

# Challenges Predicting Ligand-Receptor Interactions of Promiscuous Proteins: The Nuclear Receptor PXR

Sean Ekins<sup>1,2,3\*</sup>, Sandhya Kortagere<sup>4</sup>, Manisha Iyer<sup>5</sup>, Erica J. Reschly<sup>5</sup>, Markus A. Lill<sup>6</sup>, Matthew R. Redinbo<sup>7,8,9</sup>, Matthew D. Krasowski<sup>5</sup>

**1** Collaborations in Chemistry, Jenkintown, Pennsylvania, United States of America, **2** Department of Pharmaceutical Sciences, University of Maryland, Baltimore, Maryland, United States of America, **3** Department of Pharmacology, University of Medicine & Dentistry of New Jersey (UMDNJ)-Robert Wood Johnson Medical School, Piscataway, New Jersey, United States of America, **4** Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, Pennsylvania, United States of America, **5** Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **6** Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, Indiana, United States of America, **7** Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America, **8** Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America, **9** The Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

## Abstract

Transcriptional regulation of some genes involved in xenobiotic detoxification and apoptosis is performed via the human pregnane X receptor (PXR) which in turn is activated by structurally diverse agonists including steroid hormones. Activation of PXR has the potential to initiate adverse effects, altering drug pharmacokinetics or perturbing physiological processes. Reliable computational prediction of PXR agonists would be valuable for pharmaceutical and toxicological research. There has been limited success with structure-based modeling approaches to predict human PXR activators. Slightly better success has been achieved with ligand-based modeling methods including quantitative structure-activity relationship (QSAR) analysis, pharmacophore modeling and machine learning. In this study, we present a comprehensive analysis focused on prediction of 115 steroids for ligand binding activity towards human PXR. Six crystal structures were used as templates for docking and ligand-based modeling approaches (two-, three-, four- and five-dimensional analyses). The best success at external prediction was achieved with 5D-QSAR. Bayesian models with FCFP\_6 descriptors were validated after leaving a large percentage of the dataset out and using an external test set. Docking of ligands to the PXR structure co-crystallized with hyperforin had the best statistics for this method. Sulfated steroids (which are activators) were consistently predicted as non-activators while, poorly predicted steroids were docked in a reverse mode compared to 5 $\alpha$ -androstan-3 $\beta$ -ol. Modeling of human PXR represents a complex challenge by virtue of the large, flexible ligand-binding cavity. This study emphasizes this aspect, illustrating modest success using the largest quantitative data set to date and multiple modeling approaches.

**Citation:** Ekins S, Kortagere S, Iyer M, Reschly EJ, Lill MA, et al. (2009) Challenges Predicting Ligand-Receptor Interactions of Promiscuous Proteins: The Nuclear Receptor PXR. *PLoS Comput Biol* 5(12): e1000594. doi:10.1371/journal.pcbi.1000594

**Editor:** James M. Briggs, University of Houston, United States of America

**Received:** June 30, 2009; **Accepted:** November 3, 2009; **Published:** December 11, 2009

**Copyright:** © 2009 Ekins et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** MDK was supported by National Institutes of Health Grant K08-GM074238 and the Competitive Medical Research Fund from the University of Pittsburgh Medical Center. SK is supported by American Heart Association grant 09SDG2110054. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [ekinssean@yahoo.com](mailto:ekinssean@yahoo.com)

‡ Current address: Department of Pathology, University of Iowa Hospitals and Clinics, Iowa City, Iowa, United States of America

## Introduction

Promiscuous proteins generally bind a large array of diverse ligand structures. These proteins include enzymes like cytochrome P450s (e.g. CYP3A4, EC 14.13.97), transporters such as P-glycoprotein (ABCB1), the human ether-a-go-go related gene (hERG, K<sub>v</sub>11.1) potassium channel and nuclear hormone receptors (NHRs) such as the pregnane X receptor (PXR; NR1I2; also known as SXR or PAR) [1]. This promiscuous binding may be facilitated by a very large binding site, multiple (overlapping) binding sites, or a flexible binding site that can adjust to the size of the ligand. Intrinsic disorder in the protein may also have a role [2,3]. These proteins described above are also particularly important as xenobiotic sensors and represent key mechanisms to respond to toxic stress.

The human PXR [4–6] transcriptionally regulates genes involved in xenobiotic metabolism and excretion, as well as other

cellular processes such as apoptosis [7–11]. Human PXR has a very broad specificity for ligands as exemplified by the structurally diverse array of activators including endogenous (bile acids, steroid hormones, fat-soluble vitamins) and exogenous (prescription and herbal drugs, and environmental chemicals) compounds. Activation of human PXR can cause drug-drug interactions [4,5] or result in physiological effects ranging from ameliorating cholestatic injury to the liver, altering bone homeostasis, and causing cell proliferation [12]. As PXR represents a potential target for pharmacologic modulation in disease, it is therefore becoming even more important to develop methods that can identify whether a molecule is likely to be a PXR agonist [13]. Currently there are five high-resolution crystal structures of human PXR [14–18] available in the Protein Data Bank (PDB) (and another structure to be deposited [19]). The structures have provided atomic level details that have led to a greater understanding of the

## Author Summary

Promiscuous proteins generally bind a large array of diverse ligand structures. This may be facilitated by a very large binding site, multiple binding sites, or a flexible binding site that can adjust to the size of the ligand. These aspects also increase the complexity of predicting whether a molecule will bind or not to such proteins which frequently function as exogenous compound sensors to respond to toxic stress. For example, transporters may prevent absorption of some molecules, and enzymes may convert them to more readily excretable compounds (or alternatively activate them prior to further clearance by other detoxification enzymes). Nuclear hormone receptors may respond to ligands and then affect downstream gene expression to upregulate both enzymes and transporters to increase the clearance for the same or different molecules. We have assessed the ability of many different ligand-based and structure-based computational approaches to model and predict the activation of human PXR by steroidal compounds. We find the most effective computational approach to identify potential steroidal PXR agonists which are clinically relevant due to their widespread use in clinical medicine and the presence of mimics in the environment.

ligand binding domain (LBD) and the structural features involved in ligand-receptor interactions [9,10,12–15]. The co-crystallized ligands include the natural products hyperforin (active component of the herbal anti-depressant St. John's wort) and colupulone (from hops), the steroid 17 $\beta$ -estradiol, the synthetic compounds SR12813, T1317 and the antibiotic rifampicin. These ligands span a range of molecular sizes (M.Wt range 272.38 – 713.81Da, mean 487.58 $\pm$ 147.25Da, Table S1) and are predicted as generally hydrophobic (calculated ALogP [20] 3.54–10.11, mean 5.54 $\pm$ 2.41). The cavernous ligand binding pocket (LBP) with a volume >1350  $\text{\AA}^3$  accepts molecules of these widely varying dimensions and chemical properties, and is likely capable of binding small molecules in multiple orientations [21]. This complicates overall prediction of whether a small molecule is likely to be classified as a PXR agonist using traditional structure-based virtual screening methods like docking [13,22]. With regard to this, we have previously shown that the widely used structure-based docking methods FlexX and GOLD performed relatively poorly in predicting human PXR agonists [7,16] and this is perhaps not surprising based on the observations described above.

An alternative method, which has been found to be valuable elsewhere in drug discovery, particularly when there may not be an available crystal structure of the target protein, uses a ligand-based approach. In this case a series of small molecule structures with PXR agonist activity data can be used to facilitate a structure activity relationship (SAR). When the biological activity data is continuous this will enable a quantitative structure activity relationship (QSAR) [23–25]. One widely used computational technology produces pharmacophores [20–23], which represent models that encode the key chemical features important for biological activity. Human PXR agonist pharmacophore models have been shown to possess hydrophobic, hydrogen bond acceptor and hydrogen bond donor features, consistent with the crystallographic structures of human PXR ligand-receptor complexes [26–29]. These pharmacophore models have predominantly used structurally diverse ligands in the training set and have the limitation in most cases of compiling data from multiple laboratories using different experimental protocols, ultimately

forcing binary classifications of ligands for the training sets (i.e., activating versus non-activating). Most of the models so far use EC<sub>50</sub> data, a measure of receptor transactivation. Although binding assays have been done with human PXR, they are problematic given the low affinity of most PXR activators. As a result, there is little radioligand binding data in the literature other than competition experiments with radiolabeled SR12813.

To date there have been few attempts to build ligand-based models around a large structurally narrow set of PXR activators. The absence of large sets of quantitative data for PXR agonists has restricted QSAR models to a relatively small universe of molecules compared to the known drugs, drug-like molecules, endobiotics and xenobiotics in general [30]. The PXR data limitation has resulted in the use of various machine learning methods (e.g support vector machine, recursive partitioning etc.) when the biological data is binary in nature (e.g. activating or binding versus non-activating / non-binding) [13,22,25,30].

