Bioorganic & Medicinal Chemistry 20 (2012) 5352–5365

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Development of in silico filters to predict activation of the pregnane X receptor (PXR) by structurally diverse drug-like molecules

Hans Matter *, Lennart T. Anger †, Clemens Giegerich, Stefan Güssregen, Gerhard Hessler, Karl-Heinz Baringhaus

Sanofi-Aventis Deutschland GmbH, R&D, LGCR, Structure, Design and Informatics, Building G 878, D-65926 Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Available online 19 April 2012

Keywords:
Pregnane X receptor (PXR)
Molecular modelling
In silico filter
Antitarget
QSAR

ABSTRACT

The pregnane X receptor (PXR), a member of the nuclear hormone superfamily, regulates the expression of several enzymes and transporters involved in metabolically relevant processes. The significant induction of CYP450 enzymes by PXR, in particular CYP3A4, might significantly alter the metabolism of prescribed drugs. In order to early identify molecules in drug discovery with a potential to activate PXR as *antitarget*, we developed fast and reliable in silico filters by ligand-based QSAR techniques. Two classification models were established on a diverse dataset of 434 drug-like molecules. A second augmented set allowed focusing on interesting regions in chemical space. These classifiers are based on decision trees combined with a genetic algorithm based variable selection to arrive at predictive models. The classifier for the first dataset on 29 descriptors showed good performance on a test set with a correct classification of both 100% for PXR activators and non-activators plus 87% for activators and 83% for non-activators in an external dataset. The second classifier then correctly predicts 97% activators and 91% non-activators in a test set and 94% for activators and 64% non-activators in an external set of 50 molecules, which still qualifies for application as a filter focusing on PXR activators. Finally a quantitative model for PXR activation for a subset of these molecules was derived using a regression-tree approach combined with GA variable selection. This final model shows a predictive r^2 of 0.774 for the test set and 0.452 for an external set of 33 molecules. Thus, the combination of these filters consistently provide guidelines for lowering PXR activation in novel candidate molecules.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

The pregnane X receptor (PXR; NR1I2) was identified in 1998 as a member of the nuclear hormone receptor (NHR) superfamily.[1–4] It is expressed in liver, intestine and in organs involved in the absorption, distribution, metabolism and elimination (ADME) of structurally diverse endobiotics and drug molecules. This receptor was initially found to be activated by several drug molecules, which are known to regulate CYP3A4 gene expression and thus cause clinically relevant drug–drug interactions.[2] Since then it has been unveiled that PXR regulates the expression of several metabolizing enzymes including cytochrome P450's from the CYP3A and CYP2B subfamily,[1,5] such as CYP3A4[6] as most abundant cytochrome expressed in the liver and CYP2B6, in addition to CYP2C8/9.[7] PXR activation also induces UDP-glucuronosyl-transferases and glutathione-S-transferases as major phase-II conjugating enzymes.[8,9] It also regulates the expression of important drug transporters[10] such as P-glycoprotein and multidrug resistance proteins.[11,12] The involvement of this nuclear hormone receptor in multiple relevant ADME processes are further discussed in many recent review articles.[13–15]

The activation of PXR depends on ligand-binding to its ligand-binding domain (LDB). Following this recognition event, PXR then forms a heterodimeric complex with the retinoic X receptor (RXR; another member of the NHR superfamily[16]), which subsequently binds to PXR response elements in the 5′-flanking region of PXR target DNA sequences. Among these target genes are those encoding phase-I and phase-II metabolizing enzymes and transporters. It has been experimentally demonstrated that PXR activation in fact regulates a large network of genes. For example 138 genes were induced and 82 were repressed in rats treated with the PXR ligand pregnenolone 16α-carbonitrile.[17] However, many of these genes from this and related studies[10] were not further validated by direct mRNA quantification and thus may not be direct PXR targets.

Unlike other NHRs like PPARs and steroid receptors that interact with a higher degree of specificity with their physiological ligands, PXR ligands are structurally very diverse. This finding can be attributed to a complex defense strategy of the organism in response to

* Corresponding author. Tel.: +49 69 305 84329.
E-mail address: hans.matter@sanofi-aventis.com (H. Matter).
† Present address: Charité University Medical School Berlin, Institute of Clinical Pharmacology and Toxicology, Luisenstr. 7, D-10117 Berlin, Germany.

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.04.020

threats by multiple xenobiotics. Metabolism of drugs and other molecules in the liver is the primary defense against accumulation of potentially toxic lipophilic compounds.[15] This strategy involves xenosensors like PXR, the constitutive androstane receptor (CAR)[14,15,18–21] and the aryl hydrocarbon receptor (AhR)[22] to recognize potentially dangerous molecules and cause an increase of the concentration of metabolic enzymes by inducing their transcription. The analysis of mammalian PXR ligand-binding specificity provides also an instructive example of refinement in NHR ligand-binding during evolution.[23] The comparison of in vitro assay data between vertebrate species revealed that PXR as original '*biliary salt receptor*' underwent a significant broadening of its specificity, being able to bind diverse androstanes, pregnanes, C27 bile alcohols sulfates and xenobiotics ligands in the common vertebrate ancestor, while in amniotes the specificity is restricted to C24 bile acids.[24,25]

Many ligands for human PXR have been identified among prescription drugs; those include the antibiotics rifampicin, clotrimazole and ritonavir; the antineoplastic drugs cyclophosphamide, cyproterone acetate, taxol, tamoxifen, and RU486; the anti-inflammatory agent dexamethasone; the anti-type 2 diabetes drug troglitazone, the antihypertensive drugs nifedipine and spironolactone and the sedatives glutethimide and phenobarbital.[26,15] In addition, some commonly used herbal medicines can also activate PXR, like St. John's wort.[27]

The significant induction of CYP450 enzymes, in particular from the CYP3A family, caused by PXR might dramatically alter the metabolism of a variety of prescribed drugs, as the most abundant CYP3A4 itself participates in the metabolism of >50% of the marketed drugs.[28,29] Owing to this important role in the regulation of metabolic enzymes, PXR activation might significantly impact the plasma levels and fate of many drug molecules in clinical studies and therapy. Therefore, drugs capable of activating PXR could induce their own metabolism and transport and also interact with co-administered pharmaceuticals.

Consequently the early identification of PXR activators in lead identification and lead optimization is important to focus on compounds with fewer side effects. To this end, various in vitro assays have been introduced to monitor compounds as early as possible for their ADME and antitarget-interaction characteristics. However, these assays require the synthesis of material to be tested. To save time and resources, it is desirable to establish in silico tools into the drug development process allowing the reliable assessment of compound properties prior to synthesis, and subsequently the ranking of a priority list of the most promising compounds to be synthesized. Thus, facing the requirements of modern drug discovery, rational approaches to optimize molecules directed by quantitative structure–activity relationship (QSAR) and structure-based design require a tight interplay between multiple disciplines: medicinal chemistry, structural biology, pharmacology and pharmacokinetics.

In this report we describe our approach towards the development of a fast and reliable filter[30] for PXR activation by molecules binding to this *antitarget* receptor. To this end, we employed ligand-based QSAR techniques for deriving predictive PXR activation models. Ligand-based QSAR models are aiming to provide a reliable classification of active molecules; they might also serve to identify conserved structural features among active molecules. However their application in a drug discovery context requires careful validation using external test sets before productive use in project settings. Additionally, these models should encompass multiple chemical series assuming similar competitive binding modes in PXR, thus showing predictivity beyond a single congeneric series. However, the amount and diversity of structural and biological data in this field might not yet qualify any obtained model as truly global in its scope of applicability.

We first report on the development and validation of a classification model for PXR activation on a large literature dataset. For classification we use a well-established decision tree method, which has not been used very frequently in the chemistry field yet. This approach was modified by us to encompass a genetic algorithm (GA) based variable selection to arrive at significant and predictive models. In a second classification model this dataset is augmented by internal molecules from drug discovery programs tested for their potential to activate PXR. The idea is to add additional chemical information to those chemical space regions, which are of interest for us in current project support. Hence, the scope and diversity comes from the broader literature dataset, while more details about structure–activity relationships result from the addition of particular chemical series. Finally, we used diverse quantitative data on PXR activation for a subset of these molecules covering interesting regions in chemical space to derive a quantitative model using a well-established regression-tree approach. The chemical interpretation of these models will be discussed, showing a broader picture of structural prerequisites for PXR activation.

2. Methods

2.1. Datasets

To generate qualitative 2D models for identification of PXR activators, we first collected public compounds from literature and internal databases. Datasets from earlier QSAR publications were collected or re-built according to reported information. We started with the dataset from Ung et al.[31] with 128 PXR activators and 77 non-activators. In this dataset, compounds with a pEC_{50} value[32] >4 were classified as PXR activators and pEC_{50} <4 as non-activators. This dataset, obtained as SMILES strings from the Supplementary data, was combined with the structure compilation by Khandelwal et al.,[33] which was also summarizing some molecules from previous publications[34–36] in addition to their own data. Further compounds were extracted by us from more recent literature reports by Lemaire et al.,[36,37] Xue et al.,[38] Feng et al.,[39] Gao et al.[40] and Fotsch et al.[41] Finally we investigated the Aureus database[42] and extracted molecules, for which experimental PXR activation data are available. All datasets were merged and harmonized after careful analysis; literature classifications were checked for consistency. Each compound was assigned to one of two categories, namely *HIGH* or *LOW* with respect to experimental PXR activation in the literature. However, as assays and classification thresholds vary in different experimental studies, this causes a potential source of uncertainty. It should also be mentioned that PXR activation data could be affected by multiple mechanisms (e.g., solubility, binding to the DNA-binding domain) and not only reflect true activation via binding to the PXR ligand-binding domain. The choice of decision trees for data mining is intended to address these issues.

The combined set of molecules encompassed 434 unique chemical structures (dataset A, Supplementary data) and was used to build the first PXR classification models. The dataset was split by statistical methods employing 2D UNITY fingerprints[43] as descriptors and hierarchical clustering[44] into a training set (380 molecules), a test set for validation of the variable selection approach described below (25 molecules) and an external set (29 compounds) to estimate the classification performance of the model on novel structures. Hierarchical clustering using a complete linkage approach was performed using Selector in Sybyl.[43] The distance between the most distant pairs in both clusters is used for merging. Compounds next to the cluster centers were then selected as test or external set in separate clustering runs.

One essential prerequisite for successfully applying QSAR models to previously unknown molecules is that candidates for prediction should not be too dissimilar compared to the training set for constructing the model. By applying a statistical approach for test set selection, we ensure that validation of the model is performed within its similarity boundaries. This similarity to training set compounds then also defines the applicability domain[45,46] of model-based predictions.

For constructing a second classification model, we added internal molecules to this first dataset A to arrive at 636 unique chemical structures (dataset B). Our intention was to better cover the chemical space for interesting series in lead optimization. These additional compounds were tested in different internal or external PXR activation assays. When adding in-house data for an augmented qualitative PXR activation model, we consistently classified compounds with a pEC_{50} value $\leqslant 4.75$ as *LOW* and those with a pEC_{50} value >4.75 as *HIGH*. Compounds with low effect in single-point determinations and thus without reliable dose–response data were also classified as *LOW*. As threshold a value <20% activity compared to a reference compound (SR12813)[47] is typically used.

This dataset B with 636 unique chemical structures was again split by statistical methods into a training set (536 molecules), a test set for validation of the variable selection approach described below (50 molecules) and an external set (50 compounds) to estimate the classification performance of the model on novel structures. For consistency, the same test and external set molecules from dataset A were also included in the test and external set for dataset B, respectively.

For a significant number of compounds from dataset B, quantitative PXR activation data were collected from the literature, the Aureus database and from multiple internal PXR activation assay data. Hence, we combined this subset of dataset B as dataset C to derive a quantitative PXR activation model for a focused region of the chemical space. This dataset C with 306 unique chemical structures, which are all contained also in dataset B, was split into a training set (240 molecules), a test set for validation of variable selection (33 molecules) and an external set (33 molecules) to evaluate the performance of the model. However, it should be noted that in this dataset C the EC_{50} values originate from different assays or were estimated from % activation data in literature or internal assays. A further complication towards generation of quantitative models for a nuclear hormone receptor is that any definition of experimental activation should not only consider EC_{50} from dose–response curves, but also the level of maximal PXR activation compared to a reference, which often is rifampicin. However, this level of activation might also vary and is not always known, which again makes a clear ranking of individual molecules difficult. Hence, the investigated dataset C is expected to suffer from some of these issues. Nevertheless, the obtained quantitative model might be useful to provide first SAR trends or rankings for novel molecules, which are chemically similar to the training set.

2.2. Molecular descriptors

Pretreatment of all molecules started from the connection table and includes removal of counterions and smaller fragments, neutralization plus canonization of chemical structures. Canonical 3D geometries including hydrogen positions were then generated using the program CORINA.[48,49]

For each molecule structure-derived descriptors were computed using the following program packages, namely 185 descriptors from MOE,[50] 128 from Volsurf+[51–54] and 82 descriptors from Parasurf.[55,56] Moreover, we computed 191 CATS[57] derived topological pharmacophore descriptors based on an internal implementation.[58] In previous investigations, we have systematically compared multiple descriptor blocks with regard to their model building performance. For the PXR classification in this study, either the combination MOE, Volsurf, Parasurf or MOE, Volsurf, CATS proved effective.

Prior to calculating semiempirical Parasurf descriptors, all molecules were subjected to geometry optimization using the AM1 Hamiltonian[59] in Mopac6[60] (keywords: AM1, EF) starting from their canonical 3D CORINA conformation. Molecular surfaces and descriptors were calculated with Parasurf directly from the Mopac6 output. In the first approach, the standard descriptors as described in the Parasurf manual[56] were used. For the surface integral model, the SIM descriptors[61] were calculated. Parasurf can calculate several different types of molecular surfaces. The models described below used the marching-cube algorithm[62] for the surface calculation based on the default isodensity value of 0.003.[63] The local properties calculated are the molecular electrostatic potential (*MEP*), local ionization energy (IE_L), local electron affinity (EA_L), local hardness (η_L), local polarizability (α_L), and the local electrostatic field normal to the surface (F_N).[64] Our final parameter settings for calculating Parasurf descriptors include the following options: *surf = cube, contour = isoden, fit = isod, iso = 0.003*. The entire Parasurf descriptor calculation workflow was wrapped in internal Python scripts to allow for parallel processing.

Automation and parallelization of the entire molecular preparation and descriptor calculation workflow was done using internal Python and Perl scripts.

2.3. Dataset analysis

For analysis of the input datasets, a principal component analysis (PCA)[65–67] was performed using Sybyl[43] using MOE[50] descriptors. A PCA contracts the large number of collinear variables to a few orthogonal 'principal properties'. The original data matrix is approximated by the product of two smaller matrices, namely scores and loadings.[68] The score matrix gives a simplified picture of the objects represented by a few uncorrelated new variables (PC-Scores). The first new coordinate describes the maximum variance among all possible directions, the second one the next largest variation among all directions orthogonal to the first one. All descriptors were subjected to autoscaling for normalization.

2.4. Model building

The entire compound and descriptor matrix from individual datasets was used to build classification models from assigned activity classes using the program C5.0.[69,70] This data mining technique based on decision trees was validated and useful in other internal and external studies[71] for discovering patterns that delineate categories, assembling them into classifiers, and using these classifiers for predictions on novel molecules. Such a decision tree approach is intrinsically able to analyze high-dimensional data affected by multiple underlying mechanisms. Irrelevant descriptors are ignored during tree development and pruning. These models are less complex then those derived using neural networks or other non-linear approaches, which makes a chemical interpretation possible. No further pruning options available in C5.0 were used.

In order to improve classification model quality, expressed as total error of the C5.0 model, we implemented a genetic algorithm (GA) based descriptor selection approach using Perl based on its module *AI::Genetic*.[72] The presence of all descriptors was encoded in a bitstring (1: use; 0: ignore) and the GA was used to optimize individual bit settings. For optimization, we attempted to minimize the combined training and test set errors, to which equal weights were assigned in all runs. The GA-based feature selection was performed using a roulette-wheel approach with a two-point crossover and a crossover rate of 0.9. The GA mutation rate was set to 0.01, while the population size for individual configurations at each GA generation was set to 100. A total of 1000 iterations were performed, while typically convergence was achieved at earlier stages. This means that C5.0 runs 100 times for each iteration, depending on the population size. An additional pruning step was performed for every 10 iterations, thereby directly eliminating irrelevant descriptors in the best model by a simple scanning approach. The pruned model is then added to the start population for the next generation.

In a similar way quantitative models were built using the program CUBIST,[73,74] which handles continuous dependent variables (here: pEC_{50} for PXR activation) by a regression tree approach. The basic principle is to first construct a rule-based decision tree, where each rule has an associated MLR model describing the structure–activity relationship (SAR) for all molecules belonging to this particular node of the decision tree.[75] The tree structure determines the assignment of a molecule to a class. Hence, CUBIST classifies molecules using structural parameters according to rules and evaluates a separate SAR model for each subset, rather then fitting a single model to the entire dataset. This approach is able to overcome the lower accuracy of decision trees. For variable selection, a similar GA-based approach was implemented, attempting to maximize the sum of regression coefficients for training and test sets as target function, to which again equal weights were assigned. The GA-based feature selection was performed using the same settings and for the decision tree, namely employing a roulette-wheel approach with a two-point crossover and a crossover rate of 0.9. The GA mutation rate was set to 0.01, while the population size for individual configurations at each GA generation was set to 100. A total of 1000 iterations were performed.

3. Results and discussion

3.1. Dataset characteristics

In order to compare both input datasets A and B with 434 and 636 compounds, respectively, we computed MOE descriptors for all molecules and performed a PCA after autoscaling of the descriptor matrix. The first three PCA components were kept for interpretation of the PC score matrix, as they cumulatively explained 35.1%, 46.2% and 56.4% of the total variance in the descriptor block.

The first two PC scores in Figure 1 for all 636 molecules serve to illustrate the structural diversity of this dataset. The color-coding in the graph on the left indicates 434 molecules from dataset A in blue and the additional compounds from dataset B in red. This distribution shows that dataset A spreads out significantly in PCA space, suggesting a broader coverage of the derived model, while the added compounds for dataset B (red) are densely grouped in a region close to the center of this graph. This reflects the presence of multiple congeneric series, which then might allow adding specificity and possibly extracting more detailed SAR information in this region of covered chemical space.

The color coding in the graph on the right in Figure 1 shows the distribution of compounds from the combined dataset used for training (blue), for test set (green) and as external set (red) to only evaluate the performance of the resulting models. This graph suggests an equal coverage of the PCA space; it should be noted that test set selection was based on different descriptors (2D fingerprints and clustering).

For both datasets A and B, the distribution of characteristic properties like molecular weight, log*P* and topological PSA (TPSA),[76] is displayed in histograms in Figure 2. All data were computed using MOE. The histograms show an almost similar distribution for both datasets. A more detailed analysis shows that the internal compounds added to dataset B are predominantly drug-like molecules with mean values for molecular weight of 491, log*P* of 5.5 and TPSA of 86. However, the high mean log*P* value suggests that also some very lipophilic compounds are present.

3.2. Classification model for PXR activation from dataset A

First we developed a qualitative classifier model to identify molecules that can potentially activate PXR. To this end we combined multiple data sets from the literature and databases containing public domain structures, as described above. All datasets were merged; each compound was assigned to either *HIGH* or *LOW* categories, based on PXR activation information provided in the literature. The harmonized dataset of 434 chemical structures served to build the first PXR classification models (dataset A). The dataset was split by statistical methods into a training set (380 molecules), a test set for validation of variable selection (25 molecules) and an external set (29 compounds) to evaluate classification performance.

A total of 395 descriptors were calculated for all molecules using MOE, Volsurf and Parasurf. C5.0 classification models using all descriptors and the training set were constructed for dataset A. These initial models were optimized using the GA-variable selection procedure using equal weight of training and test set errors as target function (1000 iterations, 100 populations).

The final classification model for dataset A after GA-based variable selection contains 29 relevant descriptors: 14 derived using MOE, 12 from Volsurf and 3 from Parasurf. The internal quality of this classification model is summarized in Figure 3 (left) for

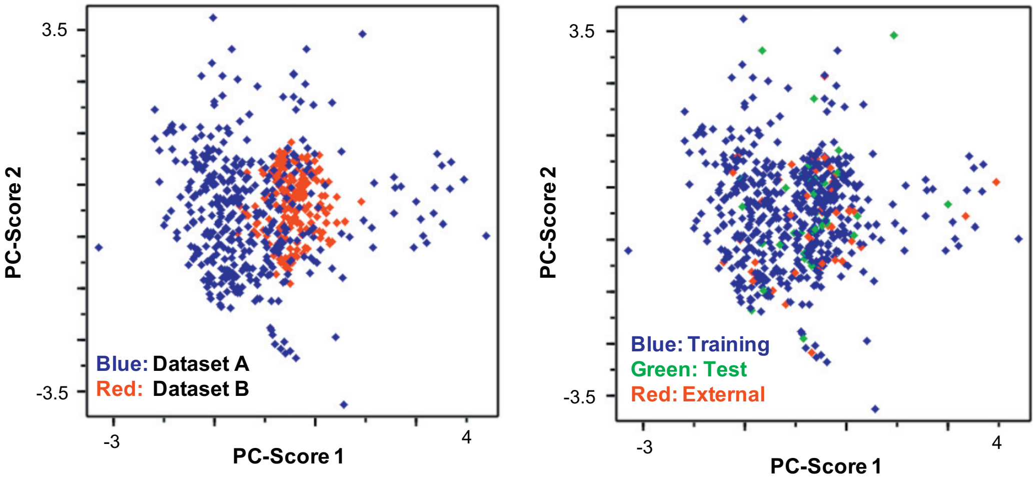


Figure 1. PCA score plot for 636 molecules to illustrate the structural diversity of this dataset. Left panel: Blue points indicate 434 molecules from dataset A, red indicates additional molecules in dataset B. Right panel: Distribution of compounds from the combined dataset used as training set (blue), as test set (green) and as external set (red).

