., and Zemaitis, M. A. (2004) Orphan nuclear receptor-mediated xenobiotic regulation in drug metabolism. *Drug Discovery Today 9*, 442−449.

(8) Ekins, S., Chang, C., Mani, S., Krasowski, M. D., Reschly, E. J., Iyer, M., Kholodovych, V., Ai, N., Welsh, W. J., Sinz, M., Swaan, P. W., Patel, R., and Bachmann, K. (2007) Human pregnane X receptor antagonists and agonists define molecular requirements for different binding sites. *Mol. Pharmacol. 72*, 592−603.

(9) Moore, L. B., Goodwin, B., Jones, S. A., Wisely, G. B., Serabjit-Singh, C. J., Willson, T. M., Collins, J. L., and Kliewer, S. A. (2000) St. John's wort induces hepatic drug metabolism through activation of the pregnane X receptor. *Proc. Natl. Acad. Sci. U.S.A. 97*, 7500−7502.

(10) Moore, J. T., and Kliewer, S. A. (2000) Use of the nuclear receptor PXR to predict drug interactions. *Toxicology 153*, 1−10.

(11) Ekins, S., and Erickson, J. A. (2002) A pharmacophore for human pregnane X receptor ligands. *Drug Metab. Dispos. 30*, 96−99.

(12) Zimniak, P., Radominska, A., Zimniak, M., and Lester, R. (1988) Formation of three types of glucuronides of 6-hydroxy bile acids by rat liver microsomes. *J. Lipid. Res. 29*, 183−190.

(13) Xie, W., Radominska-Pandya, A., Shi, Y., Simon, C. M., Nelson, M. C., Ong, E. S., Waxman, D. J., and Evans, R. M. (2001) An essential role for nuclear receptors SXR/PXR in detoxification of cholestatic bile acids. *Proc. Natl. Acad. Sci. U.S.A. 98*, 3375−3380.

(14) Sinz, M., Wallace, G., and Sahi, J. (2008) Current industrial practices in assessing CYP450 enzyme induction: preclinical and clinical. *AAPS J. 10*, 391−400.

(15) Knight, A. W., Little, S., Houck, K., Dix, D., Judson, R., Richard, A., McCarroll, N., Akerman, G., Yang, C., and Birrell, L. (2009) Evaluation of high-throughput genotoxicity assays used in profiling the US EPA ToxCast chemicals. *Regul. Toxicol. Pharmacol. 55*, 188−199.

(16) Gao, Y. D., Olson, S. H., Balkovec, J. M., Zhu, Y., Royo, I., Yabut, J., Evers, R., Tan, E. Y., Tang, W., Hartley, D. P., and Mosley, R. T. (2007) Attenuating pregnane X receptor (PXR) activation: A molecular modelling approach. *Xenobiotica 37*, 124−138.

(17) Ung, C. Y., Li, H., Yap, C. W., and Chen, Y. Z. (2007) In Silico Prediction of Pregnane X Receptor Activators by Machine Learning Approache. *Mol. Pharmacol. 71*, 158−168.

(18) Khandelwal, A., Krasowski, M. D., Reschly, E. J., Sinz, M. W., Swaan, P. W., and Ekins, S. (2008) Machine learning methods and docking for predicting human pregnane X receptor activation. *Chem. Res. Toxicol. 21*, 1457−1467.

(19) Kortagere, S., Chekmarev, D., Welsh, W. J., and Ekins, S. (2009) Hybrid scoring and classification approaches to predict human pregnane X receptor activators. *Pharm. Res. 26*, 1001−1011.