As part of an ongoing analysis of NHRs [25–28], we have generated a large cadre of experimental data for classes of steroidal compounds, namely androstanes, estratrienes, pregnanes and bile acids/salts [31]. The advantages of using steroidal compounds for QSAR are that they are amenable to common alignments based on the steroidal backbone. For, example steroids represented the first datasets used for comparative molecular fields analysis (CoMFA) [32] and have been widely used as a benchmark for other methods such as comparative molecular similarity analysis (CoMSIA)[33]. Pharmacophore methods, in contrast, generally do not require the rigid alignment methods and have found use with more diverse structure sets [28,31]. Using this large quantitative data set of PXR activators, we applied various ligand-based computational methods including Bayesian modeling with 2D fingerprints. We also compared the results from QSAR approaches to molecular docking into the six available human PXR crystal structures.

Modeling of a broad specificity receptor such as PXR represents a challenge for *in silico* modeling and it is invaluable to know what approaches prove successful, if any. Ideally, these methods will also translate to modeling approaches for other broad specificity enzymes, transporters and ion channels [1], or other promiscuous proteins [34]. We are not aware of any similar studies using a comparative approach to predicting ligand-protein interactions for promiscuous proteins. This study also provides further insights into PXR-steroid interactions which have not been well studied [19] and is clinically relevant due to the widespread use of steroidal compounds and steroid mimics (e.g. oral contraceptives [35], for inflammation and as cancer treatments etc.) in clinical medicine [36], as well as the increasing problem of environmental contamination by endocrine disruptors [24].

## Results

### Docking

All compounds shown in Table S2 were docked to the six human PXR crystal structures using GOLD which we have used previously for docking diverse compounds into the human PXR structure [22]. All six crystal structures superimposed with a backbone root mean squared deviation of 0.5  $\text{\AA}$  suggesting that they had very similar structures and their co-crystallized ligands bound to the same binding pocket (Figure S1). The docking scores for all the compounds (Table S2) were in the range of 36 to 77 for all the crystal structures and their corresponding Tanimoto similarity scores to 5 $\alpha$ -androstan-3 $\beta$ -ol and the crystal ligand 17 $\beta$ -estradiol using MDL public keys were between 0.4 and 1.

To evaluate docking results, we compared docking scores for classifying compounds as activators or non-activators of PXR.

Using an  $EC_{50}$  value of 10  $\mu\text{M}$  as a cutoff the compounds listed in Table S2 were classified as activators (30 compounds) and non-activators (89 compounds). These results were compared to the classification obtained from the docking studies. The overall accuracy (Q values) were in the range of 35 to 55 % for models that used  $5\alpha$ -androstane- $3\beta$ -ol based similarity scores as weights to the goldscore, while the Q values were in the range of 47 to 58% for models that were generated with goldscores weighted with  $17\beta$ -estradiol based similarity scores (Table 1). The Matthews coefficient C showed a modest prediction rate with the best score for docking of compounds to PXR crystal structure 1M13. Further changing the cutoff values to either 100  $\mu\text{M}$  or 40  $\mu\text{M}$  did not improve the prediction rates. The Q value for a model computed by averaging all the models with  $5\alpha$ -androstane- $3\beta$ -ol weighted goldscore was 46% and for the average model with  $17\beta$ -estradiol the weighted goldscore was 51%.

Although the overall performance of docking produced rather modest results for classification the results for individual classes of compounds was better than average. In the best classification model (compounds docked to crystal structure 1M13 and weighted with  $17\beta$ -estradiol based similarity scores), 20 out of 30 PXR activators and 49 out of 89 non-activators were predicted correctly. Among the androstanes, 6 out of 11 compounds were predicted correctly as activators and 9 out of 14 compounds were classified as non-activators (Table S2). Among the bile salts, all 4 activators and 22 out of 46 non-activators were predicted correctly. Among the estratrienes, 5 out of 7 activators were predicted correctly, while the 4 non-activators were predicted as activators (Table S2). The reason for this mis-classification was due to the high similarity scores of the estrogens with  $17\beta$ -estradiol. In the pregnane class, 4 out of 7 activators and 16 out of 20 non-activators were correctly classified (Table S2). Some examples of molecules in their binding modes with PXR structure 1M13 are shown in Figure 1.

## 2D-Classification: Bayesian Models

All 115 compounds shown in Table S2 were used to generate a Bayesian classification model [37], using a definition of active as a compound having an  $EC_{50}$  for PXR activation of less than 10  $\mu\text{M}$ . Using molecular function class fingerprints of maximum

diameter 6 (FCFP\_6) and 8 interpretable descriptors (AlogP, molecular weight, rotatable bonds, number of rings, number of aromatic rings, hydrogen bond acceptor, hydrogen bond donor and polar surface area) a model was developed with a receiver operator curve (ROC) statistic for leave one out cross validation of 0.84. In addition to the leave one out cross validation, further validation methods were undertaken. After leaving 20% of the compounds out 100 times the ROC is  $0.84\pm 0.08$ , concordance  $73.2\% \pm 8.94$ , specificity  $69.14\% \pm 12.12$ , and sensitivity  $84.11\% \pm 18.04$ . The Bayesian method appears to have good model statistics for internal cross validation of steroids. These statistics suggest the model is stable and not over-trained as the ROC values are essentially identical to that obtained with leave one out cross validation.

We have additionally used this model to classify a previously used diverse molecule test set [13,22]. After removing the steroids from the test set, the Bayesian PXR model was used to rank 123 molecules (65 activators and 58 non activators). Out of the top 30 molecules scored and ranked with this model 20 (75%) were classified as activators ( $EC_{50} < 100 \mu\text{M}$ ) (Table S3). Even though the cutoff for activity for the Bayesian model is more stringent it still appears to be able to predominantly pick out the key molecular features that contribute to activity in non-steroidal compounds.

The Bayesian model with FCFP\_6 descriptors also enabled the visualization of substructure fingerprints (Figure 2) that either contributed positively or negatively to the activity classification. It appears that all positive contributing substructures are essentially hydrophobic, while negatively contributing features possess hydroxyl or other substitutions which are likely not optimally placed to facilitate interactions with hydrogen bonding features in PXR. Therefore possession of these hydrogen bond acceptor and donor features indicated in the steroidal substructures appears to be related to loss of PXR activation. The method does not readily identify where these groups should be added in contrast to methods like docking [13,38].

## 3D-QSAR and 4D-QSAR

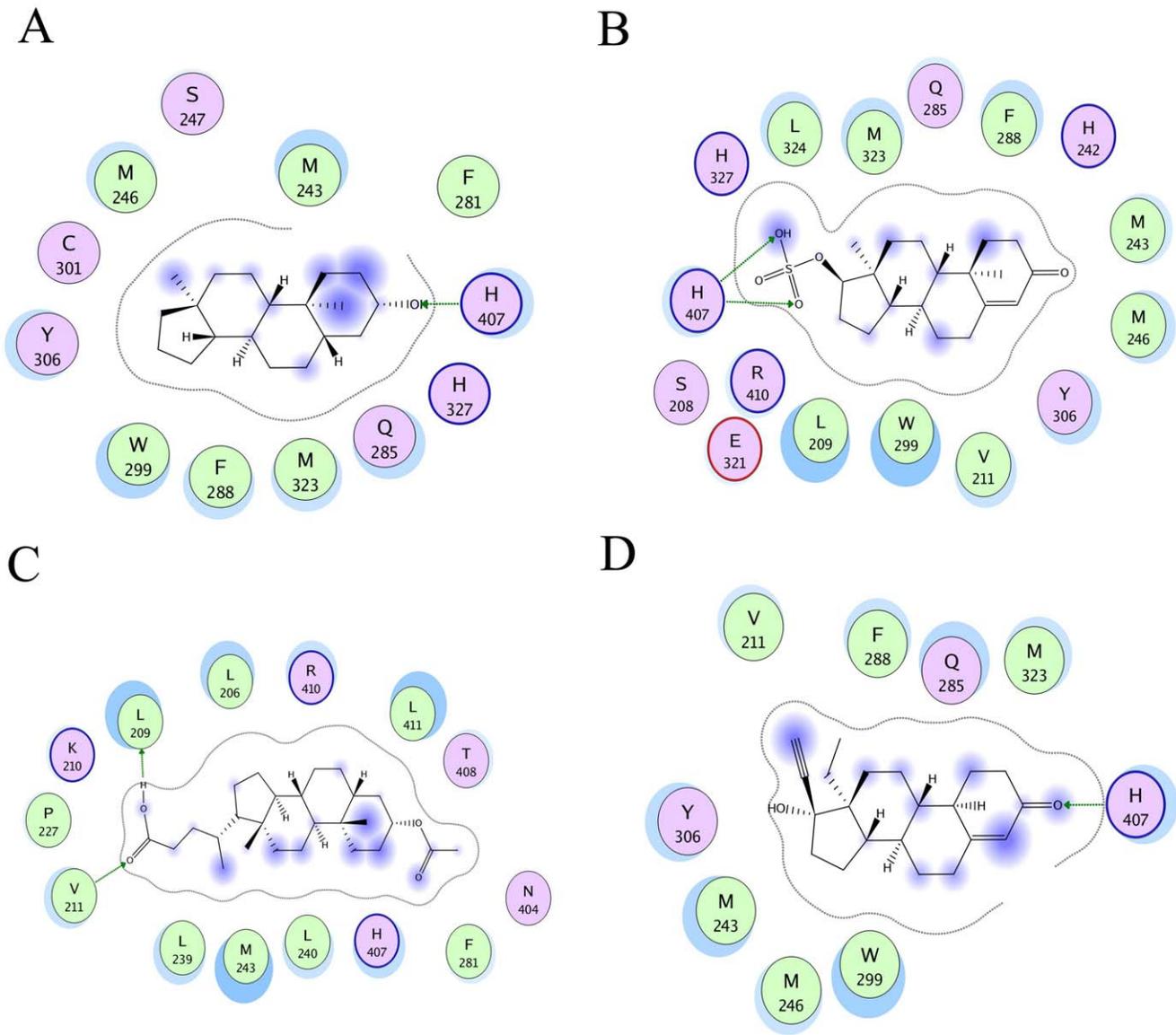
A major challenge in CoMFA and CoMSIA modeling is alignment of molecules, which must be defined by the user. As described in Text S1, multiple alignment approaches were attempted. Despite the use of multiple alignments, the best CoMFA and CoMSIA models consistently showed a large difference between the correlation  $R^2$  and cross-validated ( $XV-R^2$ ), whether modeling the entire set of steroidal compounds or the various subsets (androstanes, bile salts, pregnanes). This suggests that the CoMFA and CoMSIA models do not generalize beyond the molecules in the training set, even for a subset of steroidal compounds (Text S1, Tables S4, S5, S6, S7 and Figures S2, S3, S4, S5, S6, S7).