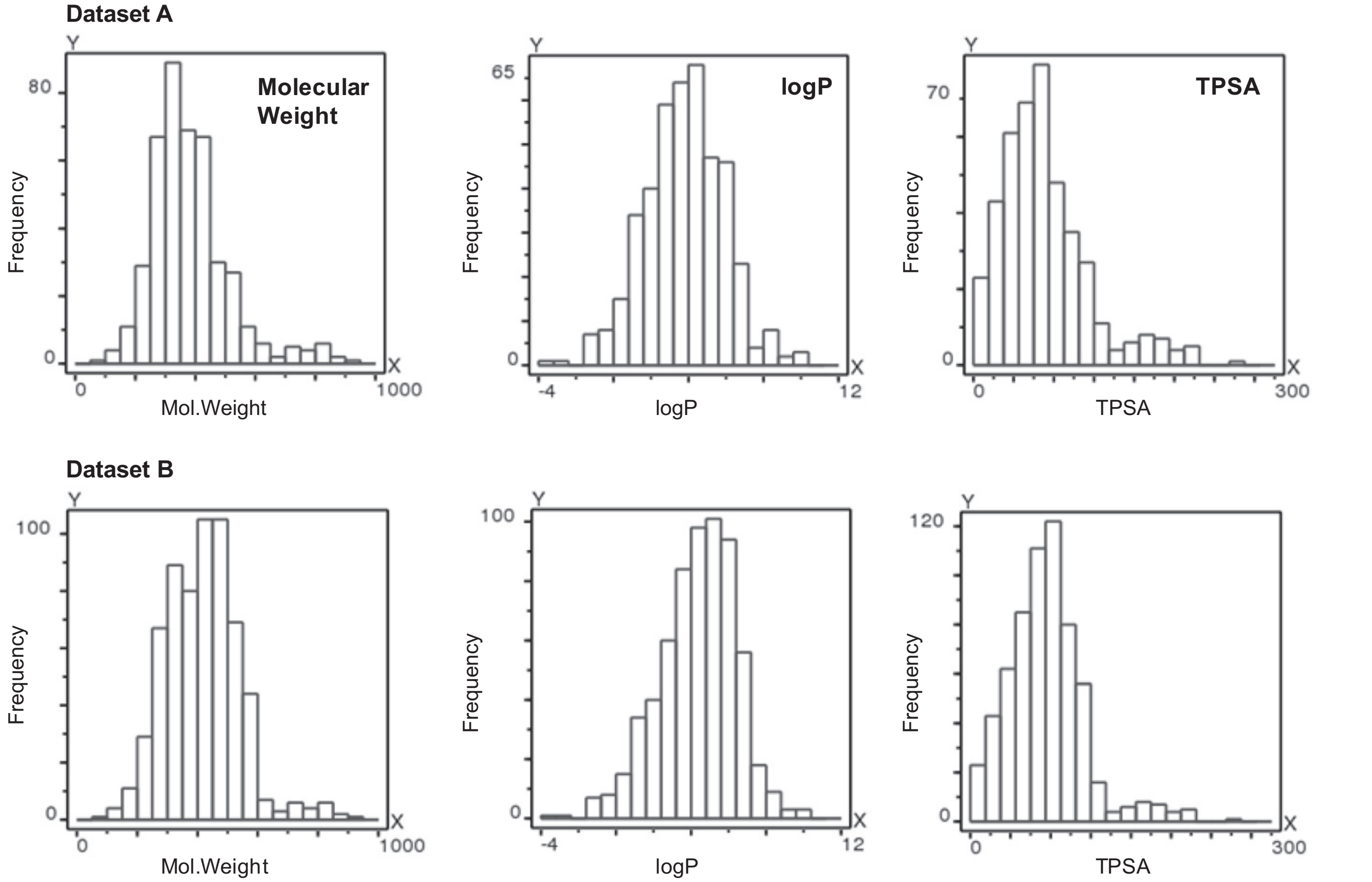
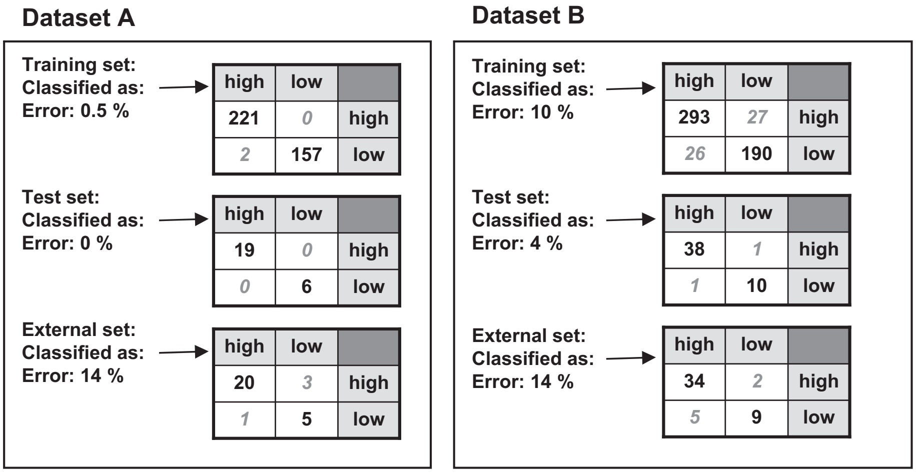


Figure 2. Distribution of characteristic physicochemical properties as histograms (left: molecular weight; middle: log*P*; right topological PSA (TPSA)) for both datasets A (upper panel) and B (lower panel).

the training and test sets. This internal model quality is very good with a correct classification of the training set of 100% for PXR activators (true high) and 99% for non-activators (true low). The corresponding confusion matrix is shown in the left upper panel in Figure 3. The model's performance for the test set is equally good with both 100% for PXR activators and non-activators (Fig. 3, left middle panel). More important for the assessment of the model quality is its performance on a true external test set, which was not used to derive the classification model or partially guide any descriptor selection. Here, the presented model correctly predicts the external test set of 29 compounds with 87% correct classification for PXR activators and 83% correct classification for non-activators, as shown in detail in the left lower panel in Figure 3. A total classification error of 14% indicates a stable and significant model. It should be noted that due to the test set selection using only chemical descriptors, an imbalance of activators and non-



Dataset A

Training set:
Classified as:
Error: 0.5 %

high	low	
221	0	high
2	157	low

Test set:
Classified as:
Error: 0 %

high	low	
19	0	high
0	6	low

External set:
Classified as:
Error: 14 %

high	low	
20	3	high
1	5	low

Dataset B

Training set:
Classified as:
Error: 10 %

high	low	
293	27	high
26	190	low

Test set:
Classified as:
Error: 4 %

high	low	
38	1	high
1	10	low

External set:
Classified as:
Error: 14 %

high	low	
34	2	high
5	9	low

Figure 3. Quality of C5.0 derived classification models for dataset A (left) and dataset B (right). For each dataset, confusion matrices for their performance on the training (upper panel), test (middle panel) and external set (lower panel) are shown.

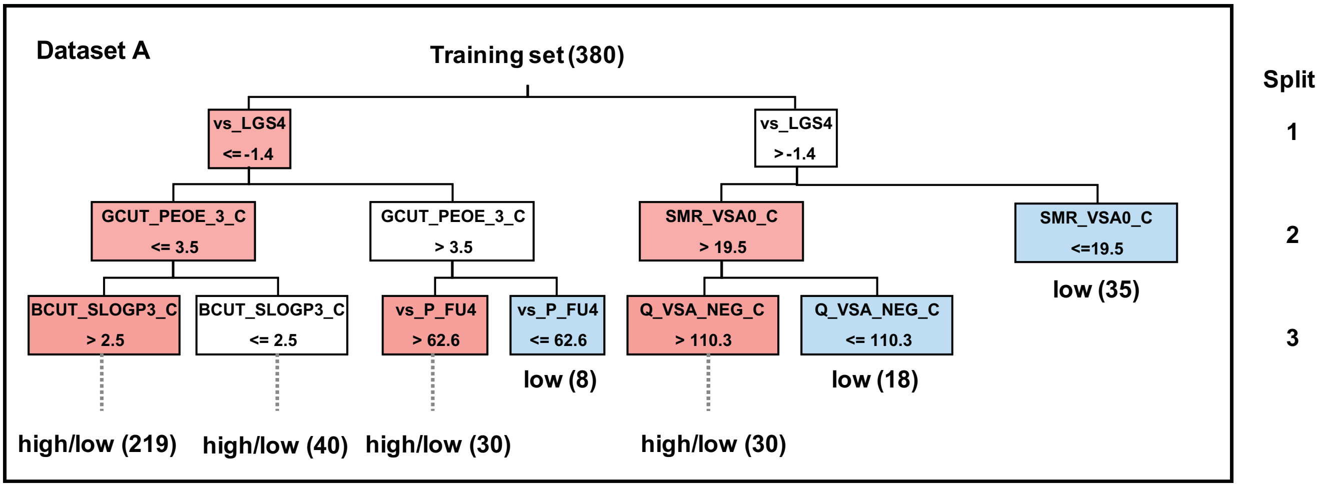


Figure 4. Final decision tree for classifier derived using dataset A with major branches in the first three splits. Nodes enriched with PXR activators are colored in red, while nodes enriched in PXR non-activators are colored blue. Terminal nodes lack dotted lines, while dotted lines indicate further branches of the final tree.

activators is present in training and validation sets, which might be related to the slightly higher classification error for non-activators.

In order to further validate this model, the assignment of biological activity class was randomized 100 times, while maintaining the actual distribution of PXR activators and non-activators. After model development, a mean classification error of 37% for the randomized training sets and 24% for the randomized test sets compared to 0.5% and 0% for the final model (see above) indicate a relevant original classifier, which appears to be not affected by chance correlation.

The final decision tree is displayed with its major branches in Figure 4. Nodes, which are enriched in PXR activators, are colored in red, while nodes enriched in PXR non-activators are colored blue. Dotted lines indicate further branches in the final decision tree, which are not shown. The entire tree for dataset A is summarized in the Supplementary data. The initial split is based on the Volsurf (prefix: *vs_*) descriptor LGS4 (solubility at pH4) with its left branch enriched in PXR activators. The next split then involves the MOE descriptor GCUT_PEOE_3, which relates to the partial charge distribution.[77] When following the left branch of the classifier, the BCUT_SLOPGP3 descriptor (atomic log*P* distribution)[78] is responsible for the next split. When following the right branch of the classifier instead, the Volsurf descriptor P_FU4 (percentage of non-ionized species at pH 4.0) provides the next split to one terminal node with non-activators (blue). When following the initial vs_LGS4 split to the right, the next relevant split is based on the descriptor SMR_VSA0 (MOE surface contribution for molecular refractivity) again providing a terminal node for PXR non-activators on the right branch. The next split on the left hand of this tree is then provided by the descriptor Q_VSA_NEG (MOE) leading once again to a terminal node. This descriptor relates to the surface area of atoms with a negative partial charge.[79]

In order to evaluate the particular influence of individual descriptors to model performance, we excluded each of the final 29 descriptors once before deriving a new classifier and analyzed the differences in classification errors for training and test sets with and without this descriptor. This classification error difference is computed by subtracting the classification error for this reduced model from the classification error for the original model, which results in negative absolute differences. Those descriptors with a very significant influence on the final classifier tend to be more important to capture the SAR in the dataset. As the implemented GA could potentially also keep descriptors without any influence on predictivity, such an analysis further highlights irrelevant descriptors. For 19 of 29 descriptors in the final model, this classification error difference is <−3 for the test set. In particular, this was observed for 9 of 13 MOE-descriptors, 7 of 12 Volsurf-descriptors and all 3 Parasurf-descriptors, respectively. All descriptors from this evaluation study are summarized in Table 1.

The most important descriptor from this analysis with a classification error differences in the test set of −20 is the MOE GCUT_-PEOE_3 descriptor on PEOE partial charge distribution, followed by Volsurf derived solubility at pH 4 (−16.0) and pH 8 (−8.0). Further descriptors with a classification error differences in the test set of −8 are the MOE SLOGP_VSA3 descriptor and the Volsurf amphiphilic moment (vs_A). These subdivided surface area descriptors like for SLOGP_VSA3 in MOE are based on an approximate accessible vdW surface area calculation for each atom along with atomic properties from a connection table approximation. Each descriptor refers to the sum of the atomic surface areas over all atoms with properties in a specific range, here a polar log*P* contribution between 0.0 and 0.1.

Other relevant contributions to the classifier are the following descriptors, all with a classification error differences in the test set of −4 plus significant contributions in the training set, namely BCUT_SLOGP3 (see above), B_TRIPLE (number of triple bonds), vs_W8 (Volsurf hydrophilic regions computed at different *OH* probe energy levels), vs_CD5 (Volsurf capacity factor for the *DRY* probe indicating the concentration of hydrophobic sites on the molecular surface at the certain energy level), vs_DD6 (Volsurf hydrophobic volume differences) and Parasurf FNvar-, which describes the variance in the negative electrostatic field normal to the molecular surface. This statistical analysis of the classifier plus important descriptors to the classification result suggests a marked influence of primarily those descriptors capturing solubility, lipophilic properties, the balance between lipophilicity and polarity and some marked electrostatic terms.

3.3. Classification model from combined dataset B

In order to augment this model and adapt the training set to interesting regions of our chemical space, we added compounds from internal projects tested for PXR activation to dataset A. This merged dataset of 636 chemical structures (dataset B) served to build an augmented PXR classification model after splitting into a training set (536 molecules), a test set for validation of variable selection (50 molecules) and an external set (50 compounds) to estimate classification performance. All members of the test and

Table 1
Summary of important descriptors for classification model of dataset A[a]

Descriptor	Type	Classification error for training set	Classification error for test set	Error difference for training set	Error difference for test set
Original	ALL	0.5	0.0	0.0	0.0
BCUT_SLOGP_3	MOE	3.2	4.0	−2.7	−4.0
B_TRIPLE	MOE	3.9	4.0	−3.4	−4.0
CHIRAL_U	MOE	0.5	0.0	0.0	0.0
GCUT_PEOE_3	MOE	4.2	20.0	−3.7	−20.0
KIERA3	MOE	0.8	4.0	−0.3	−4.0
LIP_DRUGLIKE	MOE	0.8	4.0	−0.3	−4.0
PEOE_VSA2	MOE	0.8	0.0	−0.3	0.0
PEOE_VSA_NEG	MOE	0.8	0.0	−0.3	0.0
Q_VSA_FPOL	MOE	2.6	0.0	−2.1	0.0
Q_VSA_NEG	MOE	0.8	4.0	−0.3	−4.0
SLOGP_VSA3	MOE	3.7	8.0	−3.2	−8.0
SMR_VSA0	MOE	1.1	4.0	−0.6	−4.0
SMR_VSA1	MOE	0.5	4.0	0.0	−4.0
SMR_VSA3	MOE	0.8	0.0	−0.3	0.0
vs_G	Volsurf	0.5	0.0	0.0	0.0
vs_W8	Volsurf	3.9	4.0	−3.4	−4.0
vs_D3	Volsurf	1.1	0.0	−0.6	0.0
vs_D5	Volsurf	0.5	0.0	0.0	0.0
vs_D8	Volsurf	0.5	0.0	0.0	0.0
vs_CD5	Volsurf	2.9	4.0	−2.4	−4.0
vs_A	Volsurf	3.9	8.0	−3.4	−8.0
vs_P_FU4	Volsurf	0.8	4.0	−0.3	−4.0
vs_P_FU10	Volsurf	1.1	0.0	−0.6	0.0
vs_LGS4	Volsurf	3.9	16.0	−3.4	−16.0
vs_LGS8	Volsurf	4.5	8.0	−4.0	−8.0
vs_DD6	Volsurf	4.5	4.0	−4.0	−4.0
ENEGmin	Parasurf	1.1	4.0	−0.6	−4.0
ENEGskew	Parasurf	0.3	4.0	0.2	−4.0
FNvar-	Parasurf	3.9	4.0	−3.4	−4.0

[a] Computed by subtracting the classification error for reduced model without this descriptor from the classification error for the original model.

external set for dataset A were also members of the corresponding sets for dataset B.

A total of 504 descriptors were calculated using MOE, Volsurf and CATS. C5.0 classification models were then constructed for dataset B using the same GA-descriptor selection procedure after satisfactory initial classification models. A model developed in parallel using dataset B, but MOE, Volsurf and Parasurf descriptors in accordance to the best model for dataset A revealed a slightly lower performance and was not followed further (data not shown).

The final classification model for dataset B after GA-based variable selection contains 21 relevant descriptors: 9 derived using MOE, 7 from Volsurf and 5 from CATS. The quality of this model is summarized in Figure 3 (right). This internal model quality is again acceptable with a correct classification of the training set of 92% for PXR activators (true high) and 88% for non-activators (true low), resulting in a 10% classification error. The corresponding confusion matrix is shown in the right upper panel in Figure 3. The model's performance for the test set is slightly better with a correct classification of 97% for PXR activators and 91% for non-activators, resulting in a 4% classification error (Fig. 3, right middle panel). Applying this model to the external set of 50 compounds led to a correct classification of 94% for PXR activators, but only 64% for non-activators, as shown in detail in the right lower panel in Figure 3. A total classification error of 14% for the external set indicates a stable and significant model. This model performance still qualifies its application as virtual screening filter towards a reliable alert for potential PXR activators.[30]

For further validation the biological activity class assignment was randomized for 100 times. After model development, a mean classification error of 32% for the randomized training sets and 32% for the randomized test sets compared to 10% and 4% for the final model suggests a significant original classifier unaffected by chance correlation.

In order to evaluate the influence of individual descriptors to model performance, we excluded again each of the 21 descriptors once and analyzed the differences in classification error for training and test sets with and without this descriptor. For 10 out of 21 descriptors in this model this classification error difference is <−3 for the test set. This was found for 4 of 9 MOE-descriptors, 3 of 5 CATS-descriptors and 4 of 7 Volsurf-descriptors, respectively. These results are summarized in Table 2.

The most important descriptors from this analysis with significant classification error differences are related to solubility, lipophilicity, electrostatics and size. Solubility is captured by the MOE LOGS descriptor in addition to Volsurf solubility at pH 7 (vs_LGS7). Lipophilicity is accounted for by MOE SLOGP_VSA8 as surface area of atoms with a certain $\log P$ (here for very lipophilic atoms), CATS PL6 as normalized occurrence of positive to lipophilic atom distances with a distance of 6 bonds, vs_DD4 as Volsurf hydrophobic volume differences and vs_CD7 as Volsurf capacity factor for the *DRY* probe (see above). Electrostatic properties are captured by the MOE descriptors GCUT_PEOE_0 (see above) and PEOE_VSA_FPNEG, which relates to the normalized negative polar vdW surface area for atoms with a partial charge <−0.2.

In addition to this, the size of the molecules is reflected by the CATS shape descriptors. Here SH18 refers to a CATS shape descriptors indicating large molecules. This shape implementation in our own CATS version[58] is based on a normalized spatial autocorrelation function[80] providing the normalized sum of all occurrences of heavy atom interactions at a particular graph distance. A value of 18 refers to very large molecules.

3.4. Quantitative QSAR model for PXR activation

In order to derive a quantitative model for dataset C, this dataset of 306 compounds was split into a training set (240 molecules), a test set for validation of variable selection (33 molecules) and an external set (33 compounds) to estimate the performance of the model. We then employed the program CUBIST to derive regression tree models based on the training dataset and all 395 MOE, Volsurf

Table 2
Summary of important descriptors for classification model of dataset B[a]

Descriptor	Type	Classification error for training set	Classification error for test set	Error difference for training set	Error difference for test set
Original	ALL	9.9	4.0	0.0	0.0
GCUT_PEOE_0	MOE	11.0	10.0	−1.1	−6.0
GCUT_SLOGP_2	MOE	11.4	6.0	−1.5	−2.0
LOGS	MOE	14.0	24.0	−4.1	−20.0
PEOE_VSA_0	MOE	11.6	4.0	−1.7	0.0
PEOE_VSA_FPNEG	MOE	11.4	10.0	−1.5	−6.0
Q_VSA_NEG	MOE	10.8	4.0	−0.9	0.0
SLOGP_VSA3	MOE	10.4	4.0	−0.5	0.0
SLOGP_VSA8	MOE	15.5	30.0	−5.6	−26.0
SMR_VSA6	MOE	10.1	4.0	−0.2	0.0
PL6	CATS	18.8	26.0	−8.9	−22.0
AA1	CATS	15.7	14.0	−5.8	−10.0
D	CATS	9.9	4.0	0.0	0.0
SH1	CATS	11.4	6.0	−1.5	−2.0
SH18	CATS	14.4	22.0	−4.5	−18.0
vs_W3	Volsurf	11.0	8.0	−1.1	−4.0
vs_WO1	Volsurf	10.8	12.0	−0.9	−8.0
vs_ID1	Volsurf	11.9	4.0	−2.0	0.0
vs_CD7	Volsurf	14.0	12.0	−4.1	−8.0
vs_HL1	Volsurf	10.6	4.0	−0.7	0.0
vs_LGS7	Volsurf	13.6	6.0	−3.7	−2.0
vs_DD4	Volsurf	14.4	22.0	−4.5	−18.0

[a] Computed by subtracting the classification error for reduced model without this descriptor from the classification error for the original model.

and Parasurf descriptors. After first predictive regression tree models, we selected relevant descriptors using a genetic algorithm in order to improve the statistical quality of the training set in terms of regression coefficients.

The finally selected regression tree model is based on 13 rules in a decision tree, each associated with a linear model plus 37 relevant descriptors selected in total by our GA procedure. While for the training set a r^2 value of 0.865 was obtained, the predictive r^2 of 0.774 for the test set of 33 molecules suggests a significant and useful model. The graph of predicted versus experimental pEC_{50} values is shown in Figure 5 on the left for the training set (blue) and test set compounds (red). The use of 10-fold crossvalidation employing the 240 training set compounds only indicates significant inconsistencies in this set, as indicated by a relatively low crossvalidated r^2 value of 0.292. For validation of this model, the biological activities were thus randomized 100 times, which resulted in a mean r^2 value of only 0.07 (SD 0.02) after deriving regression trees, thus suggesting the original model to be relevant and not affected by chance correlation.

This model was then applied to the dataset of 33 external compounds, which were not used either to derive the model or guide the variable selection, which resulted in a predictive r^2 value of 0.452. If this model is applied in a qualitative manner with a pEC_{50} threshold of 4.75 for PXR activators and non-activators, a total of 86% PXR activators and 73% non-activators are correctly classified.

In order to explore the leveraging effect of the five most active compounds in the training set on the model performance, those were moved for a validation run to the test set and a regression tree model was derived from 235 training and 38 test set compounds. Using the original 37 descriptors, the r^2 of 0.578 and predictive r^2 of 0.185 for the test set suggests a less significant model, primarily due to the lack of predictive power for the most active molecules with an experimental pEC_{50} between 7.8 and 8.5. In a

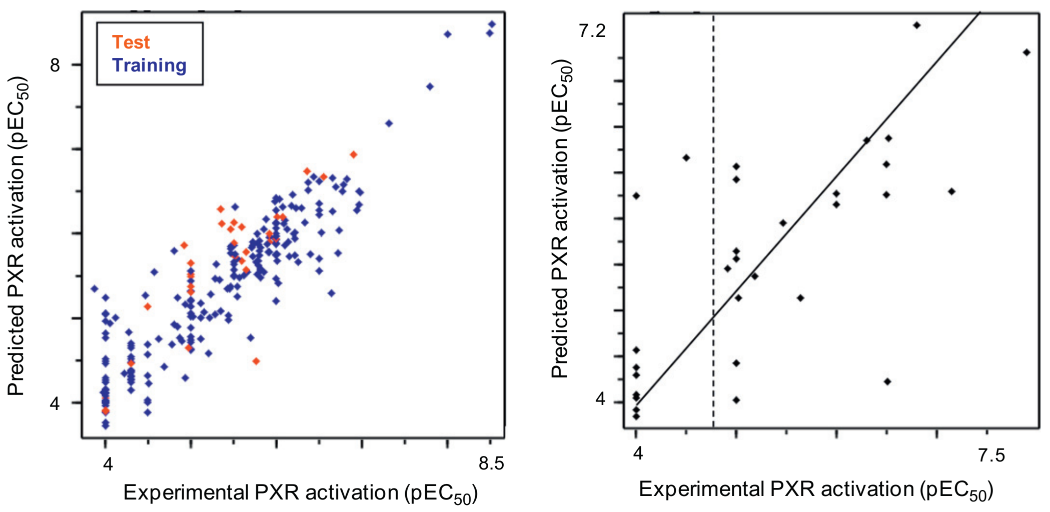


Figure 5. Predicted versus experimental pEC_{50} values for PXR activation for quantitative regression tree model from dataset C. Left: Training set (240 compounds, blue) and test set (33 compounds, red). Right: Predictions for external set (33 compounds) using this model.

second experiment, a new GA variable selection was performed on this changed training/test set assignment starting from all descriptors. The final model using 28 descriptors and 13 rules now shows only a slightly reduced predictivity compared to the original model (r^2: 0.828, q^2: 0.292, predictive r^2: 0.723), with the most active compounds predicted now with pEC_{50} values between 6.9 and 7.3. Although there is a larger deviation from the experimental values, this still suggests at least some degree of extrapolation beyond the activity range of this dataset.

Some pEC_{50} values in this entire dataset were assigned from approximate % PXR activation measurements with high experimental uncertainty. Consequently, the exclusion of 112 compounds with pEC_{50} values assigned to 4.0, 5.0 or 6.0 produces a reduced training and test set of 162 and 22 molecules, respectively. After performing a new GA variable selection, a model with 24 descriptors and 13 rules with a significantly increased statistical performance resulted (r^2: 0.884, q^2: 0.292, predictive r^2: 0.828). This clearly highlights this influence of uncertain data points to the overall regression tree model. However, as our intention is to cover a broader chemical space for predicting novel chemotypes, the model from the larger dataset is further used for analysis.

In order to evaluate the particular influence of individual descriptors to model performance, we excluded each of the 37 relevant descriptors once from the original model and analyzed the differences in corresponding correlation coefficients (r^2) for training and test sets with and without this descriptor for developing the regression tree model. For 10 out of 37 descriptors in this model this r^2 difference is <-0.12 for the test set. This was found for 3 of 14 MOE-descriptors, 2 of 13 Volsurf-descriptors and 5 of 11 Parasurf-descriptors, respectively. All results are summarized in Table 3.

The most important descriptors from this analysis with the most significant r^2 differences <-0.3 again are related to lipophilicity, electrostatics and size. Those descriptors include MOE GCUT_SLOGP_1, which indicates GCUT descriptors derived from using atomic contribution to log *P*. Electrostatic interactions are captured by the MOE PEOE_VSA_0 and PEOE_VSA_6 descriptors, each providing the sum of atomic surface area contributions with particular atomic partial charges. The Volsurf vs_DRDRAC descriptor relates to size and specific interactions: After generating all possible 3D distance triplets between individual atoms, these descriptors indicate the maximum area of the triangles considering all possible conformers derived for this class of pharmacophore triplets, as defined here for DRY–DRY-Acceptor interactions. The relevant Parasurf descriptors again relate to electrostatic effects. EALmin indicates the minimum of the local electron affinity from semiempirical calculations. Low EA_L values are induced typically by the presence of halogen atoms and strong electron-withdrawing groups. var∗balance refers to the product of the total variance of the molecular electrostatic potential (*MEP*) on the molecular surface and the electrostatic balance index ν. This product is related to the strength of non-covalent interactions of a molecule with related molecules.

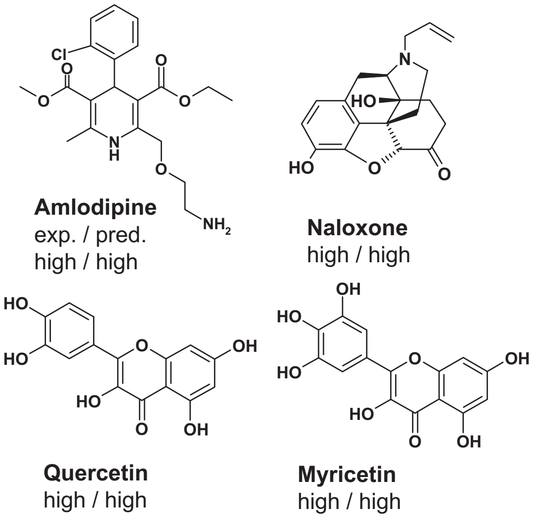
3.5. Interpretation of the classification model from dataset A

The chemical interpretation of the classification model for dataset A (Section 3.2) was performed in two different ways. First we

Table 3
Summary of important descriptors for regression tree model of dataset C[a]

Descriptor	Type	Training set r^2	Test set predictive r^2	Difference r^2 for training set	Difference r^2 for test set
Original	MOE	0.87	0.77	0.00	0.00
A_ACC	MOE	0.87	0.76	0.00	−0.02
A_IC	MOE	0.81	0.67	−0.06	−0.10
CHI0	MOE	0.85	0.77	−0.02	0.00
GCUT_SLOGP_1	MOE	0.69	0.49	−0.18	−0.28
GCUT_SMR_0	MOE	0.87	0.74	0.00	−0.04
PC_M	MOE	0.87	0.77	0.00	0.00
PEOE_VSA_0	MOE	0.72	0.35	−0.14	−0.43
PEOE_VSA_6	MOE	0.36	0.24	−0.51	−0.53
PEOE_VSA_POL	MOE	0.79	0.67	−0.07	−0.10
Q_RPC_M	MOE	0.87	0.77	0.00	0.00
RPC_M	MOE	0.87	0.77	0.00	0.00
SMR_VSA4	MOE	0.79	0.67	−0.07	−0.10
VSA_POL	MOE	0.81	0.69	−0.06	−0.09
vs_R	Volsurf	0.85	0.66	−0.02	−0.12
vs_W6	Volsurf	0.87	0.76	0.00	−0.02
vs_WO2	Volsurf	0.85	0.71	−0.02	−0.07
vs_WO3	Volsurf	0.87	0.77	0.00	0.00
vs_WN1	Volsurf	0.81	0.67	−0.06	−0.10
vs_LOGP_C_HEX	Volsurf	0.87	0.71	0.00	−0.07
vs_PSA	Volsurf	0.81	0.67	−0.06	−0.10
vs_P_FU6	Volsurf	0.85	0.74	−0.02	−0.04
vs_DRDRAC	Volsurf	0.69	0.37	−0.18	−0.40
vs_LGS7_5	Volsurf	0.87	0.71	0.00	−0.07
vs_LGBB	Volsurf	0.87	0.67	0.00	−0.10
vs_METSTAB	Volsurf	0.88	0.71	0.02	−0.07
vs_DD6	Volsurf	0.79	0.64	−0.07	−0.13
meanMEP+	Parasurf	0.79	0.61	−0.07	−0.17
var∗balance	Parasurf	0.69	0.36	−0.18	−0.41
MEPskew	Parasurf	0.85	0.66	−0.02	−0.12
IELmax	Parasurf	0.81	0.62	−0.06	−0.15
EALmin	Parasurf	0.62	0.31	−0.24	−0.46
EALskew	Parasurf	0.85	0.71	−0.02	−0.07
HARDmax	Parasurf	0.85	0.67	−−0.02	−0.10
HARDbar	Parasurf	0.83	0.71	−0.04	−0.07
HARDrange	Parasurf	0.81	0.69	−0.06	−0.09
FNvartot	Parasurf	0.77	0.64	−0.09	−0.13
FN+	Parasurf	0.87	0.77	0.00	0.00

[a] Computed by subtracting the regression coefficients (r^2) for reduced model without this descriptor from coefficient for the original model.



Decision tree node	
vs_LGS4	> -1.42517
SMR_VSA0	> 19.4683
Q_VSA_NEG	> 110.338
KIERA3	≤ 5.3651
LIP_DRUGLIKE	> 0
PEOE_VSA_NEG	≤ 205.085
PEOE_VSA2	> 12.9784
SMR_VSA3	> 21.1788

Figure 6. Chemical structures grouped together in a single terminal node from the classification model based on training set of dataset A. All compounds are PXR activators (class: HIGH) and are correctly predicted. Eight rules are used by the classifier for this grouping.

analyzed the chemical structures, which are grouped together in particular nodes from the final decision tree, derived using the training set of 380 molecules. In a second step, we systematically investigated the classification performance on chemically related compounds which were experimentally assigned to differing activity classes.

One very typical example for structures grouped in a single terminal node from the classification model of dataset A is shown in Figure 6 with the eight rules leading to this decision. These compounds are reported as PXR activators (class: *HIGH*) and are also correctly classified by the model. These molecules are chemically less related in terms of 2D similarity except for the polyphenole substances Quercetin and Myricetin, but they all share the same set of rules, reflecting specific requirements on solubility (vs_LGS4), shape (KIER3 as third kappa shape index[81] from MOE connectivity indices) and electrostatics (Q_VSA_NEG, PEOE_NSA_NEG, PEOE_VSA2) in addition to violations of the “rule-of-5”[82] (LIP_DRUGLIKE). Members for other nodes share the same or even a lower degree of 2D similarity.

Our subsequent analysis is focused on identifying compounds which are related by 2D fingerprint similarity,[43] but belong to different experimental activity classes. Predictions from the classifier for these compounds were then analyzed to unveil the performance of the classifier from Section 3.2. Five examples are shown in Figure 7A–E, all with correct classifications as either PXR activator or non-activator. Here the first activity class string (*high*/*low*) always indicates the experimental assignment, while the second value refers to the predicted class. For the following examples, we systematically analyzed descriptor values and rules in correspondence with experimental activity classes to identify discriminating features between these related molecules.

In Figure 7A, three related nitro-trifluoro-phenyl derivatives from the training set of dataset A are shown. While the oral nonsteroidal antiandrogen drug flutamide[33] is correctly predicted as PXR activator, the other two related drugs nilutamide[31] and nitisinone[31] (4-hydroxyphenylpyruvate oxidase inhibitor) are correctly predicted as non-activators. This classification is driven by the Volsurf descriptor vs_W8 (Volsurf hydrophilic regions computed at different *OH* probe energy levels), where values >1.5 indicate non-activators in a terminal node, as observed for Nilutamide and Nitisinone.

Figure 7B summarizes a series of four 2-amino-1,3-thiazol-4(5*H*)-ones reported by Fotsch et al.[41] as 11β-HSD1 inhibitors with different activities in the reported PXR luciferase reporter gene assay. While am008 and am014 are correctly predicted as PXR activators (IDs according to Fotsch et al.), am012a and am033a are much less active in the experimental assay reported in the literature. Those compounds were assigned by us to the class *LOW*; the model’s classification is therefore correct. The primary reason to predict am014 as PXR activator is that the descriptor vs_DD6 adopts values lower than 0.125 as threshold (Volsurf hydrophobic volume differences), while am012a is predicted as non-activator due to values for the descriptors vs_D8 >3.75 (Volsurf strong hydrophobic regions) and PEOE_VSA_NEG >205.08501 as part of a terminal node (MOE total vdW surface area for atoms with a negative partial charge). Finally the prediction of am033a as non-activator is driven by the Parasurf descriptor FNvar- >325.29199 as part of a terminal node (variance in field normal to the surface for all negative values).

Figure 7C summarizes a series of benzodiazepines and close analogs with a significant variation in PXR activity. While midazolam[33] and triazolam[33] are correctly classified as PXR activators, the other two molecules alprazolam[31] and diazepam[31] are correctly classified as PXR non-activators. It should be mentioned that triazolam was part of the external set for evaluation of this model, which again underscores the performance of this model. The discrimination in the training set was driven by the MOE descriptor Q_VSA_FPOL $\leqslant$ 0.129978 as part of a terminal node (MOE fractional polar vdW surface area as sum over atomic surface areas with an absolute partial charge greater than 0.2 divided by the total surface area) and the threshold of 0.059579 for vs_CD5 (Volsurf capacity factor for the DRY probe).

In Figure 7D, four related mono-substituted phtalates[31] from the training set of dataset A are shown. While the 2-ethylhexyl and the benzyl substitution at R1 forming the corresponding ester result PXR activators, the methyl and *n*-butyl esters are much less active. The final model classifies all four compounds from the training set of dataset A correctly. Here, not a single rule but the interplay of multiple descriptors provides a discrimination in activity class prediction, involving again FNvar- in a terminal node, BCUT_SLOGP_3 (MOE BCUT eigenvalues derived descriptors using log*P* atom contributions), vs_W8 (Volsurf strong hydrophilic regions) and Q_VSA_NEG (total vdW surface area for atoms with a negative partial charge).

Figure 7E finally summarizes a series of N1-substituted triazoles[36] with the first three compounds c2ba-5, c2ba-6 and

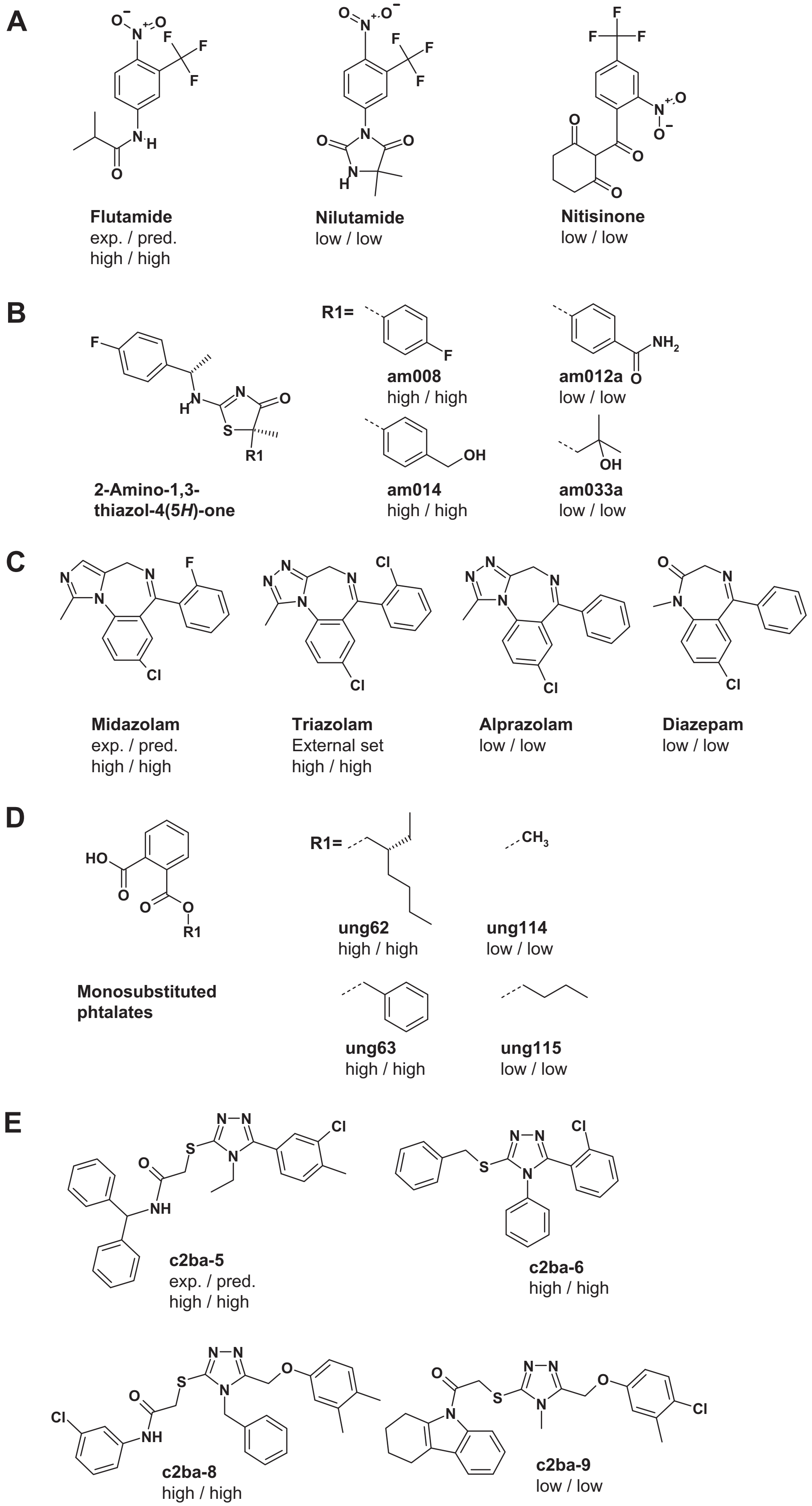


Figure 7. Representative examples of similar compounds with belonging to different experimental activity classes along with predictions from the classifier derived using dataset A. The first activity class indicates the experimental assignment, the second value always refers to the predicted class. See text for details.

c2ba-8 correctly classified as PXR activators, while small variations in particular on the left side of this series led to a correct classification for molecule c2ba-9 as PXR non-activator. This differentiation in the training set is driven primarily by the descriptor vs_P_FU4 (Volsurf percent unionized species at pH 4.0), where values <62.61949 in this terminal node indicate PXR non-activators, as observed for c2ba-9 in this series. Additional differences in other descriptors mentioned above were also found.

Hence, for the above examples, each discrimination of PXR activators versus non-activators appears to be related to different descriptors and to combinations of these differences. SAR rules for analog series can thus only be derived from an analysis of rules, as demonstrated above. It should also be mentioned that for a few other cases with large number of analog molecules belonging to a single activity class, a correct classification for the corresponding other activity class is sometimes not successful, especially, if this analog with a different class assignment is only present in the test or even external set (no data given). However, these examples collectively support our finding that a reliable classification of PXR activators might be possible within chemically related series.

3.6. X-ray crystal structures and previous QSAR models

The first determination of the three-dimensional structure of the PXR ligand-binding domain (LBD)[83] revealed its large, spherical ligand-binding cavity that allows to interact with a diverse range of ligands.[84,85] Furthermore these studies also revealed the adaptability of this binding site, when interacting with a broad range of ligands from small to very large ones. In the meanwhile several other X-ray structures for large and smaller ligands were published[86–88,38] and added a lot to our current understanding of prerequisites for binding and promiscuity[89] of this antitarget. The X-ray structure for the PXR-T0901317 complex (PDB 2O9I, resolution 2.8 Å) by Xue et al.[38] provides an instructive example of a small, but very potent PXR activator binding to the ligand-binding domain via a few hydrogen bonding interactions, aromat–aromat interactions plus a significant contribution of hydrophobic complementarity. The binding site is largely hydrophobic in nature, but contains a few polar residues able to engage in hydrogen-bonding interactions. For example hydrogen-bonds are mediated via sulfonamide oxygen atoms and a polarized benzylic hydroxyl-function in the PXR–T0901317 complex. The PXR ligand-binding domain is also reported to be highly flexible.[83] These prerequisites for ligand-binding, in particular the marked influence of more negatively charged functional groups possibly involved in hydrogen-bonding interactions, is partially reflected in the models from our study.

Previous computational models range from ligand-based pharmacophores,[34,90,91] QSAR models[92,93] and machine-learning approaches[31,33] to homology modeling with molecular dynamics[94] and protein–ligand docking workflows.[40,93,95] The majority of previous computational approaches focused on a few diverse agonist scaffolds and some structural analogues. Consensus of different models is that PXR agonists are required to match multiple hydrophobic features and at least one hydrogen-bond acceptor. In some cases an additional hydrogen-bond donor feature is also reported to be important to agonistic activity. These pharmacophore models are in general consistent with available X-ray structures of PXR–ligand complexes. However, the pharmacophore models have employed typically a limited number of structurally very diverse ligands in their training set, which are additionally measured in multiple laboratories using different experimental protocols. This limitation often only allows for a classification model. Ekins et al.[34] have described a pharmacophore model based on 30 steroids and derivatives from a consistent set of human PXR activation data, suggesting that hydrophobic interactions are essential for high PXR activity, consistent with X-ray structures and also with important descriptors in the PXR models derived in our work.

In particular the approach of computational solvent mapping was used towards the identification and characterization of protein binding site regions, which significantly contribute to free energy of binding. The results reported by Ngan et al.[89] indicated four of those hot spot regions at four different regions of the nearly spherical protein binding site, while a fifth interaction region is located close to its center. Three of these regions are already present in the apo-protein structure with the most important interaction region defined by the hydrophobic subpocket lined by Trp299, Phe288 and Tyr306. This site appears to interact with many known PXR ligands by hydrophobic and aromat–aromat interactions, which is in accord to our finding of a marked influence of hydrophobicity for PXR activator classification. Depending on their size and shape, individual PXR ligands might extend into 2, 3 or 4 more of these major interaction regions.[89]

Other QSAR approaches towards deriving classification models for larger datasets employed multiple methods and datasets,[31,33] although there have only been few attempts to derive ligand-based QSAR models around a large, structurally narrow set of PXR activators. The absence of large datasets, except for some recent comparative report by Chen et al.[96] on non-disclosed data, typically restricts QSAR approaches to a small fraction of the relevant chemical space.[97] Various PXR classification models employed a multitude of machine learning approaches (random forest, recursive partitioning, SVM). Some promising models were reported, which correctly predict 63–67% of an external test set.[33] A second study from the same group reported a test set prediction accuracy of 72–81% using SVM and different descriptors,[95] while these models do not provide a chemical interpretation. Other models have recently been derived using larger datasets for steroids and derivatives tested in the same assay system, including a predictive Bayesian classification model using 2D fingerprints and interpretable descriptors.[93] Interestingly the authors also conclude that ligand-based-QSAR methods in this case outperformed docking in PXR classification approaches.[97] All PXR positive contributing substructures from this model were reported to be essentially hydrophobic, while PXR negative contributing features contained hydroxyl-moieties or other functional groups, which might not be optimally positioned for hydrogen-bonding interactions in this site. This Bayesian approach was successfully extended to a larger and diverse training set of 177 molecules. The application of this classifier to a subset of FDA approved pharmaceuticals confirmed 9 drugs as novel PXR activators from a predicted set of 17 molecules followed by experimental testing.[98]

The PXR datasets used in our study are among the largest used to our knowledge today for ligand-based QSAR modeling; they are additionally augmented using internal data (dataset B) to focus on interesting regions in chemical space. This allows the application of either model derived from dataset A or B, depending on the potential similarity of novel prediction candidates to the employed training set. Furthermore, if there is a certain similarity to compounds from the training set of the more focused dataset C, it might be possible to obtain a more quantitative ranking and in-depth SAR insight into undesirable PXR antitarget activity in certain chemical series.

4. Conclusion

The development of decision and regression trees with a combination of relevant molecular descriptors resulted in useful antitarget in silico filters for the early identification of pregnane X receptor activators in drug discovery settings. This nuclear hormone receptor is a typical antitarget regulating the expression of several enzymes and transporters in metabolically relevant processes with CYP3A4 as most prominent enzyme induced. The applied statistical approaches provided an efficient way towards the development of meaningful QSAR models for a relatively large set of compounds without any obvious alignment rule.

Two classification models based on a diverse dataset of 434 drug-like molecules and a augmented set, with additional internal compounds in order to further investigate some regions in chemical space, were able to successfully classify novel molecules with respect to their PXR activation potential. These classifiers are based on decision trees combined with a genetic algorithm based variable selection to arrive at predictive models. Finally a predictive quantitative model for PXR activation for a subset of these molecules was derived using a regression-tree approach combined with GA variable selection.

As success rates for in silico identification of non-PXR activators for these approaches are consistently slightly lower compared to those for PXR activators, experimental testing is proposed for those compounds, which pass this filter (e.g., predicted to be non-activators of PXR), but show favorable experimental activity on the desirable biological target. Collectively these tools allow for an in silico prioritization for in vitro testing. This suggested workflow is therefore in accordance with current drug discovery scenarios. The combination of these filters consistently provide a tool identifying compounds a potential liability early in drug discovery and they also offer guidelines for lowering PXR activation in novel candidate molecules.

Acknowledgments

The authors would like to thank A. Dudda, M. Müller, R. Vaz, S. Reiling and S. Keil (all Sanofi) for many discussions.

Supplementary data

An SDF file with dataset A plus literature references, activity class and training/test/external set assignments. Three tables summarizing descriptors for model building and physicochemical meaning. A detailed description of obtained trees for classification and regression tree models associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bmc.2012.04.020.

References and notes

1. Kliewer, S. A.; Moore, J. T.; Wade, L.; Staudinger, J. L.; Watson, M. A.; Jones, S. A.; McKee, D. D.; Oliver, B. B.; Willson, T. M.; Zetterstrom, R. H.; Perlmann, T.; Lehmann, J. M. *Cell* **1998**, *92*, 73.
2. Lehmann, J. M.; McKee, D. D.; Watson, M. A.; Willson, T. M.; Moore, J. T.; Kliewer, S. A. *J. Clin. Invest.* **1998**, *102*, 1016.
3. Blumberg, B.; Sabbagh, W., Jr.; Juguilon, H.; Bolado, J.; van Meter, C. M.; Ono, E. S.; Evans, R. M. *Genes Dev.* **1998**, *12*, 3195.
4. Moore, L. B.; Maglich, J. M.; McKee, D. D.; Wisely, B.; Willson, T. M.; Kliewer, S. A.; Lambert, M. H.; Moore, J. T. *Mol. Endocrinol.* **2002**, *16*, 977.
5. Xie, W.; Barwick, J. L.; Simon, C. M.; Pierce, A. M.; Safe, S.; Blumberg, B.; Guzelian, P. S.; Evans, R. M. *Genes Dev.* **2000**, *14*, 3014.
6. Bertilsson, G.; Heidrich, J.; Svensson, K.; Asman, M.; Jendeberg, L.; Sydow-Backman, M.; Ohlsson, R.; Postlind, H.; Blomquist, P.; Berkenstam, A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12208.
7. Kliever, S. A.; Goodwin, B.; Willson, T. M. *Endocr. Rev.* **2002**, *23*, 687.
8. Mackenzie, P. I.; Gregory, P. A.; Gardner-Stephen, D. A.; Lewinsky, R. H.; Jorgensen, B. R.; Nishiyama, T.; Xie, W.; Radominska-Pandya, A. *Curr. Drug Metab.* **2003**, *4*, 249.
9. Falkner, K. C.; Pinaire, J. A.; Xiao, G. H.; Geoghegan, T. E.; Prough, R. A. *Mol. Pharmacol.* **2001**, *60*, 611.
10. Rosenfeld, J. M.; Vargas, R., Jr.; Xie, W.; Evans, R. M. *Mol. Endocrinol.* **2003**, *17*, 1268.
11. Ekins, S.; Erickson, J. A. *Drug Metab. Dispos.* **2002**, *30*, 96.
12. Xie, W.; Uppal, H.; Saini, S. P.; Mu, Y.; Little, J. M.; Radominska-Pandya, A.; Zemaitis, M. A. *Drug Discovery Today* **2004**, *9*, 442.
13. Ma, X.; Idle, J. R.; Gonzalez, F. J. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 895.
14. Timsit, Y. E.; Negishi, M. *Steroids* **2007**, *72*, 231.
15. Di Masi, A.; De Marinis, E.; Ascenzi, P.; Marino, M. *Mol. Aspects Med.* **2009**, *30*, 297.
16. *Nuclear Receptors as Drug Targets*; Ottow, E., Weinmann, H., Eds.; Wiley-VCH: Weinheim, 2008.
17. Guzelian, J.; Barwick, J. L.; Hunter, L.; Phang, T. L.; Quattrochi, L. C.; Guzelian, P. S. *Toxicol. Sci.* **2006**, *94*, 379.
18. Fuhr, U. *Clin. Pharmacokinet.* **2000**, *38*, 493.
19. Dixit, S. G.; Tirona, R. G.; Kim, R. B. *Curr. Drug Metab.* **2005**, *6*, 385.
20. Chang, T. K.; Waxman, D. J. *Drug Metab. Rev.* **2006**, *38*, 51.
21. Moreau, A.; Vilarem, M. J.; Maurel, R.; Pascussi, J. M. *Mol. Pharm.* **2008**, *5*, 35.
22. Francis, G. A.; Fayard, E.; Picard, F.; Auwerx, J. *Annu. Rev. Physiol.* **2003**, *65*, 261.
23. Markov, G.; Bonneton, F.; Laudet, V. In *Nuclear Receptors*; Bunce, C. M., Campbell, M. J., Eds.; Springer: Dordrecht, 2010; p 15.
24. Krasowski, M. D.; Yasuda, K.; Hagey, L. R.; Schuetz, E. G. *Mol. Endrocrinol.* **2005**, *19*, 1720.
25. Reschly, E. J.; Bainy, A. C. D.; Mattos, J. J.; Hagey, L. R.; Bahary, N.; Mada, S. R.; Ou, J.; Venkataramanan, R.; Krasowski, M. D. *BMC Evol. Biol.* **2007**, *7*, 222.
26. Honkakoshi, P.; Sueyoshi, T.; Negishi, M. *Ann. Med.* **2003**, *35*, 172.
27. Staudinger, J. L.; Ding, X.; Lichti, K. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 847.
28. Guengerich, F. P. *Ann. Rev. Pharmacol. Toxicol.* **1999**, *39*, 1.
29. De Groot, M. J. *Drug Discovery Today* **2006**, *11*, 601.
30. Byvatov, E.; Baringhaus, K.-H.; Schneider, G.; Matter, H. *QSAR Comb. Sci.* **2007**, *26*, 618.
31. Ung, C. Y.; Li, H.; Yap, C. W.; Chen, Y. Z. *Mol. Pharmacol.* **2007**, *71*, 158.
32. pEC_{50} computed from $-\log(EC_{50})$, with EC_{50} in [M].
33. Khandelwal, A.; Krasowski, M. D.; Reschly, E. J.; Sinz, M. W.; Swaan, P. W.; Ekins, S. *Chem. Res. Toxicol.* **2008**, *21*, 1457.
34. Ekins, S.; Chang, M.; Mani, S.; Krasowski, M. D.; Reschly, E. J.; Iyer, M.; Kholodovych, V.; Ai, N.; Welch, W. J.; Sinz, M.; Swaan, P. W.; Patel, R.; Bachmann, K. *Mol. Pharmacol.* **2007**, *72*, 592.
35. Sinz, M.; Kim, S.; Zhu, Z.; Chen, T.; Anthony, M.; Dickinson, K.; Rodrigues, A. D. *Curr. Drug Metab.* **2006**, *7*, 375.
36. Lemaire, G.; Benod, C.; Nahoum, V.; Pillon, A.; Boussioux, A.-M.; Guichou, J.-F.; Subra, G.; Pascussi, J.-M.; Bourguet, W.; Chavanieu, A.; Balaguer, P. *Mol. Pharmacol.* **2007**, *72*, 572.
37. Lemaire, G.; Mnif, W.; Pascussi, J. M.; Pillon, A.; Rabenoelina, F.; Fenet, H.; Gomez, E.; Casellas, C.; Nicolas, J. C.; Cavailles, V.; Duchesne, M. J.; Balaguer, P. *Toxicol. Sci.* **2006**, *91*, 501.
38. Xue, Y.; Chao, E.; Zuercher, W. J.; Willson, T. M.; Collins, J. L.; Redinbo, M. R. *Bioorg. Med. Chem.* **2007**, *15*, 2156.
39. Feng, D.-M.; DiPardo, R. M.; Wai, J. M.; Chang, R. K.; Di Marco, C. N.; Murphy, K. L.; Ransom, R. W.; Reiss, D. R.; Tang, C.; Prueksaritanont, T.; Pettibone, D. J.; Bock, M. G.; Kuduk, S. D. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 682.
40. Gao, Y.-D.; Olson, S. H.; Balkovec, J. M.; Zhu, Y.; Royo, I.; Yabut, J.; Evers, R.; Tan, E. Y.; Tang, W.; Hartley, D. P.; Mosley, R. T. *Xenobiotica* **2007**, *37*, 124.
41. Fotsch, C.; Bartberger, M. D.; Bercot, E. A.; Chen, M.; Cupples, R.; Emery, M.; Fretland, J.; Guram, A.; Hale, C.; Han, N.; Hickman, D.; Hungate, R. W.; Hayashi, M.; Komorowski, R.; Liu, Q.; Matsumoto, G.; Jean, D. J.; Ursu, S.; Véniant, M.; Xu, G.; Ye, Q.; Yuan, C.; Zhang, J.; Zhang, X.; Tu, H.; Wang, W. *J. Med. Chem.* **2008**, *51*, 7953.
42. AurSCOPE database; Available from Aureus Sciences, Paris, France.
43. *SYBYL (version 8.1)*; Available from Tripos Inc.: St. Louis, MO, USA.
44. Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644.
45. Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912.
46. Sushko, I.; Novotarskyi, S.; Koerner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Mueller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Oeberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. *J. Chem. Inf. Model.* **2010**, *50*, 2094.
47. Moore, L. B.; Parks, D. J.; Jones, S. A.; Bledsoe, R. K.; Consler, T. G.; Stimmel, J. B.; Goodwin, B.; Liddle, C.; Blanchard, S. G.; Willson, T. M.; Collins, J. L.; Kliewer, S. A. *J. Biol. Chem.* **2000**, *275*, 15122.
48. Sadowski, J.; Rudolph, C.; Gasteiger, J. *Anal. Chim. Acta* **1992**, *265*, 233.
49. *CORINA (version 3.4)*; Available from Molecular Networks Inc.: Erlangen, Germany.
50. *MOE (version 2009.10)*; Available from Chemical Computing Group (CCG), Montreal, Canada.
51. Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. *Theochem* **2000**, *503*, 17.
52. Cruciani, G.; Pastor, M.; Clementi, S. In *Molecular Modelling and Prediction of Bioactivity, Proceedings of the 12th European Symposium on Quantitative Structure–Activity Relationships (QSAR'98)*; Gundertofte, K., Jorgensen, F. S., Eds.; Plenum Press: New York, 2000; p 73.
53. Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. *J. Med. Chem.* **2000**, *43*, 2204.
54. *Volsurf+ (version 1.0.4.)*; Available from Molecular Discovery Ltd: Pinner, Middlesex, UK.
55. Kramer, C.; Beck, B.; Clark, T. *J. Chem. Inf. Model.* **2010**, *50*, 429.
56. *Parasurf (version 10)*; Available from Cepos InSilico Ltd: Kempston, Bedford, UK.
57. Schneider, G.; Neidhart, W.; Giller, T.; Schmidt, G. *Angew. Chem., Int. Ed.* **1999**, *111*, 3068.
58. Matter, H.; Giegerich, C. *Internal Implementation*; 2007.
59. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
60. *Mopac (version 6)*; Available from Cepos InSilico Ltd: Kempston, Bedford MK42 8BQ, UK.
61. Ehresmann, B.; de Groot, M. J.; Clark, T. *J. Chem. Inf. Model.* **2005**, *45*, 1053.
62. Heiden, W.; Goetze, T.; Brickmann, J. *J. Comput. Chem.* **1993**, *14*, 246.
63. Meyer, A. Y. *Chem. Soc. Rev.* **1985**, *15*, 449.
64. Clark, T.; Byler, K. G.; de Groot. M. J. In *Proceedings of the International Beilstein Workshop*; Bozen, Italy, 2006, Logos, Berlin, 2008, p 129.

