Pharmaceutical Research, Vol. 26, No. 4, April 2009 (© 2008)
DOI: 10.1007/s11095-008-9809-7

Research Paper

Hybrid Scoring and Classification Approaches to Predict Human Pregnane X Receptor Activators

Sandhya Kortagere,[1,2] Dmitriy Chekmarev,[1] William J. Welsh,[1] and Sean Ekins[1,3,4,5]

Received November 6 2008; accepted December 10 2008; published online December 30, 2008

Purpose. The human pregnane X receptor (PXR) is a transcriptional regulator of many genes involved in xenobiotic metabolism and excretion. Reliable prediction of high affinity binders with this receptor would be valuable for pharmaceutical drug discovery to predict potential toxicological responses

Materials and Methods. Computational models were developed and validated for a dataset consisting of human PXR (PXR) activators and non-activators. We used support vector machine (SVM) algorithms with molecular descriptors derived from two sources, *Shape Signatures* and the Molecular Operating Environment (MOE) application software. We also employed the molecular docking program GOLD in which the GoldScore method was supplemented with other scoring functions to improve docking results.

Results. The overall test set prediction accuracy for PXR activators with SVM was 72% to 81%. This indicates that molecular shape descriptors are useful in classification of compounds binding to this receptor. The best docking prediction accuracy (61%) was obtained using 1D *Shape Signature* descriptors as a weighting factor to the GoldScore. By pooling the available human PXR data sets we revealed those molecular features that are associated with human PXR activators.

Conclusions. These combined computational approaches using molecular shape information may assist scientists to more confidently identify PXR activators.

KEY WORDS: docking; hybrid methods; machine learning; pregnane X receptor; shape signatures descriptors; support vector machine.

INTRODUCTION

The transcriptional regulation of genes involved in xenobiotic metabolism and excretion is an important area of study, in particular the human pregnane X receptor, PXR (NR1I2; also known as SXR or PAR) has been a particular focus since its identification (1–6). PXR activators include a wide range of structurally diverse endogenous bile acids, hormones, dietary vitamins, prescription and herbal drugs as well as environmental chemicals. PXR activators can mediate potential drug–drug interactions and the toxic effects of environmental chemicals (7,8), hence the need to develop reliable prediction methods.

Four PXR X-ray crystal structures are available in the Protein Data Bank (PDB), which have enabled characterization of the ligand binding domain (LBD). The pocket is lined with 28 amino acid residues: 20 hydrophobic, four polar and four charged (9–14). Due to the large size of the binding pocket, molecules can bind in multiple locations, which hinders reliable prediction of PXR activators (A) or non-activators (N) using structure-based virtual screening methods. Several previous studies have constructed ligand-based computational models for human PXR employing pharmacophores (15–18), quantitative structure–activity relationships (QSARs) (19–21), and machine learning methods (21). For example, the human PXR agonist pharmacophore models contain multiple hydrophobic features, at least one hydrogen bond acceptor and, in some cases, an additional hydrogen bond donor feature which are deemed important for molecular recognition of ligands by PXR.

The absence of large biological data sets for PXR ligands has hampered efforts to build QSAR models for quantitative predictions (22). The sparse amount of data is more suitable for classification models than quantitative prediction models. Several studies using qualitative data sets (≥99 molecules) have employed machine learning methods such as recursive partitioning (RP), random forest (RF), support vector machine (SVM), K-nearest neighbors (K-NN), and probabilistic neural networks (PNN) (21,22) to distinguish between

Electronic supplementary material The online version of this article (doi:10.1007/s11095-008-9809-7) contains supplementary material, which is available to authorized users.

[1] Department of Pharmacology and Environmental Bioinformatics and Computational Toxicology Center (ebCTC), University of Medicine and Dentistry of New Jersey (UMDNJ)-Robert Wood Johnson Medical School, 675 Hoes lane, Piscataway, NJ 08854, USA.

[2] Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA 19129, USA.

[3] Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046, USA.

[4] Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD 21201, USA.

[5] To whom correspondence should be addressed. (e-mail: ekinssean@yahoo.com)

human PXR activators (A) and non-activators (N). In the latter case (22), binary classification models generated from 98 human PXR activators and 79 non-activators were used to predict between 80.8% and 85.0% of human PXR activators and 67.7–73.6% of human PXR non-activators (in the training set). The test set prediction accuracies in this same study ranged from 53.3% to 66.7% for 15 known human PXR activators across the three machine learning methods, with SVM performing the best (22). We have recently constructed human PXR classification models based on VolSurf (23) 3D descriptors employing RP, RF and SVM machine learning methods, and made predictions for a new large external test set of 145 molecules (not included in the training set) to validate and compare these three approaches (24). Our results were a significant improvement on those previously published (22). We also have previously utilized structure-based docking using FlexX combined with logistic regression; however, the overall results were disappointing and inferior to those obtained using the machine learning methods. However, when the docking predictions were correct, the docked orientations were useful for comparison with structurally similar molecules that were PXR non-activators (24). To the best of our knowledge there have been no other published direct comparisons of docking and QSAR methods for PXR with a large external test set containing diverse xenobiotics, although pharmaceutical companies are very likely to have done such studies in house.

In the current study we have greatly extended our previous work to use additional molecular descriptors, namely Shape Signatures and MOE, which have been recently applied to cardiotoxicity target proteins and blood–brain barrier data with machine learning classification methods (25,26). In these previous studies, it was found that 2D Shape Signature descriptors slightly outperformed 1D Shape descriptors with the SVM algorithm. Additionally SVM models based on Shape Signatures also performed slightly better than those developed with the MOE descriptors for the same datasets (25,26). In the current study we have also used an additional docking method GOLD (27–29) and a novel aspect of this study (with respect to PXR) is our coupling of the GoldScore with other scoring functions in an attempt to improve the overall docking results. We have also combined all the available human PXR data sets (~300 molecules) into a single model in order to identify specific molecular descriptors that correlate highly with human PXR activation. Through judicious combination of these computational approaches, we have identified new computational models that can predict human PXR activators with applicability ranging from drug discovery to toxicological sciences.

MATERIALS AND METHODS

Data Compilation

A comprehensive human PXR dataset was assembled from two previously published datasets, as described in detail elsewhere (22,24). Briefly, the first set (training compounds) retrieved 168 compounds from data published by Ung *et al.* (22) comprising 93 that were defined as activators ($EC_{50}<$ 100 μM) and 75 as inactivators ($EC_{50}>100$ μM). The second set (test compounds) taken from our previous work (24) contained 130 compounds, comprising 71 human PXR activators and 59 non-activators. Using the SMILES strings taken from the original papers as input, each compound was geometry optimized to a low energy conformation generated by CORINA (Molecular Networks GmbH, Nägelsbachstr. 25, 91052 Erlangen, Germany. http://www.mol-net.de) and assigned partial atomic charges according to the Gasteiger-Marsili scheme (30).

Molecular Descriptors

Structural analysis of the four available human PXR crystal structure complexes with their ligands from the PDB (PDB IDs: 1M13, resolution 2.00 Å (10), 1SKX, resolution 2.80 Å (14), 2O9I, resolution 2.80 Å (13) and 1NRL, resolution 2.00 Å (9)) revealed that the docking of the four agonists was dominated by electrostatic and van der Waals interactions. To capture these ligand-protein interaction features into our classification scheme we chose 20 molecular descriptors that represented shape and size (volume, weight, KierA1-A3, Kier1-3), flexibility (number of rotatable bonds, number of rings and KierFlex) and electrostatic features (logP, topological polar surface area (TPSA), logS, Lipinksi donor (lip_don), Lipinski acceptor (lip_acc), number of N atoms, and number of O atoms). Values for these specific molecular descriptors were calculated using the Molecular Operating Environment (MOE, Chemical Computing Group, Montreal, Canada) modeling program. In addition, we used a shape based descriptor method called Shape Signatures (31) to build SVM based classification models for the PXR activators and non-activator sets.

Shape Signatures Molecular Descriptors

Shape Signatures (31) is a novel shape matching algorithm that has been previously described (25,26,31,32) which allows fast comparison between any pair of typical drug-like molecules. In this method, molecular features such as shape and distribution of partial charges, which are critical for competent binding and hence the relative biological activity of the compound are encoded in the form of the one-dimensional (1D) and two-dimensional (2D) histograms. Here we will briefly outline the key steps of the algorithm. The process starts by generating a single low energy three-dimensional conformation (a default conformation) of a molecule under consideration using CORINA from a supplied SMILES string. In the second step, the solvent accessible surface (SAS) is constructed around the molecule which is followed by triangulation of the SAS by the SMART program (31). Next, a customized ray-tracing algorithm is used to explore the volume enclosed by the SAS. During this stage, a ray of light, emitted initially from a randomly selected point on the interior lining of the SAS, travels inside the molecular compartment bounded by the SAS until it strikes the opposite side, at which point it gets reflected back and propagates further in the direction determined by law of optical reflections. For each reflection point, the value of the truncated Coulomb potential or the molecular electrostatic potential (MEP), and the lengths of the incident and reflected ray segments are recorded and stored in the memory. It was determined empirically that for a typical drug-like molecule

100,000 reflections are sufficient for the paths of the rays to thoroughly survey the molecular volume enclosed by SAS. At the end of the run, the ray segments are binned by their length into a one-dimensional histogram with the predefined bin width of 0.5 Å (1D signature) and a two-dimensional histogram is constructed with the MEP values (with a step of 0.05e/Å) and the associated total length of the two path segments joined by the reflection point (2D signature). Both histograms are then normalized.

Similar to previous applications (25,26) for each compound in this study, the heights of the bins of the associated 1D (shape only) and 2D (shape and polarity) Shape Signature histograms constituted two sets of distinct molecular descriptors. It is important to stress that despite being represented as 1D and 2D histograms, these Shape Signatures descriptors are inherently three-dimensional molecular descriptors since they encode the 3D conformation and polarity of the molecule.

Docking and Scoring

A combined data set consisting of 297 molecules from the test and the training sets described above along with the four ligands (hyperforin, rifampicin, T0901317 and SR12813 that were co-crystallized with the human PXR receptor) were used for docking experiments. The molecules were docked into these four crystallized structures of human PXR (PDB IDs described above). In all cases, the crystal structure ligand was removed, and hydrogen atoms were added to the amino acids. All amino acids within 6 Å of the co-crystallized ligand were identified as the binding site. The docking program GOLD (ver 4 (29)) was used for docking all the 301 compounds to the binding sites of each PXR crystal structure. GOLD (ver 4) uses a genetic algorithm to explore the various conformations of ligands and flexible receptor side chains in the binding pocket. Further, 30 independent docking runs were performed for each ligand. The docked complexes were initially scored with GoldScore (29) and then rescored using ChemScore (33). The best ranking conformation for each ligand was chosen based on the most favorable binding energy (ΔG values from ChemScore) and the corresponding GoldScore of that conformation was used for further scoring procedures. Various cutoffs of the GoldScore were used to classify compounds into PXR activators and non-activators. The Score-1 scheme used 50% of the GoldScore of the crystal structure ligand as a cut-off, while Score-5 used the GoldScore of the crystal structure ligand −10 as a cut-off, and Score-6 used the GoldScore of the crystal structure ligand −30 as a cut-off. A number of additional customizable scoring schemes were also tested for classifying the compounds as activators and non-activators and the scoring schemes are further described below:

1. Contact scoring scheme (Score-2): The docked receptor–ligand complexes were scored using a contact based scoring function. Accordingly, an in-house program was used to examine the docked complexes for contacts between the ligand and protein atoms (34). Further, these contacts were scored based on a weighting scheme that was derived from the nature of interaction between the ligands co-crystallized with human PXR. For example, hyperforin forms hydrogen bonds with residues Gln285, His407 and Ser247 of the PXR protein in the crystal structure (PDB ID:1M13; Supplemental Figure 1). Thus the contact scoring function overweighted all those docked protein–ligand complexes that featured hydrogen bonding between the ligand and these three residues. Similarly, other non-bonded interactions were weighted based on the interactions of the ligands in the PXR crystal structures. All interaction scores were then summed and normalized for the 301 compounds against all four crystal structures. A consensus scoring scheme was developed for final classification based on the following rule: Only those compounds that equaled or exceeded half the value of the highest GoldScore and exhibited a non-zero contact score were assigned as activators while the remaining were classified as non-activators.

2. Shape based scoring scheme: In this scheme, the ligands from the combined dataset were compared with the PXR ligands from the four crystal structures for their shape based similarities using two different approaches. The first was based on 2D similarity encoded in MDL fingerprint keys calculated using Discovery Studio 2.0 (Accelrys, San Diego, CA, USA). The Tanimoto coefficient was used as the metric to compare the molecular fingerprints. The coefficients varied between 0 and 1, where 0 meant maximally dissimilar and 1 coded for maximally similar. The Tanimoto coefficient between fingerprints X and Y has been defined to be: [number of features in intersect (A, B)]/[number of features in union (A, B)], where A and B are two compounds.

   In the second approach, the 3D shapes of the molecules from the combined dataset were compared with the shapes of each of the four crystal structure ligands. This was achieved by comparing their corresponding 1D Shape Signatures and a dissimilarity score was computed for each ligand pair (see description for Shape Signatures below). The dissimilarity score was then converted to a similarity score, which was in turn used as weighting factor for the GoldScore to compute Score-4. In all these scoring schemes the consensus score was calculated as shown below in Eq. 1.

3. Molecular descriptor based scoring: In this scheme, the molecular descriptors computed using MOE were used to calculate Euclidean distances from the four crystal structure ligands. These Euclidean distances were used as weighting factors to compute Score-7. Similarly, the values of the molecular descriptors were also used to calculate Tanimoto similarity indices (35) with reference to the four crystal structure ligands. For a pair of descriptor vectors with continuous variables, the Tanimoto index ranged from −0.33 (when one vector is the inverse of another) to 1 (for two identical vectors). The values for the Tanimoto indices for each ligand in the combined set were calculated against each of the crystal structure ligands and then used as weighting factors to the GoldScores to compute Score-3.

The weighted docking score of an active compound *j* with *i* conformations was described as

$$S_{i,j} = w_i s_{ij} \tag{1}$$

where s_{ij} was the original GoldScore for the compound *i* in its *j*th conformation and w_i is the weighting factor for compound i from either of the schemes described above.

Classification by Support Vector Machines

The SVM method (36, 37) is a powerful machine learning classification technique that has been used widely to tackle complex binary classification problems (22, 38–40). Following our previous studies (25, 26) we have used the freely available program LIBSVM (C-SVM)(41) with the radial basis function (Gaussian) kernel, whose parameter γ along with the SVM penalty term C were determined in each case through a simple grid search procedure by tenfold cross validation.

In addition to the libraries of 1D and 2D Shape Signatures generated for the datasets described in the previous section, in this study we also explored a number of mixed descriptor schemes. In particular, we combined Shape Signatures and MOE molecular descriptors to investigate the performance of these classification models with those built entirely with either Shape Signatures or MOE descriptors. Also, we have combined docking scores, GoldScores and contact scores, with the set of MOE descriptors and evaluated the performance of the SVM models utilizing this hybrid descriptor scheme. In total, we have considered six different descriptor approaches (Table I): 1D Shape Signatures, 1D Shape Signatures + MOE, 2D Shape Signatures, 2D Shape Signatures + MOE, MOE alone and MOE + GOLD docking.

SVM Model Testing

The predictive nature of each SVM model described herein was assessed by computing a standard set of statistical indicators: sensitivity (SE), specificity (SP), overall prediction accuracy (Q) and Matthews correlation coefficient (C) (22, 42). These quantities are defined in terms of the numbers of true positives (TP), true PXR activators, false positives (FP), falsely classified PXR non-activators, true negatives (TN), true PXR non-activators and false negatives (FN), falsely classified PXR activators. In these notations, the total number of real experimentally documented activators is given by TP + FN whereas a corresponding number of real non-activators is TN + FP. Sensitivity, SE=TP/(TP+FN), then expresses the prediction accuracy of a classification model with respect to PXR activators while specificity reflects the prediction accuracy for non-activators: SP=TN/(TN+FP). The overall prediction accuracy is calculated as Q=(TP+TN)/B(TP+FP+TN+FN). Finally, we also report the values of Matthews correlation coefficient described as in Eq. 2.

$$C = [\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}] \Big/ [\,(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})\,]^{1/2} \tag{2}$$

This represents another measure of the overall prediction performance. For a perfect classifier with FP=FN=0 one would have C=1.0. For a random prediction, C≈0, and for a complete inversion (TP=TN=0) C=−1.0.

In order to carefully examine the quality of the proposed SVM models we used the following procedure for model validation as used previously (25,26). Prior to submitting data for SVM analysis, the dimensionality of the input datasets in

Table I. Ranking of hPXR SVM Models Based on their Performance for Different Datasets

Rank\dataset	Training	Test	Combined
1	1DSS + MOE C=0.531, Q_{LN}=77% Q_{CV}=81%	MOE C=0.464, Q_{LN}=74% Q_{CV}=78%	MOE C=0.466, Q_{LN}=74% Q_{CV}=77%
2	2DSS + MOE C=0.431, Q_{LN}=72% Q_{CV}=75%	2DSS + MOE C=0.370, Q_{LN}=69% Q_{CV}=73%	1DSS + MOE C=0.413, Q_{LN}=71% Q_{CV}=73%
3	MOE C=0.405, Q_{LN}=71% Q_{CV}=76%	MOE + DOCK C=0.327, Q_{LN}=67% Q_{CV}=72%	MOE + DOCK C=0.369, Q_{LN}=69% Q_{CV}=73%
4	1DSS C=0.392, Q_{LN}=70% Q_{CV}=74%	2DSS C=0.305, Q_{LN}=66% Q_{CV}=72%	2DSS + MOE C=0.361, Q_{LN}=69% Q_{CV}=73%
5	2DSS C=0.345, Q_{LN}=68% Q_{CV}=73%	1DSS + MOE C=0.289, Q_{LN}=66% Q_{CV}=72%	2DSS C=0.349, Q_{LN}=68% Q_{CV}=70%
6	MOE + DOCK C=0.326, Q_{LN}=67% Q_{CV}=72%	1DSS C=0.173, Q_{LN}=60% Q_{CV}=68%	1DSS C=0.283, Q_{LN}=65% Q_{CV}=70%

C and Q_{LN} were computed from 100 independent leave-20%-out runs and Q_{CV} from tenfold cross validation conducted on the entire dataset each case was reduced by means of the unsupervised forward selection (UFS) method of Livingstone and co-workers (43). This program eliminates redundancy and reduces multicollinearity of the original datasets. Two families of SVM models were then built for each dataset/descriptor set pair. The models in the first group were generated by tenfold cross validations conducted on the entire dataset. The overall prediction accuracies for these models are denoted by Q_{CV}. The models comprising the second group were averaged over a series of 100 independent leave-20%-out runs (overall accuracies Q_{LO}). The leave-20%-out tests were designed as follows: For each dataset, about 20% of the molecules were randomly picked to represent the hold-out test set and the rest of the data constituted the training set for this particular data division. The selection was carried out to approximately preserve the correct proportion of PXR activators and non-activators in both sets. Each SVM classification algorithm was then trained on the training set and applied to predict class attributes of the compounds in the test set. To obtain more reliable statistical estimates, the procedure was repeated 100 times, each time with a different composition of the test and training sets.

RESULTS

The combined dataset consisting of 301 compounds including the four PXR co-crystallized ligands were docked using GOLD to the four crystal structures of human PXR (1M13, 1NRL, 1SKX and 2O9I) available in the PDB (Supplemental Figure 1). The four co-crystallized ligands were docked to their respective crystal structures within 0.2–0.6 Å root mean square deviation (rmsd). Cross docking of all the four ligands to all the crystal structures was achieved with no significant deviations to the binding site or binding mode, except for the case of rifampicin which docked in a flipped mode to structure 1M13 (Supplemental Figure 2). The cross docking of the co-crystallized ligands yielded an rmsd in the range 0.34–1.05 Å relative to their corresponding crystal structure conformations, confirming the validity of these human PXR crystal structures for docking studies under the conditions described in this study. In all the cross-docking studies, hyperforin docked with the best score, followed by SR12813, rifampicin and T0901317 (Table II). Structural similarity analysis using MDL keys with a cut off value of 0.5 showed that only 26% of the ligands from the combined dataset shared 2D structural similarity to hyperforin and 15% of the ligands to rifampicin and little or no similarity to SR12813 and T0901317. Hence, we used all four crystal structures in order to dock all the ligands.

Table II. Raw Docking Scores (GoldScores) for the Four Crystal Structure Ligands Docked to the hPXR Crystal Structures

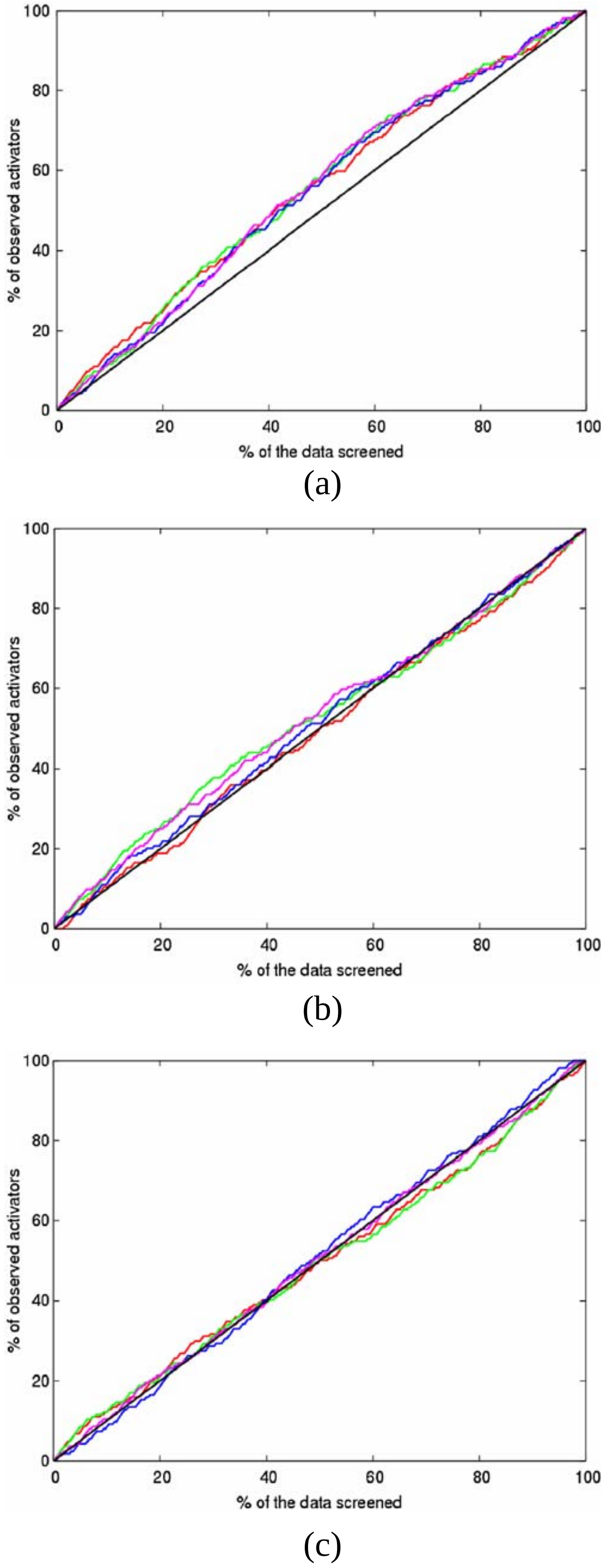
Structure	Hyperforin	SR12813	Rifampicin	T0901317
1M13	79.97	70.81	59.21	45.08
1NRL	78.28	64.04	49.05	45.97
1SKX	88.31	66.88	65.76	45.61
2O9I	87.19	71.22	51.69	47.98

It is reasonable to assume that molecules which are more structurally similar to the four crystal structure ligands (hyperforin, rifampicin, T0901317 and SR12813) would have a better chance to be successfully docked to the receptor than structures sharing less similarity with these ligands. Based on this premise, we devised and tested novel scoring schemes for molecular docking which accounts for such molecular similarity via incorporation of the specially designed weighting factors. According to this scheme, the compounds which were structurally closer to the known ligands were weighted more favorably than those that were dissimilar in the space of the chosen molecular descriptors. A set of molecular descriptors were identified that clustered PXR activators closer to the four crystal structure ligands and non-activators further away in terms of similarity. This was achieved by computing similarity scores for every molecule in the combined dataset (297 structures) with respect to each of the four co-crystallized ligands. The resulting list was ranked according to the similarity scores and the positions of PXR activators and non-activators were analyzed. Three sets of molecular descriptors were considered: 1D Shape Signatures (shape); 2D Shape Signatures (shape and electrostatics); and MOE (shape, flexibility, electrostatics and hydrogen bonding). For binned data (1D and 2D Shape Signatures), a χ^2-based distance metric commonly used for comparing discrete distributions was used. For MOE descriptors, Tanimoto similarity scores were calculated for descriptor vectors with continuous variables (35). To better understand this ranking system based on molecular similarity across all combinations of crystal-structure ligand and molecular descriptors sets, a hit rate (HR) analysis (43) was conducted as depicted in equation 3. The HR monitors the fraction of observed activators ($N_{\mathrm{act}\%}/N_{\mathrm{act}}$) at the fixed portion of the list (sorted database) screened ($N_{\%}/N_{\mathrm{tot}}$).

$$HR = (N_{act\%}/N_{act})/(N_{\%}/N_{tot}) \tag{3}$$

where N_{act}=163 molecules, N_{tot}=297 molecules, $N_{\mathrm{act}\%}$ was the number of activators listed among the top $N_{\%}$ of the ranked list screened. For a uniform distribution of PXR activators and non-activators across the sorted list, meaning that there was no preference between activators and non-activators in terms of their structural similarity to the given ligand, HR≈1 for any percent of the database screened. For a perfect separation, when all activators ranked above all non-activators, the ($N_{\mathrm{act}\%}/N_{\mathrm{act}}$) *vs.* ($N_{\%}/N_{\mathrm{tot}}$) curve would be expected to lie above the random line and plateau at 1 for $N_{\%}/N_{\mathrm{tot}}$=0.55. As can be concluded from Fig. 1, the only set of molecular descriptors for which the HR curves was consistently placed above the random selection line was the 1D Shape Signatures descriptors, which characterize molecular shape and size. In this case, at 55% of the database, 60% to 65% of all activators were recovered depending on the crystal structure ligand.

Based on the HR performance summarized in Fig. 1, 1D Shape Signatures descriptors were used to design a structural similarity weighted scheme (scheme-4) scoring function for docking. The raw docking scores for the four ligands to the four crystal structures are shown in Table II. The docked structures were classified as activators (A) and non-activators (N) based on the various scoring schemes (Table III and Supplemental Table VIII). Scoring scheme-1 was devised to



(a)

(b)

(c)

Fig. 1. Results of the hit rate analysis conducted for a number of molecular descriptors schemes for four endogenous PXR ligands: T0901317 (*red*), hyperforin (*green*), rifampicin (*blue*) and SR12813 (*purple*). **a** 1D Shape Signatures descriptors. **b** 2D Shape Signatures descriptors. **c** MOE molecular descriptors.

use 50% of the raw GoldScore (listed in Table II) as a cutoff for the A and N classification. The average Q value in this scheme for all the four structures was ~55. Similarly, scheme-5 used the raw GoldScore of the PXR crystal ligand −10 as a cutoff, while scheme-6 used raw GoldScore −30 as a cutoff. The results show a Q value in the range of 45–57% and 47–56% for schemes 5 and 6 respectively. Changing the cutoff for the raw docking scores did not enhance the classification rate as shown in scoring schemes-5 and 6.

Scoring scheme-2 which was derived based on the contact scores had Q values were in the range of 44% to 56% but the C values were mostly negative indicating a performance that was worse than random. Scoring scheme-3 that takes into account the Tanimoto indices built from molecular descriptors derived from MOE yielded an average success rate of ~48% and *C* values were negative for the four crystal structures (except 2O9I). These results suggest that combining the similarity indices with docking scores may have limited utility as a scoring scheme which was is also reflected in the SVM based model (see Supplementary Table VI). Similar results were obtained from scheme-7 (Supplementary Table VIII) that was based on intermolecular Euclidean distances instead of the Tanimoto indices.

Further analysis of the docking results showed that among all the docking scoring schemes, scheme-4 performed the best in identifying PXR activators and non-activators with a Q value of 61% (*C*=0.209) for 1M13 followed by Q value of 60% (*C*=0.211) for 1NRL crystal structures. The perfor-

Table III. Docking Results for the Combined Dataset to the Four hPXR Crystal Structures with Scoring Schemes 1–4

Scheme/structure		1M13	1NRL	1SKX	2O9I
Score-1	SE (%)	94	98	100	98
	SP (%)	11	2	3	4
	Q (%)	57	55	56	56
	C	0.092	−0.013	0.129	0.057
Score-2	SE (%)	43	61	88	64
	SP (%)	46	30	11	46
	Q	44	47	54	56
	C	−0.103	−0.096	−0.006	0.104
Score-3	SE (%)	40	38	33	41
	SP (%)	59	60	62	59
	Q	49	48	46	49
	C	−0.007	−0.019	−0.047	0.005
Score-4	SE (%)	75	63	40	57
	SP (%)	45	58	73	54
	Q	61	61	55	56
	C	0.209	0.211	0.149	0.119

Sensitivity (SE), specificity (SP), overall prediction accuracy (Q) and Matthews correlation coefficient (C) are described in the text mance of scheme-4 for 1SKX and 2O9I was very similar with an average Q value of ~55.5% (C=0.149 and 0.119 respectively). Scoring scheme-4 was constructed using 1D Shape Signatures as a weighting factor to the GoldScore. In fact, HR analysis showed that 1D Shape Signatures could correctly identify nearly 64% of the compounds as PXR activators. Thus utilizing the 1D scores of compounds to the 4 crystal structure ligands improved the classification rate of the raw docking scores (score-1) from 56% to 61% for 1M13 and 54% to 60% for 1NRL structures.

Finally, a consensus scoring approach based on a majority vote for all the crystal structures using scheme-4 was built (Supplemental Table IX). Based on the consensus analysis we found that many of the activators such as 3-ketolithocholic acid and ritonavir were correctly predicted to be activators and their interactions in the binding site mimicked those of the crystal structure ligands (Fig. 2A, B). Similarly, the non-activators such as clofazimine and 3-MC, although docked to the same binding pocket and sharing 2D structural similarities, did not have the same set of molecular interactions (and hence poor binding energetics) with the receptor as that of the crystal structure ligands (Fig. 2C, D). We also found some cases where PXR non-activators such as amiodarone (Fig. 3A) and levonorgestrel (Fig. 3B) were predicted to be activators and conversely PXR activators such as amentoflavone (Fig. 3C) and monobenzylphthalate (Fig. 3D) were classified as non-activators based on the consensus scoring scheme.

The results of SVM classifications for 18 combinations of PXR dataset—descriptor sets have been summarized in Table I and a detailed analysis for each model can be found in the Supplementary Tables I–VII. For each dataset, SVM models were ranked according to their performance in the leave-20%-out internal testing. In order to provide a comparison to the results from the previously published studies on the same dataset (22,24) we have chosen to use Q_{CV} (Q values from cross validation) for the SVM models. Overall, SVM models based on MOE descriptors and two mixed descriptor schemes 1D Shape Signatures + MOE and 2D Shape Signatures + MOE performed best. In each case, the difference in the overall

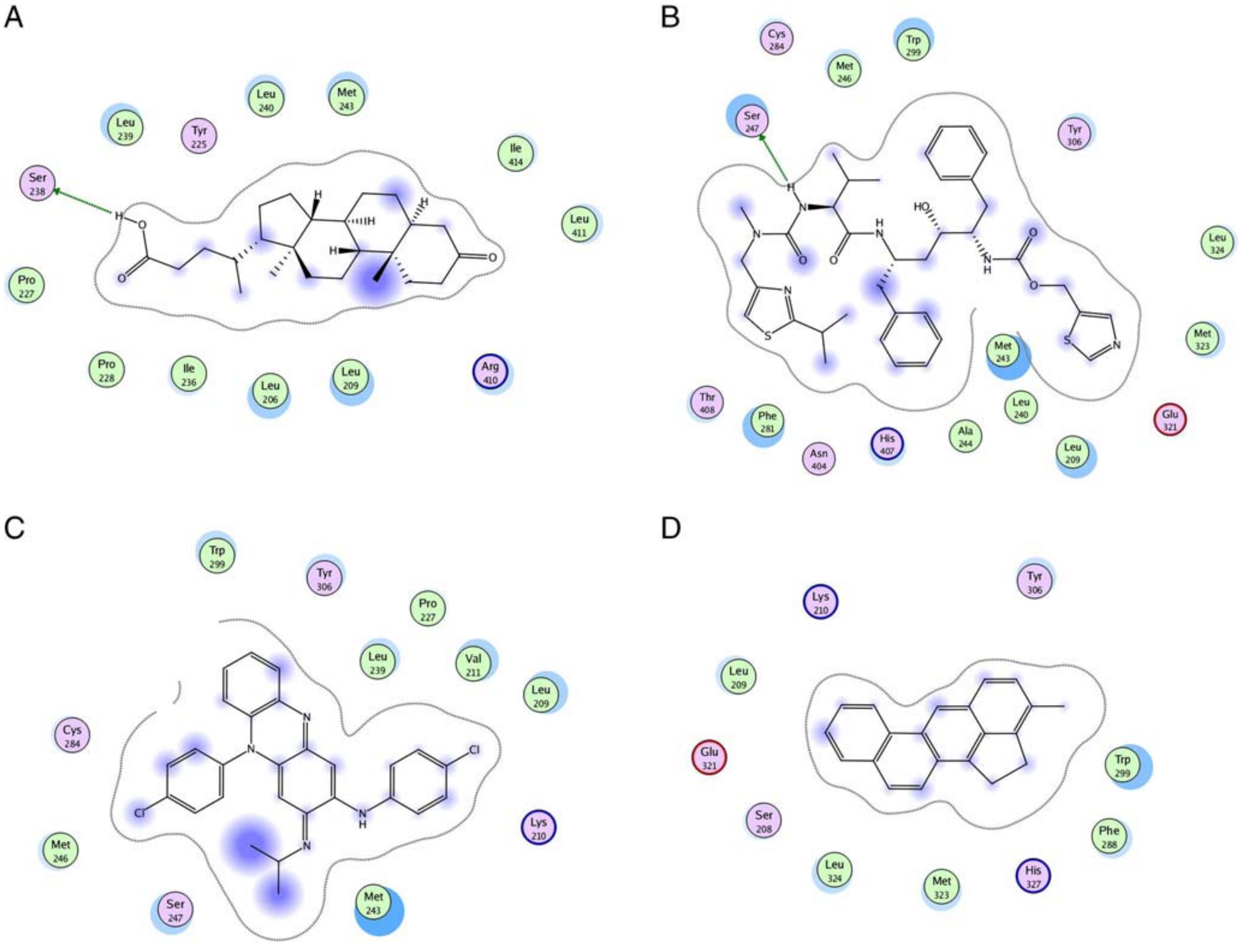


Fig. 2. Schematic representation of the binding mode of **A** 3-keto lithocholic acid **B** Ritonavir **C** Clofazimine and **D** 3-MC in the binding site of crystal structure of human PXR protein (PDB code: 1M13). The binding site residues are colored by their nature, with hydrophobic residues in *green* and charged residues in *purple*. *Blue spheres* and *contours* indicate matching regions between ligand and receptors. The schematic pictures were generated using LIGX option in MOE.

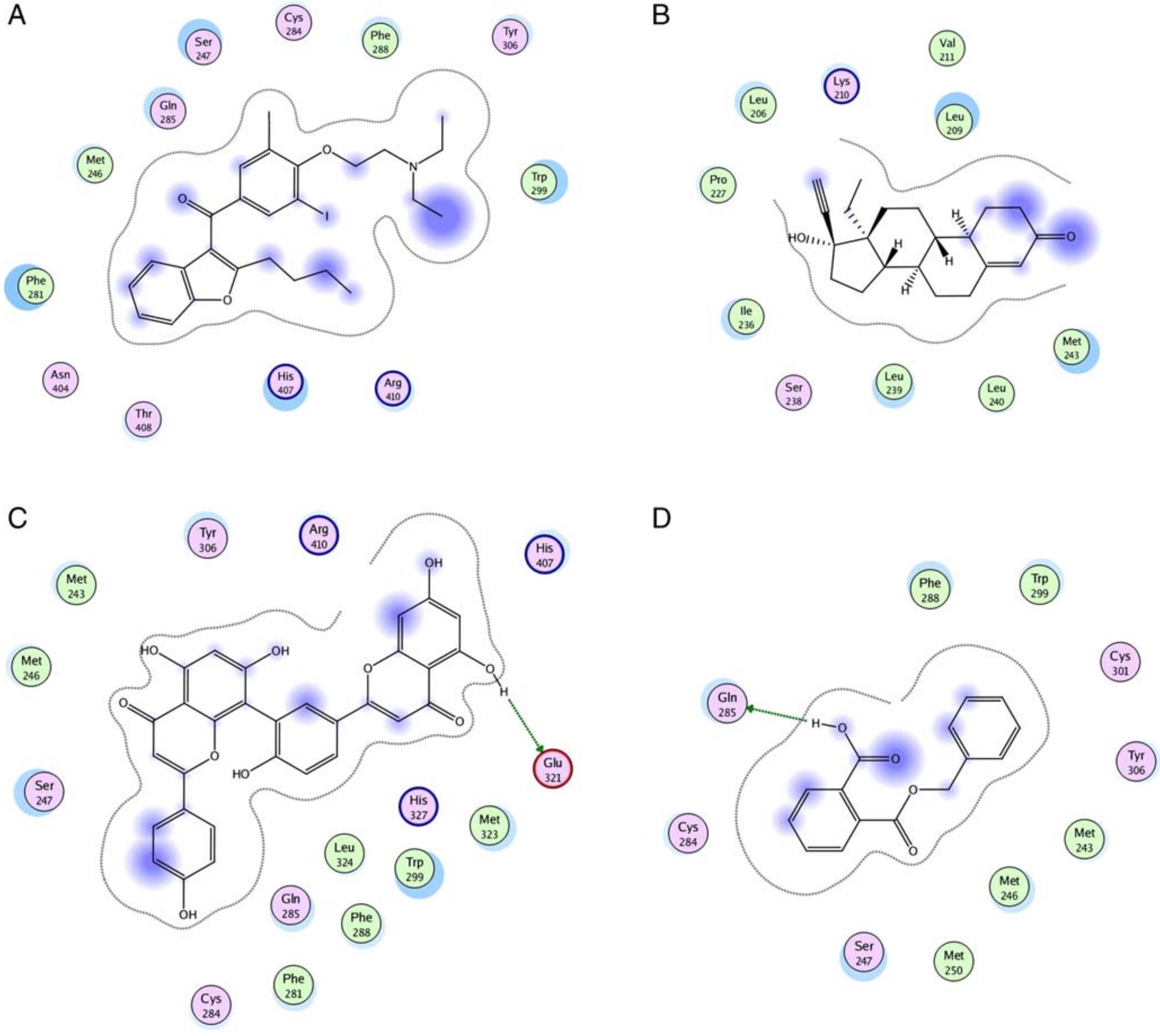


Fig. 3. Schematic representation of the binding mode of **A** Amiodarone **B** Levonorgestrel **C** Amentoflavone and **D** Monobenzyl phthalate in the binding site of crystal structure of human PXR protein (PDB code: 1M13). The binding site residues are colored by their nature, with hydrophobic residues in *green* and charged residues in *purple*. *Blue spheres* and *contours* indicate matching regions between ligand and receptors. The schematic figures were generated using LIGX option in MOE.

accuracy Q_{LN} between the best and worst models was between 9% and 14%. Further, mixing MOE descriptors with the original docking scores failed to improve the quality of SVM models (Supplemental Table VI).

Further, to probe the overlap between the regions of the chemical space occupied by molecules from the training and test sets, we conducted principal component analysis (PCA) which can provide valuable information regarding the relative position of the chemical structures in the space defined by the molecular descriptors. As clearly shown in Supplementary Figure 3A, compounds from the two sets are intermixed in the space of the MOE molecular descriptors, which validates the results of the cross-set predictions described above. Similar observations were found in previous studies using other molecular descriptor schemes (24,25). We also clustered these compounds using the molecular descriptors with the Unweighted Pair Group Method with Arithmetic Mean algorithm (UPGMA). This represents a simple agglomerative clustering method which is based on unweighted average distances between the nodes (Supplementary Figure 3B). These results confirm that our test and training sets represent similar chemical space.

DISCUSSION

PXR is a member of the nuclear receptor family of ligand-activated transcription factors and is an integral component of the body's defense mechanism against toxic endobiotics and xenobiotics. The binding of structurally diverse ligands is made possible by the large volume and shape of the largely hydrophobic ligand-binding pocket. This also presents problems for computational modeling which is compounded by the lack of specificity, generally low micromolar affinity and wide structural variability among the ligands that bind to this receptor. The main emphasis of this study was therefore to develop scoring functions and machine learning based methods that can capture the binding modes of structurally dissimilar ligands binding to PXR. Further, the compounds were classified as PXR activators and non-activators using various custom and hybrid scoring schemes. Other classification schemes such as SVM based machine learning models with a variety of molecular descriptors were also used for classification.

We have demonstrated for the first time that there is no significant difference between the four available crystal structures for PXR when using GOLD for docking ligands to the hypothesized ligand binding domain (agonist) site as evidenced by the small rmsd values in the docking and cross docking experiments. This would suggest that any of the four structures could be used for docking with GOLD without seeing appreciable differences.

We have also shown that the 2D Shape Signatures and MOE descriptor schemes (which, in addition to shape, also account for electrostatic features and other molecular properties) surprisingly performed poorly when used in an attempt to create a hybrid scoring function. These observations confirmed that the dividing boundaries between activators and non-activators were indeed likely to be complex. Hence, sophisticated machine learning methods like SVM were needed to build reasonable classification models. Scoring scheme-2 which was derived based on the contact scores did not perform as well as in previous studies with other targets such as GPCRs (34). This could be due to the fact that contact scores bias the scoring scheme by weighting those ligands that bind to the receptor in a similar mode as the crystal structure ligands. This scheme of biasing works well when the ligands have very high specificity towards the target such as GPCRs (34).

Scheme-4 performed the best in identifying PXR activators and non-activators with Q and C values that show a considerable apparent improvement when compared to the earlier FlexX docking and logistic regression scoring scheme for these same compounds (24). The previous docking efforts resulted in random results and highlighted the difficulty in docking in a large binding pocket for such a flexible protein. In the current study, the primary reason for any docking misclassifications may be due to the non-activators having very similar 1D Shape Signatures to the crystal structure ligands that were classified as activators and *vice versa*. Consequently, in the context of classification of molecules as potential PXR activators, the direct unprocessed results of docking experiments would appear of limited use as reported by us previously(24).

Our previous study (24) also used the compounds in the test set (third column in Table I) as an external test set for the classification models generated using the structures from the training set (second column in Table I). Such cross-set testing yielded Q values in the range of 63–67% depending on the classification method used (RP, RF or SVM). A similar exercise in this study in which SVM models trained on the training dataset were applied to classify molecules from the test dataset was performed. The prediction rates from our analysis indicate a *Q* value that is essentially in the same range of 57% to 67% (24) and perhaps this is indicative of an upper limit of prediction for PXR SVM based models with this dataset and the descriptors used to date in these studies.

Mixing MOE descriptors with the original docking scores did not improve the quality of SVM machine learning models. These results were in agreement with the docking results using scoring schemes 3 and 7. This is understandable since classification based on docking scores alone produced results only slightly better than random. On the other hand, merging MOE descriptors with the Shape Signatures descriptors improved the quality of predictions when compared with either 1D or 2D Shape Signature based SVM models (Supplemental Tables I–IV). Since 1D Shape Signatures account for the overall molecular shape and 2D Shape Signatures characterize molecular shape and polarity, it was interesting to find that a combination of MOE descriptors with Shape Signatures may improve performance with SVM models in this case. We have performed UFS on the mixed descriptor set to choose the combination of descriptors for each of the SVM models (see Supplemental Table VII). Although 2D Shape Signatures account for the overall MEP of the molecule, they incorporate specific hydrogen bonding features only implicitly. Hence, addition of more specific MOE descriptors, such as those associated with hydrogen bonds, improved the performance of the classification models (Table I and Supplemental Table II).

In summary, we have shown that creation of a hybrid docking and molecular descriptor-based method can be used to improve the prediction accuracy for PXR activators, which represents a major computational challenge for docking and scoring programs. The 1D Shape Signature descriptors alone were found to perform particularly well in this regard giving 61% correct predictions. In addition the overall test set prediction accuracy for PXR activators with the SVM classification method was 72% to 81% using a combination of 1D Shape Signatures descriptors and MOE descriptors, which was comparable to the results obtained previously with VolSurf descriptors and SVM (24). Our results also suggest an apparent ceiling on the prediction accuracy using the same training and test sets. Future studies will likely test the hybrid scoring/docking and classification schemes with shape-based approaches developed in this study with larger external test sets for PXR. As larger quantitative datasets of PXR activators become available in the public domain, it may be feasible to generate QSAR and quantitative scoring methods for docking that expand on the classification methods developed here and elsewhere. In addition we will assess the broader applicability of this hybrid docking and classification approach to other proteins. Shape Signatures have been previously used alone as molecular descriptors with machine learning methods (25,26). This study has further broadened the applicability of the Shape Signatures method to both PXR and in combination with the docking derived GoldScore. The approach could be useful for filtering molecule libraries for their potential to cause adverse effects or drug–drug interactions mediated by PXR or other proteins of interest in the pharmaceutical and environmental sciences fields.

ACKNOWLEDGMENTS

We gratefully acknowledge all those involved in the application and development of Shape Signatures. S.E. kindly acknowledges Dr. Akash Khandelwal, Dr. Peter Swaan and Dr. Matthew Krasowski for previous discussions on PXR and Accelrys for kindly providing Discovery Studio 2.0. Support for this work has been provided by the USEPA-funded Environmental Bioinformatics and Computational Toxicology Center (ebCTC), under STAR Grant number GAD R 832721-010, and by the Defense Threat Reduction Agency, under contract number HDTRA-BB07TAS020. This work was also funded by NIH-GM081394 from the National Institute of General Medical Sciences (to WJW). This work has not been reviewed by and does not represent the opinions of the funding agencies.

REFERENCES

1. G. Bertilsson, J. Heidrich, K. Svensson, M. Asman, L. Jendeberg, M. Sydow-Backman, R. Ohlsson, H. Postlind, P. Blomquist, and A. Berkenstam. Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction. *Proc. Natl. Acad. Sci. USA.* **95**:12208–12213 (1998).
2. B. Blumberg, W. Sabbagh Jr., H. Juguilon, J. Bolado Jr., C. M. van Meter, E. S. Ong, and R. M. Evans. SXR, a novel steroid and xenobiotic-sensing nuclear receptor. *Genes. Dev.* **12**:3195–3205 (1998).
3. S. A. Kliewer, J. T. Moore, L. Wade, J. L. Staudinger, M. A. Watson, S. A. Jones, D.D. McKee, B. B. Oliver, T. M. Willson, R. H. Zetterstrom, T. Perlmann, and J. M. Lehmann. An orphan nuclear receptor activated by pregnanes defines a novel steroid signalling pathway. *Cell.* **92**:73–82 (1998).
4. B. Goodwin, L. B. Moore, C. M. Stoltz, D. D. McKee, and S. A. Kliewer. Regulation of the human CYP2B6 gene by the nuclear pregnane X receptor. *Mol. Pharmacol.* **60**:427–431 (2001).
5. J. Staudinger, Y. Liu, A. Madan, S. Habeebu, and C. D. Klaassen. Coordinate regulation of xenobiotic and bile acid homeostasis by pregnane X receptor. *Drug Metab. Dispos.* **29**:1467–1472 (2001).
6. J. L. Staudinger, B. Goodwin, S. A. Jones, D. Hawkins-Brown, K. I. MacKenzie, A. LaTour, Y. Liu, C. D. Klaassen, K. K. Brown, J. Reinhard, T. M. Willson, B. H. Koller, and S. A. Kliewer. The nuclear receptor PXR is a lithocholic acid sensor that protects against liver toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **98**:3369–3374 (2001).
7. S. Harmsen, I. Meijerman, J. H. Beijnen, and J. H. Schellens. The role of nuclear receptors in pharmacokinetic drug–drug interactions in oncology. *Cancer Treat. Rev.* **33**:369–380 (2007).
8. T. W. Synold, I. Dussault, and B. M. Forman. The orphan nuclear receptor SXR coordinately regulates drug metabolism and efflux. *Nature Med.* **7**:584–590 (2001).
9. R. E. Watkins, P. R. Davis-Searles, M. H. Lambert, and M. R. Redinbo. Coactivator binding promotes the specific interaction between ligand and the pregnane X receptor. *J. Mol. Biol.* **331**:815–828 (2003).
10. R. E. Watkins, J. M. Maglich, L. B. Moore, G. B. Wisely, S. M. Noble, P. R. Davis-Searles, M. H. Lambert, S. A. Kliewer, and M. R. Redinbo. 2.1A crystal structure of human PXR in complex with the St John's Wort compound hyperforin. *Biochemistry.* **42**:1430–1438 (2003).
11. R. E. Watkins, S. M. Noble, and M. R. Redinbo. Structural insights into the promiscuity and function of the human pregnane X receptor. *Curr. Opin. Drug Discov. Devel.* **5**:150–158 (2002).
12. R. E. Watkins, G. B. Wisely, L. B. Moore, J. L. Collins, M. H. Lambert, S. P. Williams, T. M. Willson, S. A. Kliewer, and M. R. Redinbo. The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity. *Science.* **292**:2329–2333 (2001).
13. Y. Xue, L. B. Moore, J. Orans, L. Peng, S. Bencharit, S. A. Kliewer, and M. R. Redinbo. Crystal structure of the pregnane X receptor–estradiol complex provides insights into endobiotic recognition. *Mol. Endocrinol.* **21**:1028–1038 (2007).
14. J. E. Chrencik, J. Orans, L. B. Moore, Y. Xue, L. Peng, J. L. Collins, G. B. Wisely, M. H. Lambert, S. A. Kliewer, and M. R. Redinbo. Structural disorder in the complex of human pregnane X receptor and the macrolide antibiotic rifampicin. *Mol. Endocrinol.* **19**:1125–1134 (2005).
15. K. Bachmann, H. Patel, Z. Batayneh, J. Slama, D. White, J. Posey, S. Ekins, D. Gold, and L. Sambucetti. PXR and the regulation of apoA1 and HDL-cholesterol in rodents. *Pharmacol. Res.* **50**:237–246 (2004).
16. S. Ekins, C. Chang, S. Mani, M. D. Krasowski, E. J. Reschly, M. Iyer, V. Kholodovych, N. Ai, W. J. Welsh, M. Sinz, P. W. Swaan, R. Patel, and K. Bachmann. Human pregnane X receptor antagonists and agonists define molecular requirements for different binding sites. *Mol. Pharmacol.* **72**:592–603 (2007).
17. S. Ekins, and J. A. Erickson. A pharmacophore for human pregnane-X-receptor ligands. *Drug Metab. Dispos.* **30**:96–99 (2002).
18. D. Schuster, and T. Langer. The identification of ligand features essential for PXR activation by pharmacophore modeling. *J. Chem. Inf. Model.* **45**:431–439 (2005).
19. S. Ekins, L. Mirny, and E. G. Schuetz. A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXRa and LXRb. *Pharm. Res.* **19**:1788–1800 (2002).
20. M. N. Jacobs. In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology.* **205**:43–53 (2004).
21. S. Ekins, S. Andreyev, A. Ryabov, E. Kirillov, E. A. Rakhmatulin, S. Sorokina, A. Bugrim, and T. Nikolskaya. A combined approach to drug metabolism and toxicity assessment. *Drug Metab. Dispos.* **34**:495–503 (2006).
22. C. Y. Ung, H. Li, C.W. Yap, and Y. Z. Chen. In silico prediction of pregnane X receptor activators by machine learning approaches. *Mol. Pharmacol.* **71**:158–168 (2007).
23. G. Cruciani, P. Crivori, P. A. Carrupt, and B. Testa. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. THEOCHEM. **503**:17–30 (2000).
24. A. Khandelwal, M. D. Krasowski, E. J. Reschly, M. W. Sinz, P. W. Swaan, and S. Ekins. Machine learning methods and docking for predicting human pregnane X receptor activation. *Chem. Res. Toxicol.* **21**:1457–1467 (2008).
25. D. S. Chekmarev, V. Kholodovych, K. V. Balakin, Y. Ivanenkov, S. Ekins, and W. J. Welsh. Shape signatures: new descriptors for predicting cardiotoxicity in silico. *Chem. Res. Toxicol.* **21**:1304–1314 (2008).
26. S. Kortagere, D. Chekmarev, W. J. Welsh, and S. Ekins. New predictive models for blood–brain barrier permeability of drug-like molecules. *Pharm. Res.* **25**:1836–1845 (2008).
27. A. Evers, G. Hessler, H. Matter, and T. Klabunde. Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. *J. Med. Chem.* **48**:5448–5465 (2005).
28. A. Evers, and T. Klabunde. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J. Med. Chem.* **48**:1088–1097 (2005).
29. G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**:727–748 (1997).
30. J. Gasteiger, and M. Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron.* **36**:3219–3228 (1980).
31. R. J. Zauhar, G. Moyna, L. Tian, Z. Li, and W. J. Welsh. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **46**:5674–5690 (2003).
32. K. Nagarajan, R. Zauhar, and W. J. Welsh. Enrichment of ligands for the serotonin receptor using the Shape Signatures approach. *J. Chem. Inf. Model.* **45**:49–57 (2005).
33. M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comp.-Aided. Mol. Des.* **11**:425–445 (1997).

34. S. Kortagere, and W. J. Welsh. Development and application of hybrid structure based method for efficient screening of ligands binding to G-protein coupled receptors. *J. Comp.-Aided. Mol. Des.* **20**:789–802 (2006).
35. T. Kogej, O. Engkvist, N. Blomberg, and S. Muresan. Multi-fingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **46**:1201–1213 (2006).
36. C. Cortes, and V. Vapnik. Support vector networks. *Machine Learn.* **20**:273–293 (1995).
37. V. Vapnik. *Statistical learning theory*. Wiley, New York, 1998.
38. Y. Z. Chen, C. W. Yap, and H. Li. Current QSAR techniques for toxicology. In S. Ekins (ed.), *Computational Toxicology: risk assessment for pharmaceutical and environmental chemicals*, Wiley, Hoboken, 2007, pp. 217–238.
39. M. K. Leong. A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem. Res. Toxicol.* **20**:217–226 (2007).
40. Y. Xue, C. W. Yap, L. Z. Sun, Z. W. Cao, J. F. Wang, and Y. Z. Chen. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **44**:1497–1505 (2004).
41. C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines, 2001.
42. B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **405**:442–451 (1975).
43. D. C. Whitley, M. G. Ford, and D. J. Livingstone. Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **40**:1160–1168 (2000).