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

(20) Chen, C.-N., Shih, Y.-H., Ding, Y.-L., and Leong, M. K. (2011) Predicting Activation of the Promiscuous Human Pregnane X Receptor by Pharmacophore Ensemble/Support Vector Machine Approach. *Chem. Res. Toxicol. 24*, 1765−1778.
(21) Pan, Y., Li, L., Kim, G., Ekins, S., Wang, H., and Swaan, P. W. (2011) Identification and validation of novel human pregnane X receptor activators among prescribed drugs via ligand-based virtual screening. *Drug Metab. Dispos. 39*, 337−344.
(22) Chen, S., He, N., Chen, W., Sun, F., Li, L., Deng, R., and Hu, Y. (2014) Molecular insights into the promiscuous interaction of human pregnane X receptor (hPXR) with diverse environmental chemicals and drug compounds. *Chemosphere 96*, 138−145.
(23) Lewis, D., Jacobs, M., Dickins, M., and Lake, B. (2002) Quantitative structure−activity relationships for inducers of cytochromes P450 and nuclear receptor ligands involved in P450 regulation within the CYP1, CYP2, CYP3 and CYP4 families. *Toxicology 176*, 51−57.
(24) Fang, H., Tong, W., Welsh, W. J., and Sheehan, D. M. (2003) QSAR models in receptor-mediated effects: The nuclear receptor superfamily. *J. Mol. Struct.: THEOCHEM 622*, 113−125.
(25) Mankowski, D. C., and Ekins, S. (2003) Prediction of human drug metabolizing enzyme induction. *Curr. Drug Metab. 4*, 381−391.
(26) Schuster, D., and Langer, T. (2005) The identification of ligand features essential for PXR activation by pharmacophore modeling. *J. Chem. Inf. Model. 45*, 431−439.
(27) Stanley, L. A., Horsburgh, B. C., Ross, J., Scheer, N., and Wolf, C. R. (2006) PXR and CAR: Nuclear receptors which play a pivotal role in drug disposition and chemical toxicity. *Drug Metab. Rev. 38*, 515−597.
(28) Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res. 28*, 235−242.
(29) Lemaire, G., Benod, C., Nahoum, V., Pillon, A., Boussioux, A.-M., Guichou, J.-F. O., Subra, G., Pascussi, J.-M., Bourguet, W., Chavanieu, A., and Balaguer, P. (2007) Discovery of a highly active ligand of human pregnane X receptor: a case study from pharmacophore Modeling and virtual screening to "in vivo" biological activity. *Mol. Pharmacol. 72*, 572−581.
(30) Evans, R. M. (1988) The steroid and thyroid hormone receptor superfamily. *Science 240*, 889−895.
(31) Lemaire, G., Mnif, W., Pascussi, J. M., Pillon, A., Rabenoelina, F., Fenet, H., Gomez, E., Casellas, C., Nicolas, J. C., Cavailles, V., Duchesne, M. J., and Balaguer, P. (2006) Identification of new human pregnane X receptor ligands among pesticides using a stable reporter cell system. *Toxicol. Sci. 91*, 501−509.
(32) Sinz, M., Kim, S., Zhu, Z., Chen, T., Anthony, M., Dickinson, K., and Rodrigues, A. D. (2006) Evaluation of 170 xenobiotics as transactivators of human pregnane X receptor (hPXR) and correlation to known CYP3A4 drug interactions. *Curr. Drug Metab. 7*, 375−388.
(33) Ekins, S., Kholodovych, V., Ai, N., Sinz, M., Gal, J., Gera, L., Welsh, W. J., Bachmann, K., and Mani, S. (2008) Computational discovery of novel low micromolar human pregnane X receptor antagonists. *Mol. Pharmacol. 74*, 662−672.
(34) di Masi, A., De Marinis, E., Ascenzi, P., and Marino, M. (2009) Nuclear receptors CAR and PXR: Molecular, functional, and biomedical aspects. *Mol. Aspects Med. 30*, 297−343.
(35) Matter, H., Anger, L. T., Giegerich, C., Guessregen, S., Hessler, G., and Baringhaus, K.-H. (2012) Development of in silico filters to predict activation of the pregnane X receptor (PXR) by structurally diverse drug-like molecules. *Bioorg. Med. Chem. 20*, 5352−5365.
(36) Halgren, T. A. (1999) MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem. 20*, 720−729.
(37) SYBYL molecular simulation package, Certara Inc., St. Louis, 2014, http://www.certara.com.
(38) Ghose, A. K., Viswanadhan, V. N., and Wendoloski, J. J. (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A 102*, 3762−3772.
(39) Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N., and Villa, A. E. P. (2001) Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci. 41*, 1488−1493.
(40) Discovery Studio 3.1 Guide, Accelrys Inc., San Diego, 2012, http://www.accelrys.com.
(41) Cruciani, C., Crivori, P., Carrupt, P. A., and Testa, B. (2000) Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *J. Mol. Struct.: THEOCHEM 503*, 17−30.
(42) MOE molecular simulation package, Chemical Computing Group Inc., Montreal, Candada, 2010. http://www.chemcomp.com.
(43) Ekins, S., Shimada, J., and Chang, C. (2006) Application of data mining approaches to drug delivery. *Adv. Drug Delivery Rev. 58*, 1409−1430.
(44) Rogers, D., Brown, R. D., and Hahn, M. (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening 10*, 682−686.
(45) Chen, L., Li, Y., Zhao, Q., Peng, H., and Hou, T. (2011) ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharmaceutics 8*, 889−900.
(46) Cooper, J. A., Saracci, R., and Cole, P. (1979) Describing the Validity of Carcinogen Screening-Tests. *Br. J. Cancer 39*, 87−89.
(47) Chen, L., Li, Y., Yu, H., Zhang, L., and Hou, T. (2012) Computational models for predicting substrates or inhibitors of P-glycoprotein. *Drug Discovery Today 17*, 343−351.
(48) Tian, S., Wang, J., Li, Y., Xu, X., and Hou, T. (2012) Drug-likeness Analysis of Traditional Chinese Medicines: Prediction of Drug-likeness Using Machine Learning Approaches. *Mol. Pharmaceutics 9*, 2875−2886.
(49) Wang, S., Li, Y., Wang, J., Chen, L., Zhang, L., Yu, H., and Hou, T. (2012) ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of hERG Potassium Channel Blockage. *Mol. Pharmaceutics 9*, 996−1010.
(50) Tian, S., Li, Y., Li, D., Xu, X., Wang, J., Zhang, Q., and Hou, T. (2013) Modeling Compound-Target Interaction Network of Traditional Chinese Medicines for Type II Diabetes Mellitus: Insight for Polypharmacology and Drug Design. *J. Chem. Inf. Model. 53*, 1787−1803.
(51) Tian, S., Sun, H., Li, Y., Pan, P., Li, D., and Hou, T. (2013) Development and Evaluation of an Integrated Virtual Screening Strategy by Combining Molecular Docking and Pharmacophore Searching Based on Multiple Protein Structures. *J. Chem. Inf. Model. 53*, 2743−2756.
(52) Wang, S., Li, Y., Xu, L., Li, D., and Hou, T. (2013) Recent Developments in Computational Prediction of hERG Blockage. *Curr. Top. Med. Chem. 13*, 1317−1326.
(53) Hou, T. J., Wang, J. M., Zhang, W., and Xu, X. J. (2007) ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *J. Chem. Inf. Model. 47*, 460−463.
(54) Hou, T. J., Wang, J. M., Zhang, W., and Xu, X. J. (2007) ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model. 47*, 208−218.
(55) Hou, T. J., Wang, J. M., Zhang, W., Wang, W., and Xu, X. (2006) Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem. 13*, 2653−2667.
(56) Hou, T., and Wang, J. (2008) Structure - ADME relationship: Still a long way to go? *Expert Opin. Drug Metab. Toxicol. 4*, 759−770.

DOI: 10.1021/tx500389q
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX