Published in final edited form as:
Chem Res Toxicol. 2008 July ; 21(7): 1457–1467. doi:10.1021/tx800102e.

Machine Learning Methods and Docking For Predicting Human Pregnane X Receptor Activation

Akash Khandelwal[†], **Matthew D. Krasowski**[#], **Erica J. Reschly**[#], **Michael W. Sinz**[‡], **Peter W. Swaan**[†], and **Sean Ekins**[†,+,#,*]

[†] *Department of Pharmaceutical Sciences, University of Maryland, 20 Penn Street, Baltimore, MD 21201, USA*

[#] *Department of Pathology, University of Pittsburgh, Pittsburgh, PA 15261, USA*

[‡]*Bristol-Myers Squibb Company, Research Parkway, Wallingford, CT 06492, USA*

[+]*Department of Pharmacology, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854, USA*

[#]*Collaborations in Chemistry, 601 Runnymede Ave, Jenkintown, PA 19046, USA.*

Abstract

The pregnane X receptor (PXR) regulates the expression of genes involved in xenobiotic metabolism and transport. *In vitro* methods to screen for PXR agonists are used widely. In the current study, computational models for human PXR activators and PXR non-activators were developed using recursive partitioning (RP), random forest (RF) and support vector machine (SVM) algorithms with VolSurf descriptors. Following 10 fold randomization, the models correctly predicted 82.6 % to 98.9 % of activators and 62.0 % to 88.6 % of non-activators. The models were validated using separate test sets. The overall (n = 15) test set prediction accuracy for PXR activators with RP, RF and SVM PXR models is 80 to 93.3 %, representing an improvement over models previously reported. All models were tested with a second test set (n =145) and prediction accuracy ranged from 63−67 % overall. These test set molecules were found to cover the same area in a principal component analysis plot as the training set, suggesting the predictions were within the applicability domain. The FlexX docking method combined with logistic regression performed poorly in classifying this PXR test set compared with RP, RF and SVM, but may be useful for qualitative interpretion of interactions within the LBD. From this analysis, VolSurf descriptors and machine learning methods had good classification accuracy and made reliable predictions within the model applicability domain. These methods could be used for high throughput virtual screening to assess for PXR activation, prior to *in vitro* testing to predict potential drug-drug interactions.

Introduction

The human pregnane X receptor, PXR (NR1I2; also known as SXR or PAR) is a transcriptional regulator of a large number of genes involved in xenobiotic metabolism and excretion. The genes regulated by PXR include cytochrome P450 (CYP) 3A4 (1-3), CYP2B6 (4), aldehyde dehydrogenases, glutathione-S-transferase, sulfotransferases, organic anion transporter peptide 2, and multi drug resistance protein 1 and 2 (5,6) as well as others. Human PXR

[*] To whom correspondence should be addressed. Address: Collaborations in Chemistry, 601 Runnymede Ave, Jenkintown, PA 19046. Phone: 269−930−0974. Fax: 215−481−0159. E-mail: ekinssean@yahoo.com..

Supporting Information Available: Molecules used in Table 2 in this study are available as an SDF file, along with the supplemental tables. This material is available free of charge via the Internet at http://pubs.acs.org.

activators include a wide range of prescription and herbal drugs such as paclitaxel, troglitazone, rifampicin, ritonavir, clotrimazole, and St. John's Wort which can be involved in clinically relevant drug-drug interactions (7). In addition to xenobiotics, PXR is also activated by pregnanes, androstanes, bile acids, hormones, dietary vitamins and a wide array of endogenous molecules reviewed recently (8).

The PXR ligand binding domain (LBD) consists of 12 α-helices that fold to form a hydrophobic pocket and a short region of β-strands. The pocket is lined with twenty eight amino acid residues, twenty hydrophobic, four polar and four charged (9-13). The potential for molecules to bind in numerous locations in the LBD complicates the reliable prediction of PXR activators (A) or non-activators (N) using structure based drug design methods alone. Computational models ranging from ligand based pharmacophores (14-17), quantitative structure activity relationships (QSAR) (18-20), and machine learning methods (20), to homology modeling with molecular dynamics (21) (for identifying protein-co-repressor interactions), represent predominantly reports to predict PXR ligand binding (8) to differing degrees. These previously described computational methods focused on diverse structural types for agonists and in one case used structural analogs (8) which may have assessed specific binding locations within the LBD, such as that for steroidal compounds. A likely consensus has emerged across the different QSAR modeling methods that PXR agonists are required to fit to multiple hydrophobic features and at least one hydrogen bond acceptor (and in some cases an additional hydrogen bond donor feature) (8). A further qualitative observation from these previous studies is the dependence of the resulting agonist QSAR or pharmacophore models on the molecules used in the training set, and potential for overlap of multiple models derived from different molecules (8). It should also be noted that rarely do the published QSAR models utilize a large external test set to validate the predictive nature or assess the applicability domain (22-24) of the training and test sets, i.e. how structurally similar do the molecules in the training and test set have to be for accurate predictions. This is especially important to build confidence in the use of these methods with such structurally promiscuous proteins as PXR. One of the limitations of using published data for PXR is that only a small fraction of the data available reports quantitative EC_{50} data, (e.g. much of the work is published as greater or less than a cutoff value e.g. 100 μM). Therefore there are currently no widely available large, diverse continuous datasets to enable quantitative QSAR modeling for human PXR. Two PXR machine learning studies have been published recently with relatively large training sets (≥ 99 molecules) using recursive partitioning (19), support vector machine (SVM), K- nearest neighbors (k-NN) and probabilistic neural network (PNN) (20). In the latter case, binary classification data for 98 human PXR activators and 79 non-activators were used (20) to predict between 80.8 to 85.0 % of human PXR activators and 67.7 to 73.6 % of human PXR non-activators (in the training set); the test set prediction accuracies in this same study ranged from 53.3−66.7 % for 15 known human PXR activators across the three machine learning methods, with SVM performing the best (20).

A structure-based alternative to understanding small molecule-protein interactions is to dock molecules into proteins. These methods have resulted in the discovery of novel inhibitors for many targets (25-28) and have been applied to proteins such as drug metabolizing enzymes which are also relatively promiscuous in their ligand binding (27-29). Docking and scoring ligands in target protein binding sites is a challenging process (30-32) and the performance of different implementations of scoring functions have been found to be target dependent (33). The use of consensus scoring has also been recommended to improve the “hit” enrichment (34). Docking of a small number of molecules into the PXR structure 1NRL has been used previously to design molecules that are weaker agonists (35) and an automated docking method (GOLD (36) which maintains the protein as rigid), was used to flexibly dock azole PXR antagonists onto the outer surface at the AF-2 site (8). We are not aware of any studies using docking as a screening tool to predict the potential for PXR activation on a larger scale with diverse drug-like libraries of molecules and this may be due to the challenge of the protein promiscuity.

In the current study we have compared RP, RF and SVM machine learning methods for building human PXR models derived with VolSurf 3D descriptors (which have been widely applied for ADME/Tox and drug discovery target modeling (37-39)). In addition for comparison, we have used FlexX docking (40,41) of molecules into the crystal structures of human PXR combined with logistic regression. The predictive ability of the classification models and docking models was evaluated using a novel large external test set containing 145 human PXR activators and non-activators (using data for molecules generated in this study and published in the literature) (42-44). Our aim was to identify the most appropriate computational approach to predict whether a molecule was likely to be a human PXR agonist (activator) and in so doing, prioritize molecules for *in vitro* testing.

Materials and Methods

Reagents and Plasmids

The construction of a HepG2 (human liver) cell line stably expressing human SLC10A1, a transporter that can take up conjugated bile salts has been reported before in detail (45). Human PXR was expressed as a full-length protein and CYP3A4-PXRE-Luc, which contains promoter elements from CYP3A4 recognized by the PXR DNA-binding domain, was used as the reporter construct. The plasmids for human PXR, human SLC10A1, CYP3A4-PXRE-Luc, and empty vectors pSG5 were generously provided by S.A. Kliewer, J.T. Moore, and L.B. Moore (GlaxoSmithKline, Research Triangle Park, NC).

Reporter gene assay with HepG2 cells

Human PXR activation in the HepG2 human liver cell line was determined by a luciferase-based reporter assay as previously described (8,46).

Data sets and molecular descriptors

A previously published dataset was used for model building (training) in which compounds with EC_{50}<100 μM (n = 98) were classified as human PXR activators, whereas compounds with EC_{50}>100 μM (n = 79) were classified as human PXR non-activators (20). One small test set also published by the same group consisted of 15 human PXR activators (20). The molecular structures encoded as SMILES strings (47) were downloaded from the supplementary information tables in the original publication (20). We have additionally developed our own novel test set (n = 145) consisting of PXR activators (n = 82) and PXR non-activators (n = 63) from our exhaustive literature searching (over 5 years) and data generated in this study (see above) that are not in the current training sets (8,42,48). The SMILES string for each molecule was downloaded from either PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChemSpider (http://www.chemspider.com/) or sketched using the BUILDER module of SYBYL. The SMILES strings for the previously published training and test sets were converted to SYBYL MOL files using an in-house script. The MOL files were then subsequently minimized in SYBYL (Tripos, St Louis, MO). Energy minimizations were performed using the Tripos force field (49) and Gasteiger-Hückel charges with a distance-dependent dielectric constant and conjugate gradient method with a convergence criterion of 0.001 kcal/mol.

Eighty-six VolSurf descriptors (38) were calculated from the 3D molecular fields using VolSurf 4.0 implemented in SYBYL. Three different probes including water (OH2), dry and amphipathic (BOTH) were used for descriptor calculation. VolSurf descriptors include descriptors for size, shape, hydrophilic and hydrophobic regions, and interaction energy as well as others.

Machine learning model building and validation

RP calculations were performed using the rpart module of the R package (50). Briefly, there are (n) number of compounds and each compound contains a descriptor (x variables) and a class label (activator or non-activator). RP is a technique that builds a set of classification rules based on descriptor information. At the base of the tree, all the n compounds are mixed into a single group. The base is then split into smaller and smaller samples (every sub sample is called node) by choosing a descriptor and the corresponding threshold value to divide the sample. If the compound's descriptor value is above the threshold then it is assigned to one branch and if it is lower then it is assigned to the other branch. The ultimate goal is to separate PXR activators from non-activators. The grown tree can then be used to classify test compounds to activator or non-activator classes.

The R program was also used for RF calculations (51). In the RF method, which is an extension of RP, multiple trees are grown and predictions are made by averaging the multiple trees. The total number of trees was set to 1000. The other optimizable parameter in the random forest is m_{try} i.e. the number of descriptors (p) randomly sampled as candidates for splitting at each node. The number of descriptors was increased systematically with an increment of 5. The out of bag error (OBB) estimate can be considered as equivalent to a cross validation study. In OBB, one third of the compounds are randomly selected as a test set and a model is developed from the remaining compounds, this is then repeated multiple times. The optimum m_{try} was chosen such that %OBB is a minimum. Thus, a lower %OBB indicates higher accuracy of the model.

The Kernlab package in R was used for generating SVM models using the radial basis kernel function. The grid optimization approach was followed to obtain optimum values of C and gamma by varying these parameters from 2^{-5} to 2^{15} and 2^{-15} to 2^{3}, respectively. The optimal C and gamma values for the resultant PXR model were 8 and 0.00781, respectively based on the lowest cross-validation error (52).

Docking

The molecules in the training and validation set were docked into 4 different ligand co-crystallized structures of PXR with hyperforin (PDB ID: 1M13, resolution 2.00 Å)(10), rifampicin (PDB ID: 1SKX, resolution 2.80 Å)(53) , T0901317 (PDB ID: 2O9I, resolution 2.80 Å)(54) and SR12813 (PDB ID: 1NRL, resolution 2.00 Å)(9) using FlexX (BioSolveIT, GmbH, Sankt Augustin) (40,41). The FlexX program considers ligand flexibility by an incremental ligand placement technique while the receptor is considered as rigid. In total, 322 molecules (177 training and 145 testing) were docked into each crystal structure giving rise to 322*4 = 1288 molecules. For each ligand (n = 1288), 30 different docked poses were generated (1288*30 = 38,640 overall poses) and the best pose was selected based on the FlexX score. In all cases the active site was defined as the amino acid residues within 6Å of the co-crystallized ligand. The FlexX scores were then used as the independent (x) variable and actual PXR activators (A) / PXR non-activators (N) as the dependent (y) variable in a logistic regression analysis (Eq. 1). The PXR activators and PXR non-activators were assigned scores of 1 and 0 respectively. This logistic regression analysis was performed for molecules docked in all 4 structures. (Eq. 1).

$$\text{Prob}\,(\text{Y}=\text{A})=\frac{1}{1+\text{e}^{-(\alpha^{*}\text{FlexX}+\beta)}}\quad \text{or}\quad \text{log}\quad \text{odds}\quad (\text{Y}=\text{A})=\alpha^{*}\text{FlexX}+\beta \tag{1}$$

Where, Prob and A represents probability and activator, respectively and α and β represents adjustable parameters. Two different confusion matrices were built for each crystal structure i.e. one using only successfully docked compounds and the other from both successfully docked and failed compounds. In the latter case the molecules which failed to dock inside the LBD of PXR were considered as non-activators.

Data analysis

The performance of RP, RF, SVM and docking were evaluated using true positive (TP), true negative (TN), false positive (FP), false negative (FN), sensitivity (SE) SE=TP/(TP+FN), specificity (SP) SP=TN/(TN+FP), overall prediction accuracy (Q) Q=(TP+TN)/(TP+TN+FP+FN) and Matthew's correlation coefficient (C) $C=((TP^{*}TN)-(FN^{*}FP))/\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}$ (Table 1) (20). Sensitivity and specificity were used to represent the prediction accuracy of PXR activators and non-activators, respectively. In the present context, TP is the number of true PXR activators, TN is the number of true PXR non-activators, FP is the number of falsely classified PXR activators and FN is the number of falsely classified PXR non-activators.

Results

The human PXR data for model development (training set n = 177) and a small initial test set (n = 15 activators) was taken from a recent publication (20), while the experimental data for the large external test set (n = 145 activators and non-activators) was collated in the current study and several recent publications (8,42,48). The predictive performance of RP, RF and SVM with VolSurf descriptors for predicting PXR activators and PXR non-activators were initially based on a 10-fold cross validation metric. A 10-fold cross validation was chosen in order to make direct comparisons with previously published models in order to assess the validity of our selected descriptors and algorithms (20). The overall training set prediction accuracy of PXR models were 87.5, 73.4, and 94.3%, respectively (Table 1 and Supplemental Table 1). The C value of the RP, RF and SVM PXR models were: 0.75, 0.46 and 0.89, respectively (Table 1). For comparison the previous best SVM reported (20) had a training set accuracy of 79.6% compared with 94.3% in this study. The C value ranges between +1 and −1, where a value of +1 indicates perfect prediction, −1 represents an inverse prediction and 0 indicates that the prediction is equivalent to a completely random prediction. The previous best SVM reported (20) had a training set accuracy of C value of 0.598 compared with 0.888 in this study. Based on the data distribution, the probability of randomly selecting a PXR activator or a PXR non-activator is 55.4 % and 44.6 %, respectively. RF would therefore appear to perform poorly compared with the SVM and RP. The 10 fold cross validation results would also indicate that in all cases the classification predictions are better than random. All the models were also validated using external test sets. The first small test set (using 15 known activators from a previous study (20)) resulted in a prediction accuracy for RP, RF and SVM PXR models from 80 to 93.33 % (Supplemental Table 1). The previous best SVM reported (20) correctly predicted 66.7 % of the 15 molecules compared with 93.33 % in this study. Therefore the overall prediction accuracies for the training and test set machine learning models reported here represents an improvement over the models reported previously using other molecular descriptors (20). The large 145 molecule test set (using PXR activators and PXR non-activators) resulted in overall prediction accuracies for RP, RF, SVM and the docking PXR logistic regression model of 63.4 %, 65.5 %, 66.9 % and 51 % respectively. (Table 1 and Supplemental Table 2). Interestingly RF performs better with the external test set than with internal testing for the training set, which would suggest that this model is not over-trained. Consensus scoring has been used previously with QSAR and docking applications (34, 55-57) and in some cases can improve predictions over individual methods (58,59). In the current study a combination of the models indicated that a consensus approach did not improve the overall prediction accuracy (data not shown).

The RP method produces more interpretable models unlike those obtained from RF and SVM. The RP tree from the PXR activator and PXR non-activator model is shown in Figure 1. In general, the descriptors important for PXR activators and PXR non-activators are W1, CW1, CW5, IW7, ID1, ID7 and molecular weight (MW). These represent the first 4 descriptors from the OH2 probe and the next two are from the DRY probe. W1 accounts for polarizability and dispersion forces, CW1 and CW5 represent the extent of hydrophilic regions per surface unit, IW7 is an integer moment and measure the imbalance between the center of mass of a molecule and hydrophilic regions around it, ID1 and ID7 are also integer moments which measure the imbalance between the center of mass of a molecule and hydrophobic regions around it. The RP tree can also be used to develop simple rules for separating PXR activators from PXR non-activators.

When comparing predictions for individual molecules, we made the following observations: a) members of the initial test set of activators such as diflubenzuron, is predicted as a non-activator by all the three machine learning methods in this study. Interestingly, the same molecule is also predicted as a PXR non-activator by the PNN and k-NN models in a previous study that used the same training set (20). b) fenbuconazole, another PXR activator from the previously used test set, is predicted as a PXR non-activator by all the models used previously (20), whereas the current RP and RF models correctly predicted it as a PXR activator in this study. The large test set used in this study contains 82 PXR activators and 63 non-activators and are predominantly a sub-set of available drugs (42), imidazole derivatives similar to clotrimazole, steroids (8), molecules with different heterocyclic ring systems (48) as well as many other diverse molecules. In this large test set, the PXR non-activators 5α-petromyzonol and 5β-cholan-3α,7α,12α,24-tetrol are predicted as PXR activators by all three QSAR methods in this study. The PXR activators 7-ketolithocholic acid and 12-ketolithocholic acid are predicted as PXR non-activators by all three QSAR methods. Interestingly, the unsubstituted compound lithocholic acid is correctly predicted as a PXR activator by all three methods. The generally low affinity of PXR activators raises the possibility that for some of the molecules in the published data, toxicity of the compounds may complicate determination of PXR activation. It may also be possible to improve the predictions for PXR activation using additional different descriptors and modeling approaches.

Previously we have used molecular similarity using the Tanimoto coefficient (60), or principal component analysis (PCA) to assess test and training set similarity (61). In our hands either of these approaches are useful to understand potential outlier predictions. In this study PCA was performed using VolSurf descriptors for the PXR data. The PCA score plot provides an estimate of the descriptor space of training and test set molecules (Figure 2) (61). Predictions can be considered as an extrapolation if the test set occupies a different descriptor space to the training set. It is clear from Figure 2 that the test set molecules occupy approximately the same descriptor space as that of the training set molecules and one would therefore expect the predictions to be relatively accurate. The first three principal components of the training and test sets explain 55.6 %, to 68.6 % of the variance, respectively. PCA analysis therefore did not explain the observed versus predicted PXR activation differences.

The logistic regression analysis method is a useful tool when the outcome of an experiment is binary (62), as in the current study either PXR activator or PXR non-activator. Four different models were generated by docking molecules to the four available co-crystal structures for PXR. The FlexX scores were used as the independent variable and the PXR activators or PXR non-activators (class variables) were used as dependent variables in the logistic regression analysis. Two different confusion matrices were built for each crystal structure i.e. one using only successfully docked compounds and the other from both successfully docked and failed compounds. In the second case, molecules which failed to dock inside the LBD of PXR were considered as non-activators. The logistic regression coefficients obtained from the FlexX scores were then used as the independent variable to predict activators/non-activators in the test set. The best model was obtained using the structure with SR12813 (PDB ID: 1NRL) which resulted in Eq 2 and was used to predict the large test set (Tables 1 and 2). The results for the other 3 complexes are not shown.

$$\text{Prob}\,(Y=A)=\frac{1}{1+e^{-(0.0398^{*}\text{FlexX}-0.374)}} \quad \text{or} \quad \log \quad \text{odds} \quad (Y=A)=-0.374-0.0398^{*}\text{FlexX} \tag{2}$$

Where, Prob and A represents probability and activator, respectively. The value of the FlexX score is negative and the associated coefficient is also negative (−0.0398) suggesting that the log odds (and, therefore, the probability) of a compound being a PXR activator decrease with the increase in FlexX score (positive value). In other words, for a one unit increase in FlexX score, the odds in favor of an activator are estimated to be decreased by a multiplicative factor of 0.961 (exp-0.0398 = 0.961). Figure 3 shows several representative molecules from the test set that were correctly predicted as PXR activators by all methods in this study. Apart from the SE which was higher for the test set than the machine learning methods, the performance of RF, SVM and RP appear to be better than the docking logistic regression model, based on the prediction accuracies, SP, Q and C of the large set (Table 1). The overall poor prediction accuracy of the docking results may be due to: a) poor performance of the scoring function in characterizing ligand-receptor interactions; b) treating the protein as rigid rather than flexible; and c) the overall promiscuous nature of the PXR LBD. It should also be noted that 9 out of 145 (6.2%) test set molecules failed to dock and were classed as non-activators, whereas predictions were made with the other machine learning models for all test set molecules.

Discussion

VolSurf descriptors have been widely used for drug discovery and ADME/Tox (37-39,63). In the course of this study we have developed novel predictive global models using VolSurf descriptors with human PXR activation classification data to identify compounds that are PXR activators. Such PXR activators may in turn lead to undesirable effects on drug and endogenous compound metabolism and transport. The machine learning models we have evaluated in the course of this work can also be used for virtual screening to identify molecules previously unrecognized as PXR activators that can then be selected for *in vitro* testing. We have used a similar approach previously with transporters (64,65). In this study we have also used a large test set of 145 molecules that have not been included in model building to validate all of the computational approaches. Our classification results were a significant improvement on those published previously (20). It is important to note that the new test set was also structurally diverse yet covered the same descriptor space as the training set as analyzed using PCA analysis, which reassures us that we were unlikely to be extrapolating far beyond the training set (61). Although the results of the large test set are not as good as those achieved for internal validation of the training set, a common weakness of many computational approaches, this could be partially due to the promiscuous nature of PXR. We would add that in contrast to the previous study by Ung et al (20) who used a test set of 15 PXR activators, we not only were able to predict more of these correctly, but we also further evaluated the model with a much larger test set and obtained predictions which we felt were acceptable considering the difficulty in predicting activation with this protein. The probability of randomly selecting a PXR activator or a PXR non-activator is 56.55 % and 43.44 %, respectively. Our SVM results are therefore better than random for the large test set as we achieved approximately 68% and 65% correct for PXR activator or PXR non-activator, respectively (Supplemental Table 2). These machine learning methods may be improved by using additional descriptor types which we are currently evaluating for inclusion in future studies.

We have also utilized a structure-based docking approach using FlexX combined with logistic regression for comparison purposes which performs poorly compared with SVM, RP and RF.

This is perhaps not surprising considering the flexibility of PXR and the differences in the ligand bound structures reported (9,10,53,54). In this case the final FlexX logistic regression model was based on 1NRL which had a resolution of 2.00Å (9). To our knowledge there have been no direct comparisons of docking, QSAR or machine learning methods for PXR with a large external test set containing diverse xenobiotics. Although a recent study has used both rigid and flexible docking to destabilize the key agonist-protein interactions for PXR in a small number of molecules, no binding scores or classification was reported (35). Assessment of other docking approaches in the future with a large test set like that used in the current analysis, is justified in an attempt to improve the results and at least match the standard set by the machine learning models described herein. With several PXR crystal structures in the PDB (9,10,53, 54), this study has highlighted the difficulty in using docking to predict a large set of molecules as potential agonists and suggests that classification models developed with machine learning methods (35) may have an important role alongside pharmacophores (8,14) and smaller "local" quantitative QSAR models in prioritizing compounds for *in vitro* testing. The combination of the training and test sets into a single larger model (322 molecules) would be worth assessing in the future with other descriptors and machine learning or docking methods and may lead to the creation of custom PXR scoring functions.

Although the docking approach was not ideal for classification it could still be useful in those cases were the predictions were correct to enable a qualitative evaluation of the likely ligand-protein interactions. This could assist in defining what molecular changes can be introduced chemically to reduce or avoid PXR activation. For example, others have suggested attaching hydrogen bonding groups on one of the hydrophobic features, adding larger more rigid groups as well as removing central H-bond acceptors (35). The large test set created in this study (Table 2) can be used to find structurally similar compounds that are either observed human PXR activators or PXR non-activators (regardless of their predicted activity), and provides further instruction as to what types of structural modification can decrease activity. While the test set contains numerous β-adrenergic blockers they are all PXR non-activators, and molecules in the calcium channel blocker and proton pump inhibitor series are all PXR activators. We were able to find some examples that had differing activity and visualize those that were correctly predicted as PXR activators based on the docking orientation. For example oxycodone and morphine are PXR non-activators while naloxone is a PXR activator (Figure 4). While there are several subtle differences between the three compounds, naloxone contains an additional ethylene moiety which may interact with a hydrophobic region in the LBD in addition to being a reactive functional group. Naproxen (PXR non-activator) and nabumetone (PXR activator) are non steroidal anti-inflammatory drugs, the former possesses a carboxylic acid group while the latter has a longer alkyl chain with a methyl substituted for the hydroxyl which may serve to preferentially position the C=O hydrogen bond acceptor (Figure 4). Nevirapine is a PXR non-activator while oxcarbazepine and carbamazepine are PXR activators. The latter two molecules feature carboxamide moieties which will serve as potential hydrogen bond acceptors while in the corresponding position nevirapine has a hydrophobic group (Figure 4). A further series of tricyclic compounds includes the PXR non-activators desipramine, doxepin and amitriptyline which contain a basic primary amine group at the end of an aliphatic chain, while the PXR activator loratadine has a neutral hydrogen bond acceptor in the corresponding position as well as ring substitutions. Interestingly the training set also contains additional tricyclic molecules (protriptyline, nortriptyline, trimipramine and clomipramine) that are PXR non-activators, yet structurally they closely resemble desipramine, doxepin and amitriptyline. These qualitative examples demonstrate some of the benefits of a large database of such binary classification information as it is valuable in biasing design of future molecules away from PXR activators that may be accomplished with the assistance of computational and *in vitro* methods.

In conclusion, we have developed new machine learning models for PXR that outperform those previously developed as well as docking models when tested with a new large external test set. We suggest it is likely quite difficult to definitively describe functional groups that will globally decrease affinity due to the size and flexibility of the human PXR LBD, rather, as in this and previous cases (35) it will be important to address structural series individually using all of the available computational and *in vitro* tools at our disposal. These approaches should also enable increased efficiency of screening and testing, aiding pharmaceutical research to avoid developing potent PXR agonists.

Acknowledgment

M.D.K. is supported by K08-GM074238 from the National Institutes of Health and a Competitive Medical Research Fund (CMRF) grant from the University of Pittsburgh Medical Center. S.E. gratefully acknowledges Dr. Maggie A.Z. Hupcey for support and discussion and congratulates Dr. Kenneth Bachmann (University of Toledo) on his recent retirement.

Abbreviations

ADME/Tox, Absorption, distribution, metabolism excretion/ Toxicology; LDB, Ligand binding domain; PCA, Principal Component analysis; PXR, Pregnane X Receptor; PDB, Protein data bank; QSAR, Quantitative Structure Activity Relationship; RF, Random Forest; RP, Recursive Partitioning; SVM, Support Vector Machine..

References

1. Bertilsson G, Heidrich J, Svensson K, Asman M, Jendeberg L, Sydow-Backman M, Ohlsson R, Postlind H, Blomquist P, Berkenstam A. Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction Proc Natl Acad Sci U S A 1998;95:12208–12213.
2. Blumberg B, Sabbagh W Jr. Juguilon H, Bolado J Jr. van Meter CM, Ong ES, Evans RM. SXR, a novel steroid and xenobiotic-sensing nuclear receptor. Genes Dev 1998;12:3195–3205. [PubMed: 9784494]
3. Kliewer SA, Moore JT, Wade L, Staudinger JL, Watson MA, Jones SA, McKee DD, Oliver BB, Willson TM, Zetterstrom RH, Perlmann T, Lehmann JM. An orphan nuclear receptor activated by pregnanes defines a novel steroid signalling pathway. Cell 1998;92:73–82. [PubMed: 9489701]
4. Goodwin B, Moore LB, Stoltz CM, McKee DD, Kliewer SA. Regulation of the human CYP2B6 gene by the nuclear pregnane X receptor. Mol Pharmacol 2001;60:427–431. [PubMed: 11502872]
5. Staudinger JL, Goodwin B, Jones SA, Hawkins-Brown D, MacKenzie KI, LaTour A, Liu Y, Klaassen CD, Brown KK, Reinhard J, Willson TM, Koller BH, Kliewer SA. The nuclear receptor PXR is a lithocholic acid sensor that protects against liver toxicity. Proc Natl Acad Sci U S A 2001;98:3369–3374. [PubMed: 11248085]
6. Synold TW, Dussault I, Forman BM. The orphan nuclear receptor SXR coordinately regulates drug metabolism and efflux. Nature Medicine 2001;7:584–590.
7. Harmsen S, Meijerman I, Beijnen JH, Schellens JH. The role of nuclear receptors in pharmacokinetic drug-drug interactions in oncology. Cancer Treat Rev 2007;33:369–380. [PubMed: 17451886]
8. Ekins S, Chang C, Mani S, Krasowski MD, Reschly EJ, Iyer M, Kholodovych V, Ai N, Welsh WJ, Sinz M, Swaan PW, Patel R, Bachmann K. Human pregnane X receptor antagonists and agonists define molecular requirements for different binding sites. Mol Pharmacol 2007;72:592–603. [PubMed: 17576789]
9. Watkins RE, Davis-Searles PR, Lambert MH, Redinbo MR. Coactivator binding promotes the specific interaction between ligand and the pregnane X receptor. J Mol Biol 2003;331:815–828. [PubMed: 12909012]
10. Watkins RE, Maglich JM, Moore LB, Wisely GB, Noble SM, Davis-Searles PR, Lambert MH, Kliewer SA, Redinbo MR. 2.1A crystal structure of human PXR in complex with the St John's Wort compound hyperforin. Biochemistry 2003;42:1430–1438. [PubMed: 12578355]

11. Watkins RE, Noble SM, Redinbo MR. Structural insights into the promiscuity and function of the human pregnane X receptor. Curr Opin Drug Discov Devel 2002;5:150–158.

12. Watkins RE, Wisely GB, Moore LB, Collins JL, Lambert MH, Williams SP, Willson TM, Kliewer SA, Redinbo MR. The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity. Science 2001;292:2329–2333. [PubMed: 11408620]

13. Xue Y, Moore LB, Orans J, Peng L, Bencharit S, Kliewer SA, Redinbo MR. Crystal structure of the pregnane X receptor-estradiol complex provides insights into endobiotic recognition. Mol Endocrinol 2007;21:1028–1038. [PubMed: 17327420]

14. Ekins S, Erickson JA. A pharmacophore for human pregnane-X-receptor ligands. Drug Metab Dispos 2002;30:96–99. [PubMed: 11744617]

15. Bachmann K, Patel H, Batayneh Z, Slama J, White D, Posey J, Ekins S, Gold D, Sambucetti L. PXR and the regulation of apoA1 and HDL-cholesterol in rodents. Pharmacol Res 2004;50:237–246. [PubMed: 15225665]

16. Schuster D, Laggner C, Steindl TM, Palusczak A, Hartmann RW, Langer T. Pharmacophore modeling and in silico screening for new P450 19 (aromatase) inhibitors. J Chem Inf Model 2006;46:1301–1311. [PubMed: 16711749]

17. Ekins S, Mirny L, Schuetz EG. A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXRa and LXRb. Pharm Res 2002;19:1788–1800. [PubMed: 12523656]

18. Jacobs MN. In silico tools to aid risk assessment of endocrine disrupting chemicals. Toxicology 2004;205:43–53. [PubMed: 15458789]

19. Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, Sorokina S, Bugrim A, Nikolskaya T. A Combined Approach to Drug Metabolism and Toxicity Assessment. Drug Metab Dispos 2006;34:495–503. [PubMed: 16381662]

20. Ung CY, Li H, Yap CW, Chen YZ. In silico prediction of pregnane X receptor activators by machine learning approaches. Mol Pharmacol 2007;71:158–168. [PubMed: 17003167]

21. Wang CY, Li CW, Chen JD, Welsh WJ. Structural model reveals key interactions in the assembly of the pregnane X receptor/corepressor complex. Mol Pharmacol 2006;69:1513–1517. [PubMed: 16452398]

22. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? Drug Discov Today 2006;11:700–707. [PubMed: 16846797]

23. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan O. A stepwise approach for defining the applicability domain of SAR and QSAR models. J Chem Inf Model 2005;45:839–849. [PubMed: 16045276]

24. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. J Chem Inf Comput Sci 2004;44:1912–1928. [PubMed: 15554660]

25. Muegge I, Enyedy IJ. Virtual screening for kinase targets. Curr Med Chem 2004;11:693–707. [PubMed: 15032724]

26. Kubinyi, H. Success stories of computer-aided design. In: Ekins, S., editor. Computer Applications in Pharmaceutical Research and Development. John Wiley and Sons; Hoboken: 2006. p. 377-424.

27. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol 2007;152:21–37. [PubMed: 17549046]

28. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 2007;152:9–20. [PubMed: 17549047]

29. Costache AD, Trawick D, Bohl D, Sem DS. AmineDB: large scale docking of amines with CYP2D6 and scoring for druglike properties--towards defining the scope of the chemical defense against foreign amines in humans. Xenobiotica 2007;37:221–245. [PubMed: 17624022]

30. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 2004;3:935–949. [PubMed: 15520816]

31. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Curr Opin Chem Biol 2006;10:194–202. [PubMed: 16675286]

32. Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. J Med Chem 2006;49:5851–5855. [PubMed: 17004700]

33. Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP. Comparison of automated docking programs as virtual screening tools. J Med Chem 2005;48:962–976. [PubMed: 15715466]

34. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42:5100–5109. [PubMed: 10602695]

35. Gao YD, Olson SH, Balkovec JM, Zhu Y, Royo I, Yabut J, Evers R, Tan EY, Tang W, Hartley DP, Mosley RT. Attenuating pregnane X receptor (PXR) activation: a molecular modelling approach. Xenobiotica 2007;37:124–138. [PubMed: 17484516]

36. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748. [PubMed: 9126849]

37. Doddareddy MR, Cho YS, Koh HY, Kim DH, Pae AN. In silico renal clearance model using classical Volsurf approach. J Chem Inf Model 2006;46:1312–1320. [PubMed: 16711750]

38. Cruciani G, Crivori P, Carrupt PA, Testa B. Molecular Fields in Quantitative Structure-Permeation Relationships:The VolSurf Approach. THEOCHEM 2000:17–30.

39. Ooms F, Weber P, Carrupt PA, Testa B. A simple model to predict blood-brain barrier permeation from 3D molecular fields. Biochim Biophys Acta 2002;1587:118–125. [PubMed: 12084453]

40. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 1999;37:228–241. [PubMed: 10584068]

41. Zhang T, Zhou JH, Shi LW, Zhu RX, Chen MB. 3D-QSAR studies with the aid of molecular docking for a series of non-steroidal FXR agonists. Bioorg Med Chem Lett 2007;17:2156–2160. [PubMed: 17307356]

42. Sinz M, Kim S, Zhu Z, Chen T, Anthony M, Dickinson K, Rodrigues AD. Evaluation of 170 xenobiotics as transactivators of human pregnane X receptor (hPXR) and correlation to known CYP3A4 drug interactions. Curr Drug Metab 2006;7:375–388. [PubMed: 16724927]

43. Faucette SR, Zhang TC, Moore R, Sueyoshi T, Omiecinski CJ, LeCluyse EL, Negishi M, Wang H. Relative activation of human pregnane X receptor versus constitutive androstane receptor defines distinct classes of CYP2B6 and CYP3A4 inducers. J Pharmacol Exp Ther 2007;320:72–80. [PubMed: 17041008]

44. Zhu Z, Kim S, Chen T, Lin JH, Bell A, Bryson J, Dubaquie Y, Yan N, Yanchunas J, Xie D, Stoffel R, Sinz M, Dickinson K. Correlation of high-throughput pregnane X receptor (PXR) transactivation and binding assays. J Biomol Screen 2004;9:533–540. [PubMed: 15452340]

45. Krasowski MD, Yasuda K, Hagey LR, Schuetz EG. Evolution of the pregnane X receptor: adaptation to cross-species differences in biliary bile salts. Mol Endocrinol 2005;19:1720–1739. [PubMed: 15718292]

46. Krasowski MD, Yasuda K, Hagey LR, Schuetz EG. Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). Nucl Recept 2005;3:2. [PubMed: 16197547]

47. Weininger D. SMILES 1. Introduction and encoding rules. J Chem Inf Comput Sci 1988;28:31.

48. Lemaire G, Benod C, Nahoum V, Pillon A, Boussioux AM, Guichou JF, Subra G, Pascussi JM, Bourguet W, Chavanieux A, Balaguer P. 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 2007;72:572–581. [PubMed: 17573484]

49. Clark MA, Cramer RD, van Op den Bosch N. Validation of the general purpose Tripos 5.2 force field. J Comput Chem 1989;10:982–1012.

50. Therneau, TM.; Atkinson, EJ. An introduction to recursive partitioning using the RPART routines, Department of health Sciences Research: Mayo clinic. 1997.

51. Liaw A, Wiener M. Classification and regression by random forest. R News 2002;2/3:18–22.

52. Hsu, C-W.; Chang, C-C.; Lin, C-J. A Practical Guide to Support Vector Classification. 2008. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdfhttp://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

53. Chrencik JE, Orans J, Moore LB, Xue Y, Peng L, Collins JL, Wisely GB, Lambert MH, Kliewer SA, Redinbo MR. Structural disorder in the complex of human pregnane X receptor and the macrolide antibiotic rifampicin. Mol Endocrinol 2005;19:1125–1134. [PubMed: 15705662]

54. Xue Y, Chao E, Zuercher WJ, Willson TM, Collins JL, Redinbo MR. Crystal structure of the PXR-T1317 complex provides a scaffold to examine the potential for receptor antagonism. Bioorg Med Chem 2007;15:2156–2166. [PubMed: 17215127]

55. Krovat EM, Langer T. Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. J Chem Inf Comput Sci 2004;44:1123–1129. [PubMed: 15154781]

56. Oloff S, Mailman RB, Tropsha A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. J Med Chem 2005;48:7322–7332. [PubMed: 16279792]

57. Zhang Q, Muegge I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. J Med Chem 2006;49:1536–1548. [PubMed: 16509572]

58. Ekins S, Durst GL, Stratford RE, Thorner DA, Lewis R, Loncharich RJ, Wikel JH. Three dimensional quantitative structure permeability relationship analysis for a series of inhibitors of rhinovirus replication. J Chem Inf Comput Sci 2001;41:1578–1586. [PubMed: 11749585]

59. So S-S, Karplus M. A comparitive study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. J Comp-Aided Mol Des 1999;13:243–258.

60. Ekins S, Balakin KV, Savchuk N, Ivanenkov Y. Insights for human Ether-a-Go-Go-Related Gene Potassium Channel inhibition using recursive partitioning, Kohonen and Sammon mapping Techniques. J Med Chem 2006;49:5059–5071. [PubMed: 16913696]

61. Khandelwal A, Bahadduri P, Chang C, Polli JE, Swaan P, Ekins S. Computational Models to Assign Biopharmaceutics Drug Disposition Classification from Molecular Structure. Pharm Res 2007;24:2249–2262. [PubMed: 17846869]

62. Jones DR, Ekins S, Li L, Hall SD. Computational approaches that predict metabolic intermediate complex formation with CYP3A4 (+b5). Drug Metab Dispos 2007;35:1466–1475. [PubMed: 17537872]

63. Berellini G, Cruciani G, Mannhold R. Pharmacophore, drug metabolism, and pharmacokinetics models on non-peptide AT1, AT2, and AT1/AT2 angiotensin II receptor antagonists. J Med Chem 2005;48:4389–4399. [PubMed: 15974591]

64. Chang C, Ekins S, Bahadduri P, Swaan PW. Pharmacophore-based discovery of ligands for drug transporters. Adv Drug Del Rev 2006;58:1431–1450.

65. Ekins S, Johnston JS, Bahadduri P, D'Souzza VM, Ray A, Chang C, Swaan PW. In Vitro And Pharmacophore Based Discovery Of Novel hPEPT1 Inhibitors. Pharm Res 2005;22:512–517. [PubMed: 15846457]

66. Lemaire G, Benod C, Nahoum V, Pillon A, Boussioux AM, Guichou JF, Subra G, Pascussi JM, Bourguet W, Chavanieux A, Balaguer P. 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. 2007

67. Maier A, Zimmermann C, Beglinger C, Drewe J, Gutmann H. Effects of budesonide on P-glycoprotein expression in intestinal cell lines. Br J Pharmacol 2007;150:361–368. [PubMed: 17179942]

68. Luo G, Cunningham M, kim S, Burn T, Lin J, Sinz M, Hamilton GA, Rizzo C, Jolley S, Gilbert D, Downey A, Mudra D, Graham R, Carroll K, Xie J, Madan A, Parkinson A, Christ D, Selling B, LeCluyse EL, Gan L-S. CYP3A4 induction by drugs: correlation between a pregnane X receptor reporter gene assay and CYP3A4 expression in human hepatocytes. Drug Metab Dispos 2002;30:795–804. [PubMed: 12065438]

69. Hartley DP, Dai X, Yabut J, Chu X, Cheng O, Zhang T, He YD, Roberts C, Ulrich R, Evers R, Evans DC. Identification of potential pharmacological and toxicological targets differentiating structural analogs by a combination of transcriptional profiling and promoter analysis in LS-180 and Caco-2 adenocarcinoma cell lines. Pharmacogenet Genomics 2006;16:579–599. [PubMed: 16847427]

70. Lindley C, Hamilton G, McCune JS, Faucette S, Shord SS, Hawke RL, Wang H, Gilbert D, Jolley S, Yan B, LeCluyse EL. The effect of cyclophosphamide with and without dexamethasone on cytochrome P450 3A4 and 2B6 in human hepatocytes. Drug Metab Dispos 2002;30:814–822. [PubMed: 12065440]

71. Chang TK, Waxman DJ. Synthetic drugs and natural products as modulators of constitutive androstane receptor (CAR) and pregnane X receptor (PXR). Drug Metab Rev 2006;38:51–73. [PubMed: 16684648]

72. Duret C, Daujat-Chavanieu M, Pascussi JM, Pichard-Garcia L, Balaguer P, Fabre JM, Vilarem MJ, Maurel P, Gerbal-Chaloin S. Ketoconazole and miconazole are antagonists of the human glucocorticoid receptor: consequences on the expression and function of the constitutive androstane receptor and the pregnane X receptor. Mol Pharmacol 2006;70:329–339. [PubMed: 16608920]

73. Prueksaritanont T, Richards KM, Qiu Y, Strong-Basalyga K, Miller A, Li C, Eisenhandler R, Carlini EJ. Comparative effects of fibrates on drug metabolizing enzymes in human hepatocytes. Pharm Res 2005;22:71–78. [PubMed: 15771232]

74. Drocourt L, Pascussi JM, Assenat E, Fabre JM, Maurel P, Vilarem MJ. Calcium channel modulators of the dihydropyridine family are human pregnane X receptor activators and inducers of CYP3A, CYP2B, and CYP2C in human hepatocytes. Drug Metab Dispos 2001;29:1325–1331. [PubMed: 11560876]

75. Ogino M, Nagata K, Yamazoe Y. Selective suppressions of human CYP3A forms, CYP3A5 and CYP3A7, by troglitazone in HepG2 cells. Drug Metab Pharmacokinet 2002;17:42–46. [PubMed: 15618651]

76. Cerveny L, Svecova L, Anzenbacherova E, Vrzal R, Staud F, Dvorak Z, Ulrichova J, Anzenbacher P, Pavek P. Valproic acid induces CYP3A4 and MDR1 gene expression by activation of constitutive androstane receptor and pregnane X receptor pathways. Drug Metab Dispos 2007;35:1032–1041. [PubMed: 17392393]

77. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–1612. [PubMed: 15264254]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

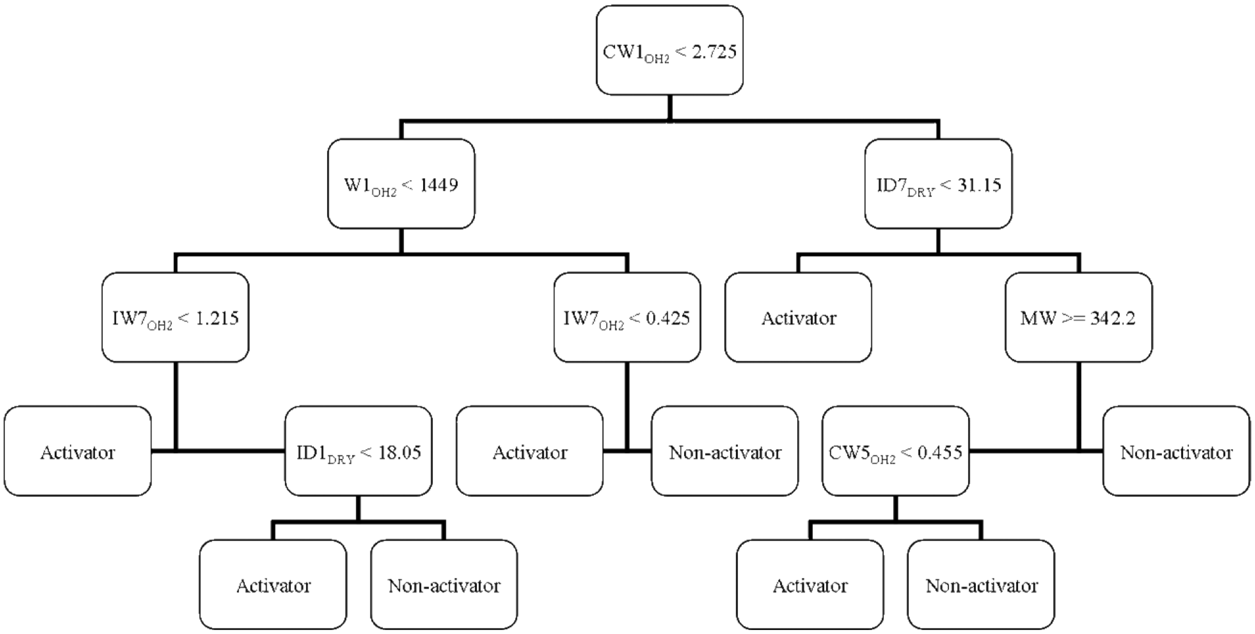

Figure 1.

The recursive partitioning tree model separating PXR activators and PXR non-activators developed in this study. The tree method is described in the Materials and Methods. Molecules are assigned to branches when descriptor values are greater than or less than a threshold value described in the preceding node. The subscripts represent the OH2 and dry probe used in VolSurf calculations.

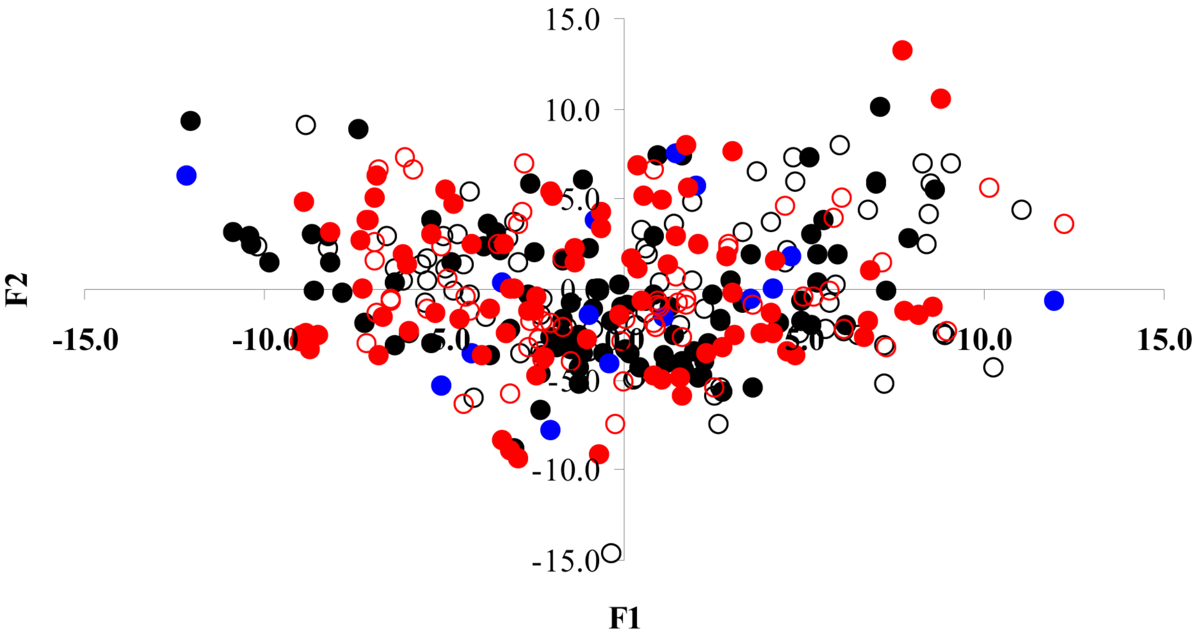

Figure 2.
The principal component analysis scores plot of PXR training (20) and test and validation sets using the VolSurf descriptors. The PXR activators are shown as filled circles training (black), small test set ((20), blue) and large test set (red), whereas, the non-activators are shown as empty circles training (black) and large test set (red). This orthogonal linear transformation, transforms the data to a new coordinate system. The x and y axes represent PCA scores from first and second principal components, respectively. The first three principal components of the training and test sets explain 55.6 %, to 68.6 % of the variance, respectively.

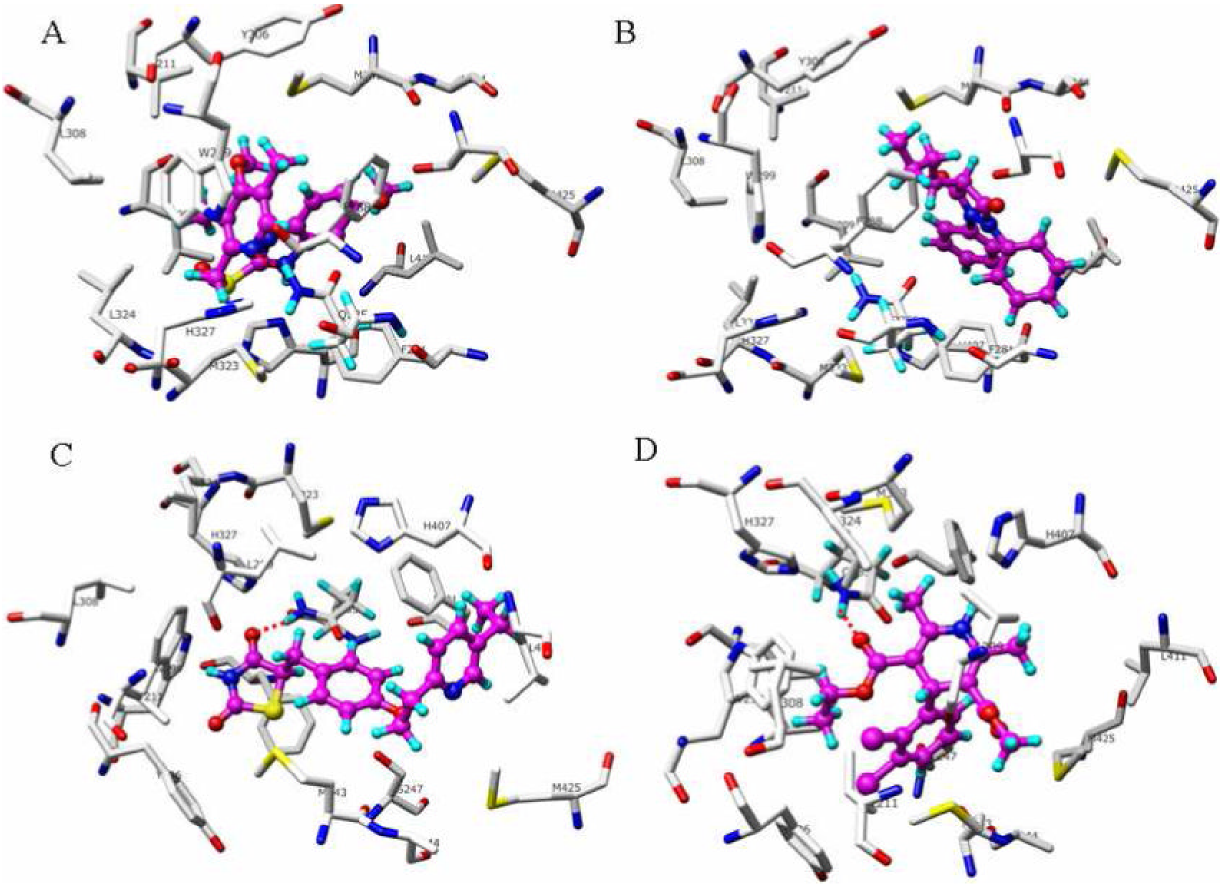

Figure 3.

The FlexX docked (1NRL) conformations of selected correctly predicted PXR activators - omeprazole (A), phenylbutazone (B), pioglitazone (C), and felodipine (D). Omeprazole, pioglitazone and felodipine are predicted to form hydrogen bonding interactions with the side chain NH of Gln-285. The hydrogen bonding interaction is shown as red dotted lines. The amino acid residues are shown in the stick mode whereas the PXR activators are shown in the ball and stick mode (purple). The 3D images were created using Chimera.(77)

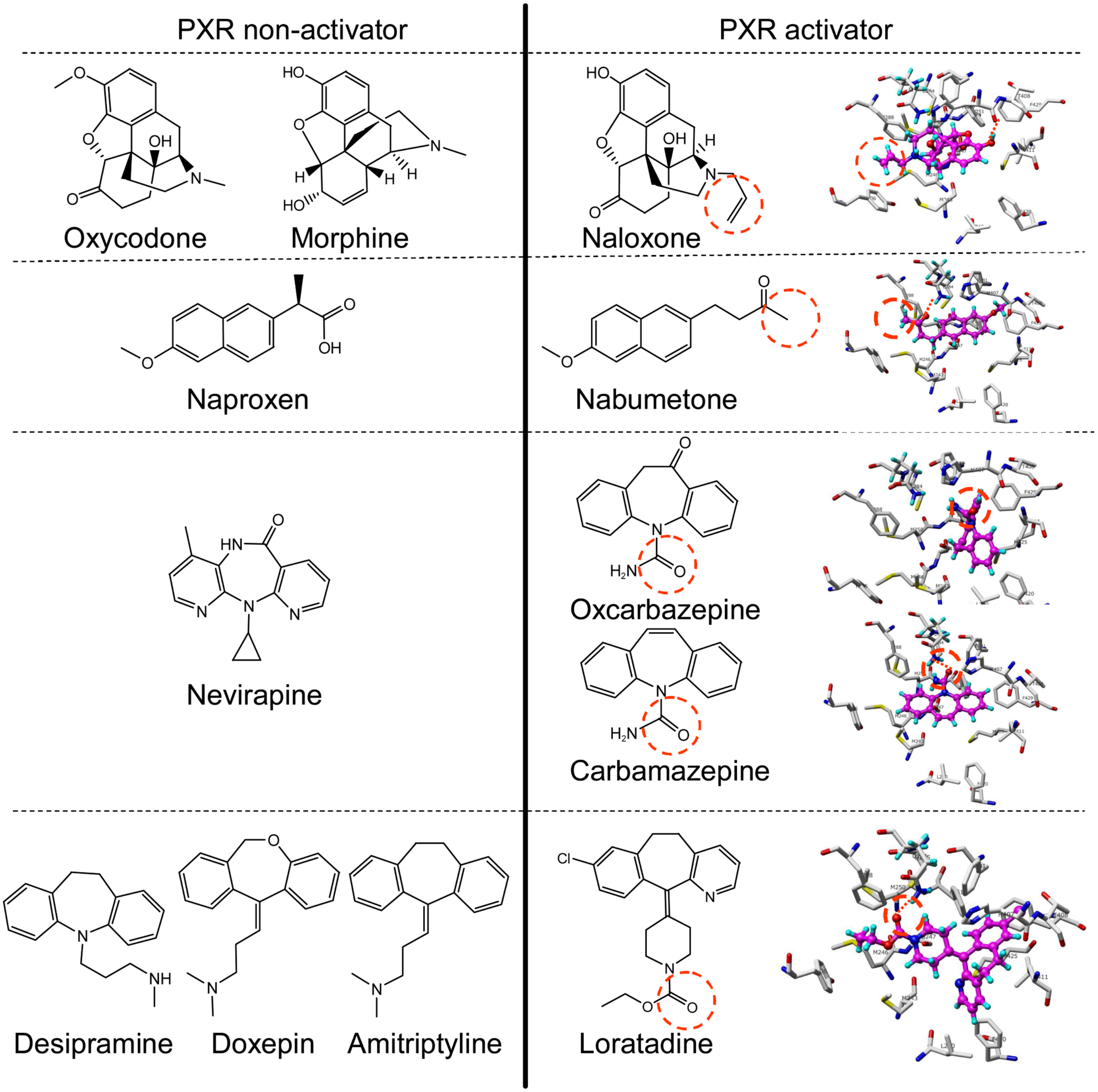

Figure 4.

Examples of PXR non-activator and PXR activator compounds in the test set that are structurally similar. Red dashed circles highlight likely important structural features that may contribute to activity. Corresponding PXR activator compounds (that had correctly predicted classifications) were docked in the PXR structure (1NRL) with FlexX (as described in the Materials and Methods). Carbamazepine (circled group), loratadine (circled group) and nabumetone (C=O group next to circled group) are predicted to form hydrogen bonding interactions with the side chain NH of Gln-285, additionally the methyl group in nabumetone may interact with the Phe288. Naloxone is predicted to form a H-bond with the side chain C=O and His407 and the ethylene group may also have a hydrophobic interaction with the Phe288. Oxcarbazepine docks differently to Carbamazepine and may also form a H-bond between the side chain C=O and His407. The hydrogen bonding interactions are shown as a red dotted line. The amino acid residues are shown in the stick mode whereas the PXR activators are shown in the ball and stick mode (purple). The 3D images were created using Chimera.(77)

Table 1

The predictive performance of RP, RF, SVM and docking with logistic regression for the training set (n = 177, 10-fold cross-validation study) and test set (n = 145). The values in the parenthesis (docking and logistic regression) include both failed and successfully docked compounds. RP = recursive partitioning; RF = random forest; SVM = support vector machine

Method	% SE[a]	% SP[b]	% Q[c]	C[d]
Training				
RP	95.92	77.21	87.57	0.754
RF	82.65	62.02	73.45	0.459
SVM	98.98	88.61	94.35	0.888
Docking + logistic regression	60.0 (45.92)	40.28 (45.57)	50.34 (45.76)	0.002 (−0.085)
Test set				
RP	64.63	61.9	63.45	0.264
RF	64.63	66.67	65.52	0.310
SVM	68.29	65.08	66.9	0.332
Docking + logistic regression	73.42 (70.73)	17.54 (25.4)	50.0 (51.03)	−0.106 (−0.043)

True positive (TP), true negative (TN), false positive (FP), false negative (FN), sensitivity (SE), specificity (SP), overall prediction accuracy (Q) and Matthew's correlation coefficient (C).