Using the pharmacophore approach for the individual steroids, the training set  $r$  values were quite low but increased upon inclusion of excluded volumes with variable weight and tolerances (0.81–0.93) (Table S8). All PXR pharmacophores (Figure S8) had at least 2 hydrophobes and a hydrogen bond acceptor in common (Text S1). Using the pharmacophores derived from training sets based on subsets of steroidal compounds (e.g., androstanes only) to predict the other respective subsets did not result in reliable correlations (data not shown), suggesting that highly specific pharmacophores were generated or this may be due to the addition of the excluded volumes which limits the chemical space of molecules mapping to the features. These class-specific pharmacophores may therefore only be useful for making predictions of very closely related molecules and even crossing steroidal classes may be extrapolating too far beyond the training sets.

**Table 1.** Docking results for all of the 119 molecules to the six hPXR crystal structures and the combined model in terms of statistical parameters namely sensitivity (SE), specificity (SP), overall prediction accuracy (Q) and matthews correlation coefficient (C) are listed for models predicted with  $17\beta$ -estradiol similarity weighted goldscores and the values in parenthesis are for models predicted with  $5\alpha$ -androstane- $3\beta$ -ol similarity weighted goldscores.

Structure	SE (%)	SP (%)	Q (%)	C (%)
1M13 (hyperforin)	66.67(56.7)	55.06(53.9)	57.98(54.6)	0.19(0.09)
1NRL (SR12813)	46.67(60)	47.19(41.6)	47.06(46.2)	-0.05(0.01)
1SKX (rifampicin)	53.33(70)	50.56(39.3)	51.26(47.1)	0.03(0.08)
2O9I (T0901317)	53.33(40)	50.56(40.4)	51.26(40.3)	0.03(-0.17)
2QNV (colupulone)	53.33(46.7)	52.81(31.5)	52.94(35.3)	0.05(-0.19)
EST (estradiol)	53.33(56.67)	50.56(49.44)	51.26(51.26)	0.02(0.05)
AVG	52 (55)	50.34 (42.69)	50.76(45.79)	0.02(-0.02)

The values in AVG represent the average prediction rates.  
doi:10.1371/journal.pcbi.1000594.t001



**Figure 1. Schematic representation of the binding mode of A. 5 $\alpha$ -Androstan-3 $\beta$ -ol B epitestosterone sulfate C lithocholic acid acetate and D levonorgestrol 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 representations were generated using the LIGX option in MOE.  
doi:10.1371/journal.pcbi.1000594.g001

4D-QSAR performed somewhat better than CoMSIA and CoMFA in modeling the compounds in the training sets using three atom alignments (Table S9). One potential advantage of 4D-QSAR relative to standard 3D-QSAR methods is the ability to consider an ensemble of different ligand conformations, theoretically increasing the chances of defining the active conformation. The best 4D-QSAR models are found in Table S10 and Figure S9, and generally predict steric/non-polar interactions between ligand and receptor. Although the XV- $R^2$  for the best 4D-QSAR models are better than for CoMFA and CoMSIA models of the same training sets, the 4D-QSAR were poorly predictive of the activity of compounds in the test set (Table S10).

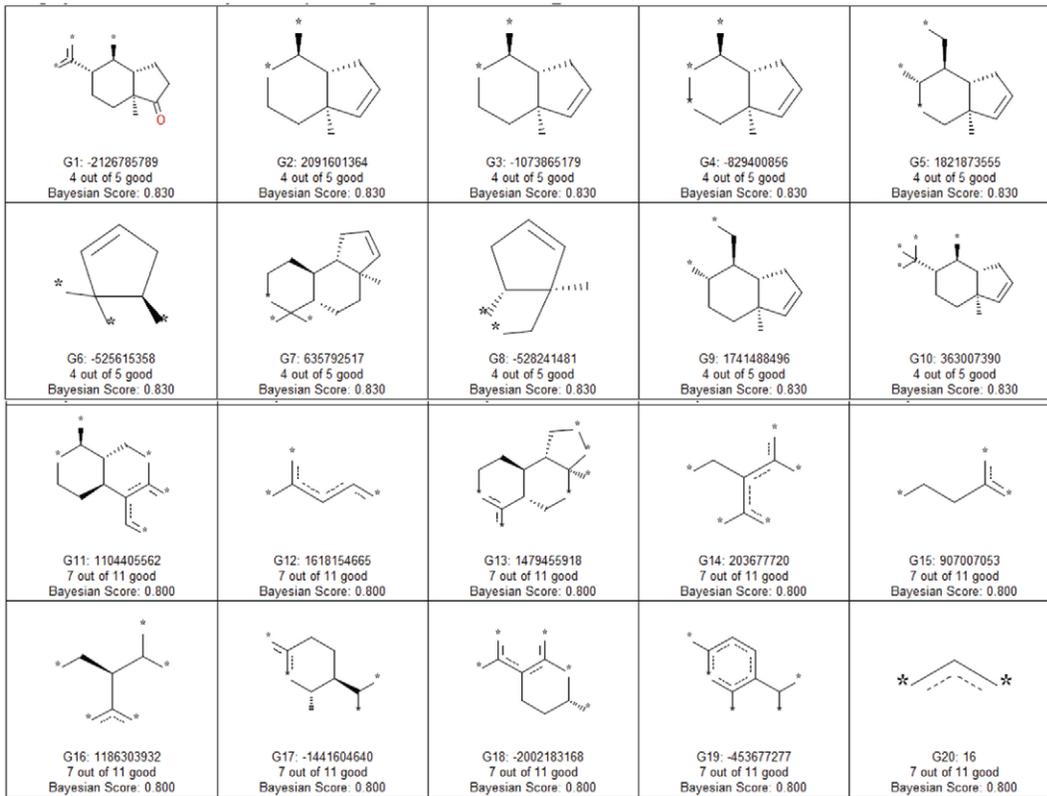
#### 5D - QSAR

4D- and 5D-QSAR have the advantage of being able to select the bioactive conformation from a pool of possible binding modes

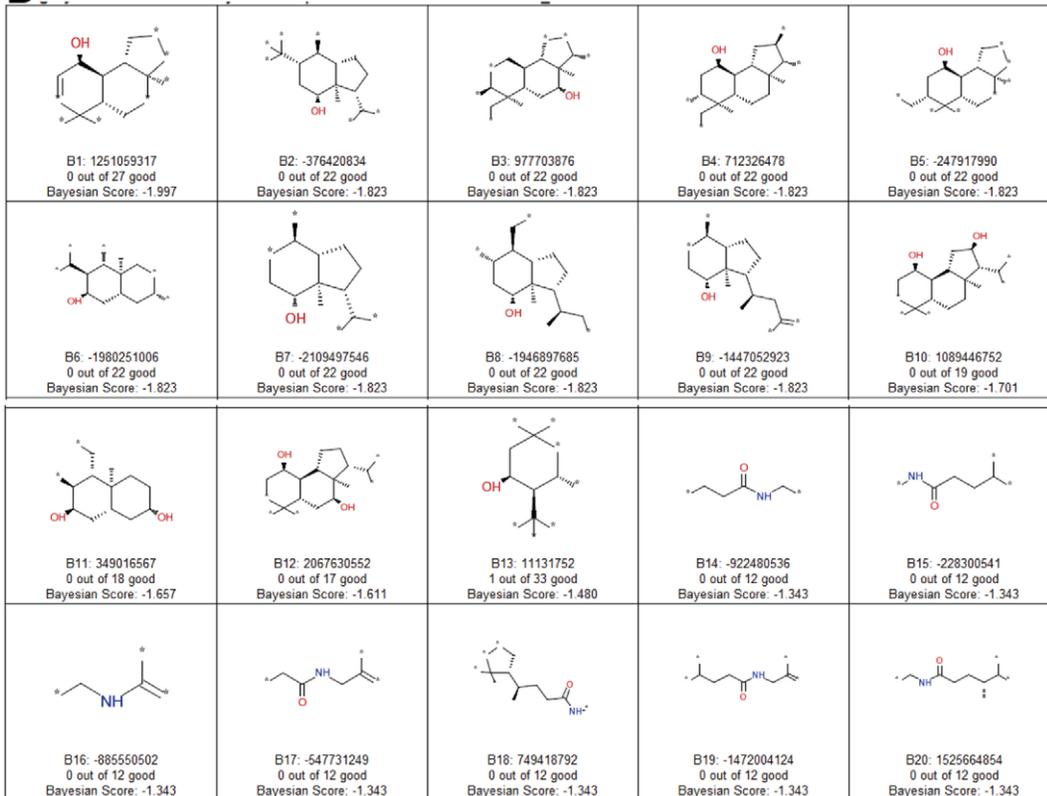
in parallel to the QSAR modeling stage. We have tested three different alignment protocols in conjunction with the 5D-QSAR technique Raptor.