65. Dillon, W. R.; Goldstein, M. *Multivariate Analysis: Methods and Applications*; Wiley: New York, 1984.
66. Cramer, R. D., III *J. Am. Chem. Soc.* **1980**, *102*, 1837.
67. Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johanson, E.; Lindberg, W.; Sjöström, M. In *Chemometrics: Mathematics and Statistics in Chemistry*, Kowalski, B. R. (Ed.); NATO, ISI Series C 138, D. Reidel Publ. Co.: Dordrecht, Holland, 1984; p 17.
68. Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. *J. Chemometrics* **1994**, *8*, 111.
69. Quinlan, J. R. In *Proceedings ML'93*, Utgoff, P. E., Ed.; Morgan Kaufmann, Los Altos, CA, 1993.
70. *C5.0 (version 2.05)*; Available from RuleQuest Research Pty Ltd: St Ives NSW, Australia.
71. Gupta, R. R.; Gifford, E. M.; Liston, T.; Waller, C. L.; Hohman, M.; Bunin, B. A.; Ekins, S. *Drug Metab. Dispos.* **2010**, *38*, 2083.
72. Qumsieh, A., 2005. Documentation and download: http://search.cpan.org/~aqumsieh/AI-Genetic-0.05/Genetic.pm (accessed January 2012).
73. Quinlan, J. R. *Mach. Learn.* **1991**, *6*, 93.
74. Quinlan, J. R. In *Proc. AI'92, 5th Australian Joint Conference on Artificial Intelligence*, Adams, A., Sterling, L., Eds.; World Scientific: Singapore, 1992, p 343.
75. Butina, D.; Gola, J. M. R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837.
76. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.
77. GCUT descriptors are derived by graph theory, they are calculated from the eigenvalues of a modified graph distance adjacency matrix. The diagonal takes the value of the PEOE partial charges, as implemented in MOE. The resulting eigenvalues are sorted and the eigenvalues are reported.
78. BCUT_SLOPGP3 refers to BCUT values from eigenvalues of a modified adjacency matrix using atomic contributions to log*P*.
79. Q_VSA_NEG refers to the fractional negative vdW surface area as sum of atomic surface areas with a negative partial charge divided by the total surface area.
80. Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem.* **1984**, *19*, 66.
81. Hall, L. H.; Kier, L. B. *Rev. Comput. Chem.* **1991**, *2*, 367.
82. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3.
83. Watkins, R. E.; Wisely, G. B.; Moore, L. B.; Collins, J. L.; Lambert, M. H.; Williams, S. P.; Willson, T. M.; Kliewer, S. A.; Redinbo, M. R. *Science* **2001**, *292*, 2329.
84. Jacobs, M. N.; Dickens, M.; Lewis, D. F. *J. Steroid Biochem. Mol. Biol.* **2003**, *84*, 117.
85. Kliewer, S. A.; Goodwin, B.; Willson, T. M. *Endocr. Rev.* **2002**, *23*, 687.
86. Watkins, R. E.; Maglich, J. M.; Moore, L. B.; Wisely, G. B.; Noble, S. M.; Davis-Searles, P. R.; Lambert, M. H.; Kliewer, S. A.; Redinbo, M. R. *Biochemistry* **2003**, *42*, 1430.
87. Chrencik, J. E.; Orans, J.; Moore, L.; Xue, Y.; Peng, L.; Collins, J. L.; Wisely, G. B.; Lambert, M. H.; Kliewer, S. A.; Redinbo, M. R. *Mol. Endocrinol.* **2005**, *19*, 1125.
88. Teotico, D. G.; Bischof, J. J.; Peng, L.; Kliewer, S. A.; Redinbo, M. R. *Mol. Pharmacol.* **2008**, *74*, 1512.
89. Ngan, C.-H.; Beglov, D.; Rudnitskaya, A. N.; Kozakov, D.; Waxman, D. J.; Vajda, S. D. *Biochemistry* **2009**, *48*, 11572.
90. Ekins, S.; Erickson, J. A. *Drug Metab. Dispos.* **2002**, *30*, 96.
91. Schuster, D.; Langer, T. *J. Chem. Inf. Model.* **2005**, *45*, 431.
92. Ekins, E.; Andreyev, S.; Ryabov, A.; Kirillov, E.; Rakhmatulin, E. A.; Sorokina, S.; Bugrim, A.; Nikolskaya, T. *Drug Metab. Dispos.* **2006**, *34*, 495.
93. Ekins, S.; Kortagere, S.; Iyer, M.; Reschly, E. J.; Lill, M. A.; Redinbo, M. R.; Krasowski, M. D. *PLOS Comput. Biol.* **2009**, *5*, e1000594.
94. Wang, C. Y.; Li, C. W.; Chen, J. D.; Welch, W. J. *Mol. Pharmacol.* **2006**, *69*, 1513.
95. Kortagere, S.; Chekmarev, D.; Welch, W. J.; Ekins, S. *Pharm. Res.* **2009**, *26*, 1001.
96. Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. *J. Chem. Inf. Model.* **2012**, *52*, 792.
97. Ekins, S.; Kortagere, S.; Krasowski, M. D.; Williams, A. J.; Xu, J. J.; Zientek, M. In *RSC Drug Discovery Series: Drug Design Strategies: Quantitative Approaches*, Livingstone, D. J., Davis, A. M., Eds., RSC Publishing: Cambridge, UK, 2012; Vol. 13, pp 312.
98. Pan, Y.; Li, L.; Kim, G.; Ekins, S.; Wang, H.; Swaan, P. W. *Drug Metab. Dispos.* **2011**, *39*, 337.