[a] SE=TP/(TP+FN)

[b] SP=TN/(TN+FP)

[c] Q=(TP+TN)/(TP+TN+FP+FN)

[d] $C = ((TP^{*}TN) - (FN^{*}FP)) / \sqrt{(TP+FN)\,(TP+FP)\,(TN+FN)\,(TN+FP)}$

Table 2

Prediction results for the large 145 molecule test set. Classifications: A = PXR activator (agonist); N = PXR non-activator. RP = recursive partitioning; RF = random forest; SVM = support vector machine

Compounds	**Classification**	**Docking + logistic regression**	**RP**	**RF**	**SVM**	**Reference / EC$_{50}$ value determined in this study (μM)**
1,9-Dideoxydforskolin	A	N	A	A	A	(42)
Amlodipine	A	A	A	A	N	(42)
Bergamottin	A	N	A	A	A	(42)
Celecoxib	A	A	A	N	N	(42)
Diltiazem	A	A	A	N	A	(42)
Felodipine	A	A	A	A	A	(42)
Fluvastatin	A	A	N	A	A	(42)
Forskolin	A	A	A	A	A	(42)
Glimepiride	A	N (failed)	N	A	A	(42)
Haloperidol	A	A	A	N	N	(42)
Lansoprazole	A	A	A	N	N	(42)
Mevastatin	A	N	A	A	A	(42)
Montelukast	A	A	N	A	A	(42)
Omeprazole	A	A	A	A	A	(42)
Phenylbutazone	A	A	A	A	A	(42)
Pioglitazone	A	A	A	A	A	(42)
Rabeprazole	A	A	A	A	N	(42)
Reserpine	A	N	A	N	N	(42)
Rifabutin	A	N	A	A	A	(42)
Rifapentine	A	N (failed)	A	N	N	(42)
CDD3508	A	A	A	A	A	(8)
CDD3532	A	A	A	A	A	(8)
CDD3538	A	A	A	A	A	(8)
CDD3543	A	A	A	A	A	(8)
CDD3501	A	A	A	A	A	(8)
CDD3530	A	A	A	A	A	(8)
CDD3536	A	A	A	A	A	(8)
CDD3540	A	A	A	A	A	(8)
5β-Androstan-3α-ol	A	A	N	A	A	(8)
Petromyzonol	N	N	A	A	A	(8)
17β-Dihydroandrosterone	A	N	A	A	A	(8)
Dihydrotestosterone	A	N	A	A	A	(8)
Etiocholanolone	A	A	A	A	A	(8)
Taurochenodeoxycholic acid	N	A	A	N	N	(8)
Lithocholic acid	A	N	A	A	A	(8)
7-Ketolithocholic acid	A	A	N	N	N	(8)
12-Ketolithocholic acid	A	N	N	N	N	(8)
Estrone	A	A	A	A	A	(8)
Estriol	A	N	N	A	A	(8)
Allocholic acid	N	N	A	N	N	(8)
Cortolone	A	A	A	A	A	(8)
Estetrol	A	N	N	A	A	(8)
Epitestosterone sulfate	A	N	N	A	A	(8)
5β-Pregnan-3α,20α-diol	A	N	A	A	A	(8)
5α-Androstan-3α-ol	A	N	A	A	A	(8)
Lithocholicacid acetate	A	N (failed)	A	A	A	(8)
5β-Cholan-3α,7α,12α,24-tetrol	N	N	A	A	A	(8)
ω-Muricholicacid	A	N	A	A	A	(8)
16,(5α)-Androsten-3β-ol	A	N	A	A	A	(8)

Compounds	Classification	Docking + logistic regression	RP	RF	SVM	Reference / EC$_{50}$ value determined in this study (μM)
Tauro-β-muricholic acid	N	N	A	N	N	(8)
C2BA-10	A	A	A	N	N	(66)
C2BA-11	N	A	A	A	A	(66)
C2BA-12	N	A	A	A	A	(66)
C2BA-13	A	A	N	A	N	(66)
C2BA-248	A	A	N	A	N	(66)
C2BA-251	A	A	N	A	A	(66)
C2BA-3	N	A	A	A	N	(66)
C2BA-5	A	A	A	A	A	(66)
C2BA-6	A	A	A	N	A	(66)
C2BA-7	A	A	A	A	A	(66)
C2BA-8	A	A	A	A	A	(66)
C2BA-9	N	A	A	N	N	(66)
Budesonide	A	A	A	A	A	(67)
Carbamazepine	A	A	N	N	A	(68)
Chlorpromazine	N	A	N	N	N	(69)
Cyclophosphamide	A	N	A	A	A	(70)
Efavirenz	A	A	N	N	N	(71)
Etoposide	A	N	N	A	A	(71)
Fluconazole	N	A	N	N	N	(72)
Gemfibrozil	N	A	N	N	N	(73)
Ketoconazole	N	N	N	A	A	(71)
Nicardipine	A	A	A	N	N	(74)
Rosiglitazone	A	A	A	N	A	(75)
Topotecan	A	N	N	N	A	(71)
Valproic acid	A	N	N	A	A	(76)
Bupropion	A	A	N	N	N	(42)
Diclofenac	A	A	N	N	A	(42)
Flutamide	A	A	N	N	N	(42)
Isotretinoin	A	A	N	A	A	(42)
Loratadine	A	A	A	N	N	(42)
Meclizine	A	A	A	N	N	(42)
Meloxicam	A	A	A	A	A	(42)
Midazolam	A	A	A	N	N	(42)
Nabumetone	A	A	A	A	A	(42)
Naloxone	A	A	N	N	N	(42)
Ondansetron	A	A	N	N	N	(42)
Quinapril	A	A	A	N	N	(42)
Raloxifene	A	A	A	N	A	(42)
Sildenafil	A	A	N	N	N	(42)
Simvastatin	A	N	A	A	A	(42)
Terbinafine	A	A	N	N	N	(42)
Triazolam	A	A	N	N	N	(42)
Zolpidem	A	A	A	A	A	(42)
Acetaminophen	N	A	A	A	A	(42) and Inactive this study
Acyclovir	N	A	N	N	N	(42)
Albuterol	N	A	N	N	N	(42)
Amitriptyline	N	A	N	N	N	(42)
Atenolol	N	A	N	N	N	(42)
Azithromycin	N	N (failed)	N	N	A	(42)
Captopril	N	A	N	N	N	(42)
Cimetidine	N	A	A	A	A	(42)

Compounds	Classification	Docking + logistic regression	RP	RF	SVM	Reference / EC_{50} value determined in this study (μM)
Ciprofloxacin	N	A	A	N	N	(42)
Clarithromycin	N	N (failed)	A	A	N	(42)
Dapsone	N	A	A	A	N	(42)
Desipramine	N	A	N	N	N	(42)
Doxazosin	N	N	N	N	N	(42)
Doxepin	N	A	N	N	N	(42)
Doxorubicin	N	A	N	N	N	(42)
Enalapril	N	N	A	A	A	(42)
Erythromycin	N	N (failed)	A	A	A	(42)
Ethosuximide	N	A	N	N	A	(42)
Etodolac	N	A	N	N	A	(42)
Famotidine	N	A	N	N	N	(42)
Fexofenadine	N	N (failed)	A	N	N	(42)
Furosemide	N	A	N	N	N	(42)
Ibuprofen	N	A	N	A	A	(42)
Lamotrigine	N	A	A	N	N	(42)
Levofloxacin	N	A	N	N	N	(42)
Linezolid	N	A	N	A	A	(42)
Lisinopril	N	N (failed)	N	N	N	(42)
Metoprolol	N	A	N	N	N	(42)
Metronidazole	N	A	N	N	N	(42)
Morphine	N	A	N	N	N	(42)
Nadolol	N	A	N	N	A	(42)
Naproxen	N	A	N	A	A	(42)
Nevirapine	N	A	A	A	A	(42)
Paroxetine	N	A	N	N	N	(42)
Propranolol	N	A	N	N	N	(42)
Ranitidine	N	N	N	N	N	(42)
Risperidone	N	A	A	N	N	(42)
Sumatriptan	N	A	N	N	N	(42)
Tacrine	N	A	A	A	A	(42)
Terazosin	N	A	A	A	A	(42)
Theophylline	N	A	A	A	N	(42)
Timolol	N	A	N	N	N	(42)
Tocainide	N	A	N	N	N	(42)
Tolmetin	N	A	N	N	N	(42)
Valsartan	N	A	N	N	A	(42)
Venlafaxine	N	N	N	N	N	(42)
Zomitriptan	N	A	N	N	N	(42)
Cholecalciferol	N	N (failed)	A	A	A	Inactive this study
Flurbiprofen	A	A	N	N	N	80.1 μM this study
Mycophenolic acid	N	N	N	A	A	Inactive this study
Oxcarbazepine	A	A	N	A	A	18.1 μM this study
Oxycodone	N	A	N	N	N	Inactive this study