- (1) The top-1, top-2, top-5, top-10 and top-20 docking poses for each ligand from our docking studies on 2QNV were superimposed producing an alignment containing 115 – 2300 conformations for the 115 compounds in our dataset.
- (2) All 115 compounds were automatically aligned onto 17 $\beta$ -estradiol extracted from the complex structure with PXR. Already aligned compounds were automatically added as templates for aligning subsequent compounds. The order of compounds in the alignment was determined based on their experimental affinity towards PXR.
- (3) As in (2) but each of the four substrate classes was aligned separately on a template of the class selected by the lowest

A



B



**Figure 2. Good and bad molecular features identified in the Bayesian model using FCFP<sub>6</sub> fingerprints.** A. Good features from FCFP<sub>6</sub> Bayesian model, B. Bad features from FCFP<sub>6</sub> Bayesian model. Asterisks can represent any atom. Numbers represent how many molecules out of the total number possessing the fingerprint are active (good) or inactive (bad).  
doi:10.1371/journal.pcbi.1000594.g002

binding affinity. The individual templates are first aligned onto 17 $\beta$ -estradiol. The four individual alignments are then combined into a 4D set for all 115 compounds.

In the alignment protocols (2) and (3) the protein crystal structure was used as a forbidden (excluded) region. A penalty was added to the similarity score for alignment solutions that overlapped with the protein, thus physically impossible solutions were removed from the alignment. As significant protein flexibility is observed on the side chain level, all crystal structures were aligned using PyMol [39]. Side chains that have different rotamer states for different co-crystallized ligands were removed from the forbidden region definition.

Our multidimensional QSAR study (software *Raptor* [40]) was based on the same set of 115 molecules as described in the CoMFA and CoMSIA studies. The dataset was split into 95 training set compounds, and 20 test set compounds identical to the separation used in the CoMFA and CoMSIA studies. For 33 compounds only an upper limit for their  $K_i$  values has been experimentally determined. These molecules defined the “threshold class” (26 training, 7 test). A threshold value of 100  $\mu$ M was chosen considering that the lowest affinities were measured for this dataset at approximately this value. To allow for topological and physicochemical variation at the true biological receptor with different ligands bound, the *Raptor* results were averaged over 10 individual models defining a surrogate conformational family.

For alignment (1) we were not able to derive QSAR models with predictive models for leave-5-groups-out ( $r^2_{CV-5}$ ) or cross-validation values (i.e. >0.3). This is not surprising, as the identification of bioactive binding modes using docking is difficult for this system (see docking results). If we use an alignment with only the top-1 or top-2 solutions, we most probably end up with an alignment containing incorrect binding modes. Using the top-10 or top-20 binding modes generates too large a variety of contacts between ligand and binding site model that the QSAR algorithm is not able to extract the critical interactions throughout the binding site modeling phase.

For alignment (2) a QSAR model with a  $r^2_{CV-5}$  value of 0.55 could be generated, but with no observed correlation for the test set. For alignment (3) a QSAR model with an  $r^2_{CV-5}$  of 0.56 was derived with a predictive  $r^2$  for the test set of 0.45. The superior model based on alignment (3) was due to the focused class-based alignment process (Figure 3). The maximum deviation of predicted from experimentally measured  $EC_{50}$  is 5.6 and 3.0 fold for training and test set, respectively. Significantly higher regression coefficients can hardly be expected for this dataset considering the fact that the threshold compounds have to be removed from the calculation of the regression coefficients yielding a rather small range in  $EC_{50}$  of 2.2 log units (Figure S10, Table S11). This is in contrast to the CoMFA and CoMSIA simulations where the threshold compounds have been assigned an  $EC_{50}$  value of 10,000  $\mu$ M yielding a range of 4.1 log units. All except one of the 33 threshold compounds have been predicted with an  $EC_{50}$  value lower than the given threshold or maximally a factor of 6.6 fold higher. Only 5 $\alpha$ -Androstane was predicted to have a 46 fold higher value than the threshold. Thus, the model was able to predict the affinity of compounds accurately and at the same time was able to classify weak- or non-binding molecules correctly.

## Discussion

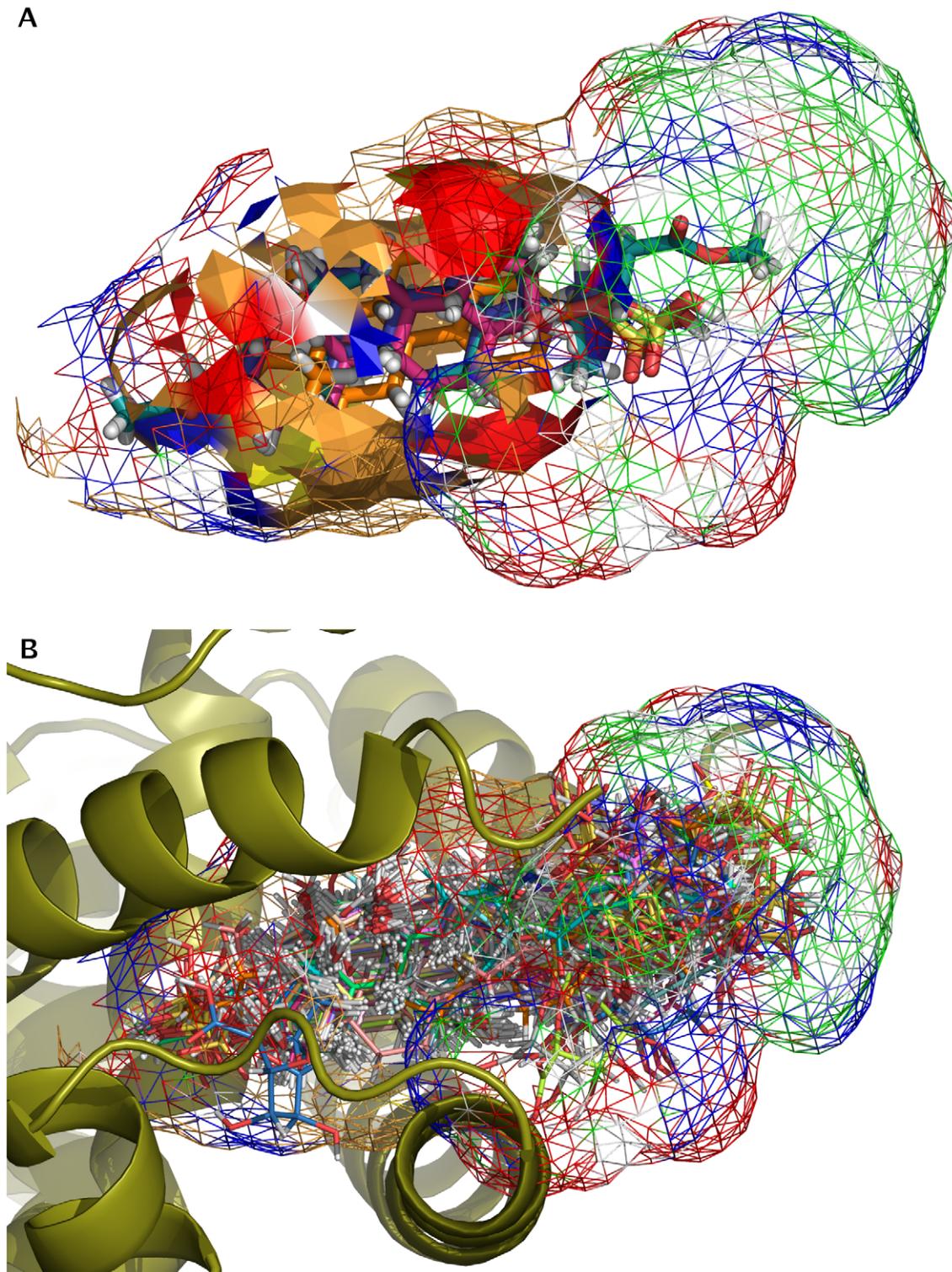
It has been suggested that PXR forms a heterotetramer and exhibits a range of motions which are key for its functioning and preparing for coactivator binding at the Activator Function (AF-2) site [41]. The large and promiscuous ligand binding pocket of PXR

accepts molecules of widely varying sizes (Table S1), and is likely capable of binding small molecules in multiple orientations. Furthermore, movement of regions of this pocket may be translated elsewhere in the protein to influence protein-protein interactions. Thus, the identification of the bioactive conformation of a ligand binding to PXR (and the effect it might have as an agonist, antagonist or allosteric antagonist [10]) and development of a ligand alignment based on these conformations represents a challenge for any computational technique. A realistic ligand alignment, however, is the basis for a reliable 3D-QSAR model. Computational methods including QSAR (3D, 4D and 5D), pharmacophores and machine learning classification models for PXR can assist in rapid prediction of whether a compound is likely to be an agonist (activator), however each method has its limitations and advantages (Table 2). For example a previous study used human PXR activation data for 30 steroidal compounds (including 9 bile acids) to create a pharmacophore with four hydrophobic features and one hydrogen bond acceptor [27]. This pharmacophore contained 5 $\alpha$ -androstane-3 $\beta$ -ol ( $EC_{50}$  0.8  $\mu$ M) which contains one hydrogen bond acceptor, indicating that in contrast to the crystal structure of 17 $\beta$ -estradiol (published  $EC_{50}$  20  $\mu$ M) bound to human PXR with two hydrogen bonding interactions [19], hydrophobic interactions may therefore be more important for increased affinity [27]. This and other pharmacophores have been used to predict PXR interactions for antibiotics [35] which were verified *in vitro*, suggesting one use for computational approaches in combination with experimental methods.

To our knowledge there has been no comparative analysis of the steroidal classes with respect to their use as PXR agonists. The use of the Bayesian classification with 2D fingerprints represents a low computational cost approach [42] which has been used frequently with large molecule datasets [43–46]. Using 2D-molecular fingerprint descriptors identified regions in the training set molecules that were predominantly hydrophobic and that were important for PXR activation. Substructures with free hydroxyls as hydrogen bonding features were associated with compounds that were not activators. This is in general agreement with other studies which have used docking to try to help design out PXR activation [38]. This model was able to successfully rank a large test set (Table S3) of non-steroidal molecules, indicative that the molecular descriptors adequately captured the global properties of PXR agonists and suggests some utility.

The current study suggests that while it is generally possible to create 3D-QSAR (CoMFA, CoMSIA, Catalyst) and 4D-QSAR models that can be cross-validated, these models perform poorly when used to predict external molecules. Only the 5D-QSAR model generated displays some success in predicting external test set steroidal compounds. Three main differences between the 5D-QSAR and the 3D-QSAR studies that might contribute to the difference in performance are the less rigid alignment using Symposar [40], the possibility to present a ligand in more than one binding pose and the better treatment of weak or non-binding compounds.

Pharmacophore models for the 4 classes of steroidal compounds possessed some of the features in the published human PXR crystal structures, however the models contained two or three hydrophobic regions (rather than four as shown previously) [27,28,31] and one to two hydrogen bond acceptors or a hydrogen bond acceptor and hydrogen bond donor (compared to one hydrogen bond acceptor as shown previously). This might suggest that the steroids evaluated occupy just a part of the ligand binding pocket while larger molecules like rifampicin occupy most of the binding pocket and have subsequently many more interactions with the protein [17]. The addition of the excluded volumes to the pharmacophores was shown to improve the



**Figure 3. Receptor model for PXR obtained using Raptor (beige-brown, hydrophobic properties; red, hydrogen bond acceptor; blue, hydrogen-bond donor; and green, hydrogen bond donor/acceptor).** The most active ligand of each of the four substrate classes aligned to each other is displayed as sticks. A: Inner shell is displayed in surface representation, outer shell in wireframe. B: The bulky right portion of the outer shell corresponds to the solvent exposed region of the ligand alignment. It is dominated by a mixed hydrogen bond donor/acceptor character in agreement with solvent exposure.  
doi:10.1371/journal.pcbi.1000594.g003

**Table 2.** Summary of the different methods used in this study.

Method	Advantages	Limitations
Bayesian Classification with 2D Fingerprints and interpretable descriptors	Computationally fast and cheap model generation, illustration of features important for activity	Cannot deal with stereoisomers, not quantitative, requires quite large training sets
3D-QSAR :Catalyst	Can use structurally diverse molecules, can add excluded volumes, quantitative, interpretable, starts from multiple conformations.	Models may not be useful beyond a narrow compound class, conformations may not be biologically relevant
3D-QSAR: CoMFA, COMSIA	Widely used methods and useful for drug design and analog modification.	Molecules require manual alignment and this may be a major limitation in this study
4D-QSAR	Considers an ensemble of different ligand conformations to define the active conformation	Computationally expensive, Alignment strategy may be a limitation
5D-QSAR	Considers an ensemble of different ligand conformations to define the active conformation in parallel, less rigid alignment, better treatment of weak binders.	Computationally expensive, Alignment strategy may be a limitation. For alignment crystal structures may not amply take into account the protein flexibility however in Raptor this is treated explicitly, exclusion areas could be too harsh.
GOLD docking and scoring	May provide potential binding orientation with respect to pocket which could be verified by site directed mutagenesis	Relatively slow, defining the binding site is key in such a large pocket

doi:10.1371/journal.pcbi.1000594.t002

correlation for the training sets and likely acts in a similar manner to using the crystal structures in 5D-QSAR.

Consistent with the QSAR findings were those from docking studies that though modest in success overall, fare much better with individual classes of compounds. The classification was performed using two similarity weighted scoring schemes: one based on a highly potent compound  $5\alpha$ -androstane- $3\beta$ -ol and the other based on a structurally relevant compound  $17\beta$ -estradiol. The goal was to test the utility of biasing the scoring scheme with either a structurally relevant compound or a functionally significant compound.

However, in this case  $17\beta$ -estradiol and  $5\alpha$ -androstane- $3\beta$ -ol share nearly 75% structural similarity (using MDL Keys and Tanimoto similarity coefficient). The results from the classification studies showed that biasing the scoring scheme with a structurally relevant compound ( $17\beta$ -estradiol) produced classification rates with sensitivity and specificity values averaging at 52% and 50% respectively with slightly better prediction accuracy (Table 1). These results unfortunately cannot be compared with our recent docking study [47] as a different co-crystal ligand was used for the scoring scheme. Although the structure biased scoring scheme performed better among all the compounds, both the scoring schemes performed equally well when individual classes were considered. In the case of androstanes, 6 out of 11 compounds were correctly predicted as activators in docking studies.  $5\alpha$ -Androstane- $3\beta$ -ol that had the lowest  $EC_{50}$  value (described earlier) was predicted to be an activator in all structures.  $5\alpha$ -Androstane- $3\beta$ -ol binds with very high docking scores and has a hydrogen bond interaction with His407, a key interaction of PXR (Figure 1A). This interaction was consistent among all the androstane activators. However, epitestosterone sulfate has an  $EC_{50}$  of  $3.39\ \mu\text{M}$  and was misclassified in the combined model using predictions from all structures as a non-activator. Docking studies show that epitestosterone sulfate has a consistently reversed docking pose (when compared with  $5\alpha$ -Androstane- $3\beta$ -ol) in all the models and the sulfate group is predicted to make a hydrogen bond interaction with His407, as opposed to the steroid ester in 1M13 structure (Figure 1B). A few other misclassified activators were docked in reversed poses and often had favorable hydrogen bonding partners such as sulfates that probably influence the binding mode of these steroids. This is a surprising and novel

finding of this study and other researchers should be aware of this when docking similar compounds with this functional group.

Among the bile salts, all four activators were correctly predicted and the ligands bind in a conserved mode with the steroid esters participating in favorable interactions with the side chain of His407 and Arg410, and the steroid rings with hydrophobic groups such as Leu411, Leu239 and Phe281 (Figure 1C). The pregnanes had similar activation patterns as the bile salts and docking studies could predict 4 out of the 9 compounds correctly. Among the misclassified compounds, levonorgestrol was predicted to be an activator in three models, and a non-activator in three models and hence could not be classified with high confidence. Levonorgestrol has an  $EC_{50}$  of  $4.30\ \mu\text{M}$  and is predicted to have favorable interactions with hPXR as shown in Figure 1D. Despite this, the similarity weighted scoring functions generally performed well in classifying activators as described in the examples above and by the sensitivity values in Table 1. The paucity of available PXR binding data may limit some of the insights from docking experiments performed to date.

It is not surprising that CoMFA and CoMSIA do not perform well as they use rigid alignments of the molecules. This is potentially a serious limitation given that the binding pocket of PXR may accommodate multiple orientations of the steroids (Figure 1A vs. Figure 1B). Theoretically, 4D- and 5-QSAR should perform better by considering an ensemble of ligand conformations and in fact 4D-QSAR does well within subsets (especially androstanes) but like all methods extrapolates poorly. 5D-QSAR appears to perform the best with the test set. Alignment independent methods like Catalyst which can deal with structurally diverse molecules can generate pharmacophores for the individual classes of compounds but their inter-class predictivity is limited. Another alignment independent method such as using 2D fingerprints and descriptors with the Bayesian classification approach may represent a fast approach to screen for potential PXR agonists, but like all methods their applicability domain [48,49] is dependent on the training set. In this case the set of steroids would be expected to limit the utility of such models to a relatively narrow class of compounds, although it may be picking up key features in more diverse molecules (Table S3) suggesting overlap in the chemical space.

This study shows the inherent difficulty of producing predictive ligand or structure-based computational models for PXR. Some of

the methods used are ligand alignment dependent while others are alignment independent, and each has limitations when used with flexible proteins. These computational models also confirm some of the molecular features (hydrophobicity and hydrogen bond acceptors) identified in previous models and structures, while using a large quantitative dataset to create new QSAR, classification and pharmacophore models to test docking and scoring. The study represents an initial step comparing multiple methods focused on steroidal compounds rather than a more diverse series of drug-like molecules. Using a more diverse series of molecules would have been expected to present even more difficulty for the alignment dependent methods such as CoMFA and CoMSIA. There are also many more commercial computational methods that could be evaluated and compared, although we have used several 3D, 4D, 5D-QSAR methods, machine learning with 2D descriptors, pharmacophore and GOLD docking and scoring methods in this study. The results from these methods could be used in combination as part of a consensus approach or Pareto optimization [50]. The provision of the 115 molecule human PXR dataset is potentially useful as a benchmark PXR set for testing further methods in future. For example, flexible docking methods [51] could be used as well as algorithms that could differentiate multiple binding mechanisms [52].

In conclusion, there are many promiscuous proteins [34] where the modeling of ligand-protein interactions is complicated by a large binding site, multiple binding pockets, protein flexibility or all of the preceding. We have applied several different computational approaches which could also be applied to other proteins like CYPs, transporters and ion channels. This work is therefore more broadly applicable in an attempt to predict whether molecules bind in such flexible proteins, and which methods perform the best. Depending on the desired use of such information, different modeling methods may be appropriate and required. While 2D methods do not encode 3D information like shape [53] they are fast and they can highlight important features likely interacting with the protein. 3D-5D methods provide more shape based information but they are fragile, with a narrow applicability domain and may not be able to differentiate close analogs. Docking is also limited unless key interactions with the protein are already known. Our results suggest that even in the presence of multiple crystal structures, the full range of protein motions may not be captured. As we have previously shown, when docking classification predictions are correct the binding conformation information alone may be instructive [13]. This current analysis indicates that using many different computational approaches (both alignment dependent and alignment independent) may be necessary and expectations should be scaled accordingly if some do not work with such promiscuous proteins. Even with their respective limitations, these methods have provided some useful information of general interest that could be applicable beyond PXR.

## Methods

### Experimental Methodology and Datasets

Human PXR activation was determined by a luciferase-based reporter assay as has been previously described [21,33,34]. The datasets modeled in this study were collected by a consistent protocol and have been previously published [31,54]. Experimental data for four classes of steroidal compounds, namely androstanes, estratrienes, pregnanes and bile acids/ salts are shown in Table S2.

### *In Silico* Methodology: Docking and Scoring

All molecules described in Table S2 were used for docking experiments. The molecules were docked into these six crystallized

structures of human PXR (PDB IDs 1M13, 1NRL, 1SKX, 2O9I, 2QNV and one structure co-crystallized with 17 $\beta$ -estradiol that is not in the PDB identified here as EST). 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 [55]) was used for docking all compounds to the binding sites of each PXR crystal structure. GOLD uses genetic algorithm to explore the various conformations of ligands and flexible receptor side chains in the binding pocket. Further, 20 independent docking runs were performed for each ligand. The docked complexes were scored with goldscore [55] and then rescored using similarity weighted scoring scheme (SWscore). For each ligand, the best ranking conformation's goldscore denoted by  $S_i$  was used to derive the SWscore shown in equation 1. The similarity scores  $W_i$  were computed based on 2D similarity encoded in MDL fingerprint keys calculated using Discovery Studio 2.1 (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:  $[\text{number of features in intersect (A, B)}] / [\text{number of features in union (A,B)}]$ , where A and B are two compounds.

So the SWscore is given by,  $\text{SWscore} = W_i * S_i$ , where  $W_i$  was the similarity score of compound  $i$  against 5 $\alpha$ -Androstan-3 $\beta$ -ol which had the best  $EC_{50}$  value of 0.8  $\mu\text{M}$  for PXR or 17 $\beta$ -estradiol which had a steroid core that was present in most of the compounds. Further, the quality of the scoring function was assessed using standard statistical indicators namely sensitivity (SE), specificity (SP), overall prediction accuracy (Q) and Matthews correlation coefficient (C) (Table 1) and were derived as described previously [22].

### *In Silico* Methodology: Machine Learning with 2D Descriptors

Bayesian models were generated using Discovery Studio 2.1 (Accelrys, San Diego, CA) Laplacian-corrected Bayesian classifier [37,42,43,45,56]. FCFP\_6 fingerprints, AlogP, molecular weight, number of rotatable bonds, number of rings, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors and molecular fractional polar surface area were calculated from the input sdf file using the "calculate molecular properties protocol". The "create Bayesian model protocol" was used for model generation and a custom protocol for validation (leave out 20% 100 times) was used.

### *In Silico* Methodology: 5D-QSAR - Symposar and Raptor

5D-QSAR studies were performed using Raptor [40]. Raptor includes the possibility of representing each ligand molecule as an ensemble of conformations, orientations, stereoisomers and protonation states (4D-QSAR), thereby reducing the bias in identifying the bioactive conformer. In addition, it explicitly allows for induced fit by a dual-shell representation of the three-dimensional binding-site model, onto which the physicochemical properties (hydrophobicity and hydrogen-bonding propensity) are mapped (5D-QSAR). The inner shell is tailored using the most potent ligand of the training set, the outer shell accommodates the topology of all molecules from the training set. The adaptation of both field and topology of the receptor surrogate to each ligand is achieved by combining a steric adjustment to the topology of every ligand and a term due to the attraction or repulsion between ligand and receptor model. The latter is obtained by correlating their physicochemical properties (hydrophobicity and hydrogen-

bond propensity) in 3D space. Since the mapping of properties onto the shells is not unambiguously determinable, different models with similar predictive power can be identified. *Raptor* generates a family of receptor models. Such model families may be interpreted to represent the various configuration states of the true biological receptor. The obtained binding affinities are averaged over the individual models.

The underlying scoring function for evaluating ligand-protein interactions includes directional terms for hydrogen bonding ( $\Delta G_{\text{Hbond}}$ ), hydrophobicity ( $\Delta G_{\text{Hphob}}$ ) as well as terms for the cost of the topological adaptation ( $\Delta G_{\text{TF}}$ ) and the changes in entropy ( $T\Delta S$ ) upon ligand binding:  $\Delta G_{\text{binding}} = \Delta G_{\text{constant}} + \Delta G_{\text{Hbond}} + \Delta G_{\text{Hphob}} - T\Delta S + \Delta G_{\text{TF}}$ .

Experimental determination of binding affinity for weak inhibitors is often prevented due to limited solubility or limited sensitivity. Thus, only an upper limit ('threshold') for  $K_i$  values is accessible. To prevent artificial assignment of affinities in a QSAR study including weak binders, the *Raptor* concept allows the use of a threshold option: the optimization algorithm forces the model to reproduce the binding affinity of the weak- and non-binding ligand molecules to be lower than the experimental limit. Obviously, compounds which are experimentally measured to bind weaker than a threshold  $K_i(t)$  and are correctly classified during the model optimization, no penalty is added to the lack-of-fit value, if, on the other hand, the binding affinity of the ligand is predicted higher than the threshold, the lack-of-fit function applies a penalty proportional to  $\Delta G_{\text{binding}(t)} - \Delta G_{\text{binding}}$ .

4D sets of alternative conformations for each ligand as input for *Raptor* were performed with Symposar [57]. In Symposar the ligand molecules are superimposed onto one or several template molecules, first, on the basis of fuzzy-like 2D substructure similarities and, subsequently, in 3D space with respect to their similarity of physicochemical fields. This two-step process combines the speed of a 2D similarity search with the accuracy and authenticity of protein-ligand interactions in 3D space. The molecules are thereby treated as flexible and are fully relaxed at the end of the alignment process.

## Supporting Information

**Table S1** Molecular descriptors for PXR crystal structure ligands calculated with Discovery Studio ver 2.1 (Accelrys, San Diego, CA).

Found at: doi:10.1371/journal.pcbi.1000594.s001 (0.01 MB PDF)

**Table S2** Dataset of human PXR activation used for modeling studies.

Found at: doi:10.1371/journal.pcbi.1000594.s002 (0.04 MB PDF)

**Table S3** Test set [13] prediction with Bayesian model (activator =  $EC_{50} < 100 \mu\text{M}$ , non-activator =  $EC_{50} > 100 \mu\text{M}$ ).

Found at: doi:10.1371/journal.pcbi.1000594.s003 (0.02 MB PDF)

**Table S4** CoMFA Test Set Predictions

Found at: doi:10.1371/journal.pcbi.1000594.s004 (0.02 MB PDF)

**Table S5** To identify which outliers are bringing down the  $XV-R_2$  of the CoMFA model, the following table lists the activities as predicted by the cross-validated PLS model: The standard deviation of the residuals in the following table is 0.825, and accordingly, the two possible outliers are highlighted. The two outliers are the only two inactives in the training set.

Found at: doi:10.1371/journal.pcbi.1000594.s005 (0.02 MB PDF)

**Table S6** CoMFA outlier analysis.

Found at: doi:10.1371/journal.pcbi.1000594.s006 (0.02 MB PDF)

**Table S7** Test set predictions for CoMFA and CoMSIA models. Found at: doi:10.1371/journal.pcbi.1000594.s007 (0.02 MB PDF)

**Table S8** Best model training set correlation ( $r$ ) values and model statistics (total cost and null cost) for Catalyst Hypogen hypotheses.

Found at: doi:10.1371/journal.pcbi.1000594.s008 (0.01 MB PDF)

**Table S9** Three-ordered atom alignments (based on the steroidal core) used in the 4D-QSAR analysis.

Found at: doi:10.1371/journal.pcbi.1000594.s009 (0.01 MB PDF)

**Table S10** External Validation test Set Predictions for 4D-QSAR

Found at: doi:10.1371/journal.pcbi.1000594.s010 (0.02 MB PDF)

**Table S11** Experimental versus predicted pEC50 values for 115 compounds binding to PXR divided into four different substrate classes - 5D-QSAR.

Found at: doi:10.1371/journal.pcbi.1000594.s011 (0.03 MB PDF)

**Text S1** In silico methodology: 3D-QSAR - CoMFA, CoMSIA, In silico methodology: 3D-QSAR - Catalyst, In silico methodology: 4D-QSAR, Supplemental results: CoMFA, CoMSIA and Catalyst. Supplemental data - pharmacophores output files from Discovery Studio Catalyst.

Found at: doi:10.1371/journal.pcbi.1000594.s012 (0.08 MB PDF)

**Figure S1** Structural superposition of six PXR crystal structures are shown in ribbon models and colored 1M13 (red), 1NRL (orange), 1SKX (cyan), 2O9I (blue), 2QNV (yellow) and PXR-EST (brown). The co-crystallized ligands are shown as sticks and colored blue for rifampicin, orange for colupulone, dark green for hyperforin, light green for N-{4-[2,2,2-trifluoro-1-hydroxy-1-(trifluoromethyl)-ethyl]phenyl}benzenesulfonamide and pink for 17 $\beta$ -estradiol.

Found at: doi:10.1371/journal.pcbi.1000594.s013 (0.68 MB TIF)

**Figure S2** CoMFA models for androstanes. A 5 $\alpha$ -Androstan-3 $\beta$ -ol (pIC50 = 6.1) shown with the steric component of the CoMFA model. Green denotes areas where steric bulk is favorable for bioactivity while yellow shows areas where steric bulk is not favored. B 5 $\alpha$ -Androstan-3 $\beta$ -ol shown with the electrostatic component of the CoMFA model. Blue denotes areas where positive charge is favorable for bioactivity while red shows areas where negative charge is favored.

Found at: doi:10.1371/journal.pcbi.1000594.s014 (0.22 MB TIF)

**Figure S3** CoMSIA models for androstanes. A - 17 $\beta$ -dihydroandrosterone (pIC50 = 5.38) with the steric component of the CoMSIA model. Blue denotes areas where steric bulk is favorable for bioactivity while red shows areas where steric bulk is not favored. B 17 $\beta$ -dihydroandrosterone with the hydrophobic component of the CoMSIA model. Purple denotes areas where hydrophobic groups are favorable for bioactivity while grey shows areas where hydrophobic groups are not preferred. C 17 $\beta$ -dihydroandrosterone with the hydrogen bond acceptor component of the CoMSIA model. Blue denotes areas where acceptor groups are favorable for bioactivity while red shows areas where acceptor groups are not preferred.

Found at: doi:10.1371/journal.pcbi.1000594.s015 (0.24 MB TIF)

**Figure S4** CoMFA models for pregnanes. A Pregnanedione (pIC50 = 5.59) shown with the steric component of the CoMFA model. Green denotes areas where steric bulk is favorable for bioactivity while yellow shows areas where steric bulk is not favored. B Pregnanedione shown with the electrostatic component of the CoMFA model. Blue denotes areas where positive charge is

favorable for bioactivity while red shows areas where negative charge is favored.

Found at: doi:10.1371/journal.pcbi.1000594.s016 (0.26 MB TIF)

**Figure S5** CoMSIA models for Pregnanes. A. Inactive training set molecule Pregnenolone Carbonitrile (PCN) (pIC<sub>50</sub> = 2.00) with the steric component of the CoMSIA model. Blue denotes areas where steric bulk is favorable for bioactivity while red shows areas where steric bulk is not favored. B. Inactive training set molecule PCN shown with the electrostatic component of the CoMSIA model. Blue denotes areas where positive charge is favorable for bioactivity while red shows areas where negative charge is favored. C. Inactive training set molecule PCN with the hydrophobic component of the CoMSIA model. Purple denotes areas where hydrophobic groups are favorable for bioactivity while grey shows areas where hydrophobic groups are not preferred.

Found at: doi:10.1371/journal.pcbi.1000594.s017 (0.23 MB TIF)

**Figure S6** A. CoMFA models for bile acids and bile salts. Lithocholic acid acetate (pIC<sub>50</sub> = 5.92) shown with the steric component of the CoMFA model. Green denotes areas where steric bulk is favorable for bioactivity while yellow shows areas where steric bulk is not favored. B. Lithocholic acid acetate shown with the electrostatic component of the CoMFA model. Blue denotes areas where positive charge is favorable for bioactivity while red shows areas where negative charge is favored.

Found at: doi:10.1371/journal.pcbi.1000594.s018 (0.27 MB TIF)

**Figure S7** CoMSIA models of bile acids and bile salts. Using the PLS focused region, CoMSIA components were calculated. A. Hyodeoxycholic acid (pIC<sub>50</sub> = 4.42) shown with electrostatic components of the CoMSIA model. Blue denotes areas where positive charge is favorable for bioactivity while red shows areas where negative charge is favored. B. Hyodeoxycholic acid with the hydrophobic component of the CoMSIA model. Purple denotes areas where hydrophobic groups are favorable for bioactivity while grey shows areas where hydrophobic groups are not preferred. C. Hyodeoxycholic acid with the hydrogen bond donor component

of the CoMSIA model. Blue denotes areas where donor groups are favorable for bioactivity while red shows areas where donor groups are not preferred.

Found at: doi:10.1371/journal.pcbi.1000594.s019 (0.31 MB TIF)

**Figure S8** Catalyst PXR pharmacophores A. Bile acids, B. Estratrienes, C. Androstanes, D Pregnanes pharmacophore features represent Green = hydrogen bond acceptor, purple = hydrogen bond donor, blue = Hydrophobic, gray = excluded volumes.

Found at: doi:10.1371/journal.pcbi.1000594.s020 (0.15 MB TIF)

**Figure S9** A. 4D-QSAR for androstanes showing the active conformation of 5 $\alpha$ -Androstan-3 $\beta$ -ol, B 4D-QSAR for pregnanes showing the active conformation of pregnanolone, C 4D-QSAR for bile acids/salts showing the active conformation of lithocholic acid acetate.

Found at: doi:10.1371/journal.pcbi.1000594.s021 (0.19 MB TIF)

**Figure S10** 5D-QSAR Experimental versus predicted pEC<sub>50</sub> values for 115 compounds binding to PXR. Training set compounds are displayed in green, test set compounds in red. Threshold compounds are placed at an experimental pEC<sub>50</sub> value of 2 to better separate them visually from the other molecules.

Found at: doi:10.1371/journal.pcbi.1000594.s022 (0.11 MB TIF)

## Acknowledgments

S.E. gratefully acknowledges Accelrys for making Discovery Studio Catalyst available to us. M.D.K. and M.I. gratefully acknowledge Prof. A.J. Hopfinger (The Chem21 Group, Lake Forest, IL) for providing 4D-QSAR software.

## Author Contributions

Conceived and designed the experiments: SE SK MAL MDK. Performed the experiments: SE SK MI EJR MAL MDK. Analyzed the data: SE SK MAL MDK. Contributed reagents/materials/analysis tools: SE SK EJR MAL MRR MDK. Wrote the paper: SE SK MAL MRR MDK.

## References

- Ekins S (2004) Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discovery Today* 9: 276–285.
- Krasowski MD, Reschly EJ, Ekins S (2008) Intrinsic disorder in nuclear hormone receptors. *J Proteome Res* 7: 4359–4372.
- Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. *Biochemistry* 41: 6573–6582.
- Bertilsson G, Heidrich J, Svensson K, Asman M, Jendeberg L, et al. (1998) Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction. *Proc Natl Acad Sci U S A* 95: 12208–12213.
- Blumberg B, Sabbagh W Jr, Jugulion H, Bolado J, Jr., van Meter CM, et al. (1998) SXR, a novel steroid and xenobiotic-sensing nuclear receptor. *Genes Dev* 12: 3195–3205.
- Kliwer SA, Moore JT, Wade L, Staudinger JL, Watson MA, et al. (1998) An orphan nuclear receptor activated by pregnanes defines a novel steroid signalling pathway. *Cell* 92: 73–82.
- Verma S, Tabb MM, Blumberg B (2009) Activation of the steroid and xenobiotic receptor, SXR, induces apoptosis in breast cancer cells. *BMC Cancer* 9: 3.
- Gupta D, Venkatesh M, Wang H, Kim S, Sinz M, et al. (2008) Expanding the roles for pregnane X receptor in cancer: proliferation and drug resistance in ovarian cancer. *Clin Cancer Res* 14: 5332–5340.
- Zhou J, Liu M, Zhai Y, Xie W (2008) The antiapoptotic role of pregnane X receptor in human colon cancer cells. *Mol Endocrinol* 22: 868–880.
- Biswas A, Mani S, Redinbo MR, Krasowski MD, Li H, et al. (2009) Elucidating the 'Jekyll and Hyde' nature of PXR: The case for discovering antagonists. *Pharm Res* 26: 1807–1815.
- Urquhart BL, Tirone RG, Kim RB (2007) Nuclear receptors and the regulation of drug-metabolizing enzymes and drug transporters: implications for interindividual variability in response to drugs. *J Clin Pharmacol* 47: 566–578.
- Mani S, Redinbo MR, Krasowski MD, Biswas A, Li H, et al. (2009) Elucidating the 'Jekyll and Hyde' Nature of PXR: The Case for Discovering Antagonists *Pharm Res* Submitted.
- Khandelwal A, Krasowski MD, Reschly EJ, Sinz MW, Swaan PW, et al. (2008) Machine learning methods and docking for predicting human pregnane X receptor activation. *Chem Res Toxicol* 21: 1457–1467.
- Watkins RE, Davis-Searles PR, Lambert MH, Redinbo MR (2003) Coactivator binding promotes the specific interaction between ligand and the pregnane X receptor. *J Mol Biol* 331: 815–828.
- Watkins RE, Maglich JM, Moore LB, Wisely GB, Noble SM, et al. (2003) 2.1A crystal structure of human PXR in complex with the St John's Wort compound hyperforin. *Biochemistry* 42: 1430–1438.
- Xue Y, Chao E, Zuercher WJ, Willson TM, Collins JL, et al. (2007) Crystal structure of the PXR-T1317 complex provides a scaffold to examine the potential for receptor antagonism. *Bioorg Med Chem* 15: 2156–2166.
- Chrencik JE, Orans J, Moore LB, Xue Y, Peng L, et al. (2005) Structural disorder in the complex of human pregnane X receptor and the macrolide antibiotic rifampicin. *Mol Endocrinol* 19: 1125–1134.
- Teotico DG, Bischof JJ, Peng L, Kliwer SA, Redinbo MR (2008) Structural basis of human pregnane X receptor activation by the hops constituent colupulone. *Mol Pharmacol* 74: 1512–1520.
- Xue Y, Moore LB, Orans J, Peng L, Bencharit S, et al. (2007) Crystal structure of the pregnane X receptor-estradiol complex provides insights into endobiotic recognition. *Mol Endocrinol* 21: 1028–1038.
- Ghose AK, Viswanadhan VN, Wendoloski JJ (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J Phys Chem* 102: 3762–3772.
- Watkins RE, Wisely GB, Moore LB, Collins JL, Lambert MH, et al. (2001) The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity. *Science* 292: 2329–2333.
- Kortagere S, Chekmarev D, Welsh WJ, Ekins S (2008) Hybrid scoring and classification approaches to predict human pregnane X receptor activators. *Pharm Res* In press.
- Ekins S, Mirny L, Schuetz EG (2002) A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXR $\alpha$  and LXR $\beta$ . *Pharm Res* 19: 1788–1800.

24. Jacobs MN (2004) In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology* 205: 43–53.
25. Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, et al. (2006) A Combined Approach to Drug Metabolism and Toxicity Assessment. *Drug Metab Dispos* 34: 495–503.
26. Bachmann K, Patel H, Batayneh Z, Slama J, White D, et al. (2004) PXR and the regulation of apoA1 and HDL-cholesterol in rodents. *Pharmacol Res* 50: 237–246.
27. Ekins S, Chang C, Mani S, Krasowski MD, Reschly EJ, et al. (2007) Human pregnane X receptor antagonists and agonists define molecular requirements for different binding sites. *Mol Pharmacol* 72: 592–603.
28. Ekins S, Erickson JA (2002) A pharmacophore for human pregnane-X-receptor ligands. *Drug Metab Dispos* 30: 96–99.
29. Schuster D, Langer T (2005) The identification of ligand features essential for PXR activation by pharmacophore modeling. *J Chem Inf Model* 45: 431–439.
30. Ung CY, Li H, Yap CW, Chen YZ (2007) In silico prediction of pregnane X receptor activators by machine learning approaches. *Mol Pharmacol* 71: 158–168.
31. Ekins S, Reschly EJ, Hagey LR, Krasowski MD (2008) Evolution of pharmacologic specificity in the pregnane X receptor. *BMC Evol Biol* 8: 103.
32. Cramer RD, Patterson DE, Bunce JD (1988) Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110: 5959–5967.
33. Yates CR, Chang C, Kearbey JD, Yasuda K, Schuetz EG, et al. (2003) Structural determinants of P-glycoprotein-mediated transport of glucocorticoids. *Pharm Res* 20: 1794–1803.
34. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, et al. (2007) Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* 2: 874–880.
35. Yasuda K, Ranade A, Venkataraman R, Strom S, Chupka J, et al. (2008) A comprehensive in vitro and in silico analysis of antibiotics that activate pregnane X receptor and induce CYP3A4 in liver and intestine. *Drug Metab Dispos* 36: 1689–1697.
36. Mnif W, Pascussi JM, Pillion A, Escande A, Bartegi A, et al. (2007) Estrogens and antiestrogens activate PXR. *Toxicol Lett* 170: 19–29.
37. Hassan M, Brown RD, Varma-O'Brien S, Rogers D (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* 10: 283–299.
38. Gao YD, Olson SH, Balkovec JM, Zhu Y, Royo I, et al. (2007) Attenuating pregnane X receptor (PXR) activation: a molecular modelling approach. *Xenobiotica* 37: 124–138.
39. DeLano WL (2002) The PyMOL Molecular Graphics System.
40. Lill MA, Vedani A, Dobler M (2004) Raptor: combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling: application toward the simulation of structurally diverse ligand sets. *J Med Chem* 47: 6174–6186.
41. Teotico DG, Frazier ML, Ding F, Dokholyan NV, Temple BR, et al. (2008) Active nuclear receptors exhibit highly correlated AF-2 domain motions. *PLoS Comput Biol* 4: e1000111.
42. Rogers D, Brown RD, Hahn M (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen* 10: 682–686.
43. Prathipati P, Ma NL, Keller TH (2008) Global Bayesian models for the prioritization of antitubercular agents. *J Chem Inf Model* 48: 2362–2370.
44. Metz JT, Huith JR, Hajduk PJ (2007) Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J Comput Aided Mol Des* 21: 139–144.
45. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, et al. (2007) Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* 2: 861–873.
46. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 46: 1124–1133.
47. Kortagere S, Chekmarev D, Welsh WJ, Ekins S (2009) Hybrid scoring and classification approaches to predict human pregnane X receptor activators. *Pharm Res* 26: 1001–1011.
48. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI (2006) Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov Today* 11: 700–707.
49. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, et al. (2008) Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* 48: 1733–1746.
50. Suarez M, Tortosa P, Carrera J, Jaramillo A (2008) Pareto optimization in computational protein design with multiple objectives. *J Comput Chem* 29: 2704–2711.
51. Lill MA, Dobler M, Vedani A (2006) Prediction of Small-Molecule Binding to Cytochrome P450 3A4: Flexible Docking Combined with Multidimensional QSAR. *ChemMedChem* 6: 73–81.
52. Feng J, Sanil A, Young SS (2006) PharmID: pharmacophore identification using Gibbs sampling. *J Chem Inf Model* 46: 1352–1359.
53. Kortagere S, Krasowski MD, Ekins S (2009) The importance of discerning shape in molecular pharmacology. *Trends Pharmacol Sci* 30: 138–147.
54. Krasowski MD, Yasuda K, Hagey LR, Schuetz EG (2005) Evolution of the pregnane x receptor: adaptation to cross-species differences in biliary bile salts. *Mol Endocrinol* 19: 1720–1739.
55. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 267: 727–748.
56. Klon AE, Lowrie JF, Diller DJ (2006) Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J Chem Inf Model* 46: 1945–1956.
57. Lill MA, Vedani A (2006) Combining 4D pharmacophore generation and multidimensional QSAR: Modeling ligand binding to the Bradykinin B2 receptor. *J Chem Inf Model* 46: 2135–2145.