LASSO-ing Potential Nuclear Receptor Agonists and Antagonists: A New Computational Method For Database Screening

Sean Ekins¹, Michael-R. Goldsmith²,³, Aniko Simon⁴, Zsolt Zsoldos⁴, Orr Ravitz⁴, and Antony J. Williams⁵

¹ Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526, U.S.A.
² National Exposure Research Laboratory, US-Environmental Protection Agency, 109 TW Alexander Dr., Research Triangle Park, NC 27711, U.S.A.
³ To whom correspondence should be addressed. (e-mail: goldsmith.rocky@epa.gov)
⁴ SimBioSys, Inc. 135 Queen’s Plate Drive, Suite 520, Toronto, Ontario M9W 6V1, Canada.
⁵ Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC 27587, U.S.A.

DISCLAIMER: This document has been subjected to review by the U.S. Environmental Protection Agency and Approved for publication.

Running Head: Models for Nuclear Receptor Agonists and Antagonists
Abstract

Nuclear receptors (NRs) are important biological macromolecular transcription factors that are implicated in multiple biological pathways and may interact with other xenobiotics that are endocrine disruptors present in the environment. Examples of important NRs include the androgen receptor (AR), estrogen receptors (ER) and the pregnane X receptor (PXR). In this study we have utilized the Ligand Activity by Surface Similarity Order (LASSO) method, a ligand-based virtual screening strategy to derive structural (surface/shape) molecular features used to generate predictive models of biomolecular activity for AR, ER and PXR. For PXR, twenty five models were built using between 8 to 128 agonists and tested using 3000, 8000 and 24,000 drug-like decoys including PXR inactive compounds (N=228). Preliminary studies with AR and ER using LASSO suggested the utility of this approach with 2 fold enrichment factors at 20%. We found that models with 64-128 PXR actives provided enrichment factors of 10 fold (10% actives in the top 1% of compounds screened). The LASSO models for AR and ER have been deployed and are freely available at www.chemspider.com and represent a ligand-based prediction method for putative NR activity of compounds in this database.

Keywords

Androgen receptor; Estrogen receptor; Ligand Activity by Surface Similarity Order; Ligand-based models; Pregnane X receptor
Introduction

The nuclear receptor (NRs) family of transcription factors are important targets for therapeutic interventions for multiple diseases [1] and also may interact with other xenobiotics that are endocrine disruptors present in the environment [2]. It is therefore important to identify compounds that may specifically bind NRs and act as endocrine disruptors as well as develop synthetic compounds that can selectively (in a cell-type and/or tissue-selective manner) modulate NR pharmacology (reviewed in [3-9]). NRs including the androgen receptor (AR; NR3C4), estrogen receptor α and β (ERα and ERβ; NR3A1 and NR3A2) and pregnane X receptor (PXR; NR1I2) are particularly important as both therapeutic targets and for xenobiotics to mediate off-target effects.

The ERs are activated by 17β-estradiol while the AR is activated by testosterone and dihydrotestosterone and these receptors are transcriptional regulators of many genes [10] with important physiological functions [11-16]. The human PXR [17-19] similarly transcriptionally regulates genes involved in xenobiotic metabolism and excretion, as well as other cellular processes, including apoptosis [20-24]. Human PXR is a broad specificity NR, binding a wide variety of molecules [25] and the activation of this NR can cause drug-drug interactions [23].

Multiple QSAR and machine learning models have been described for these NRs, to address endocrine disruptor risk assessment [26-28] and toxicological screening [29]. For example, a recent QSAR analysis of 74 natural or synthetic estrogens provided information on structural features for the activation of ERα and ERβ [30]. Non-linear statistical machine learning methods have been applied to separate NR activators from non-activators [31]. A virtual screening protocol identified ERβ specific ligands from a
plant product-based database [32], from 12 candidates evaluated by a fluorescence polarization binding assay, 3 had >100-fold selectivity to ERβ over ERα. The same approach has also been used to find compounds with good selectivity for ERα over ERβ [32, 33]. Bisson et al., have used computational methods that led to a non-steroidal antiandrogen with improved AR antagonistic activity based on an initial screening of FDA approved new drugs [33]. Several groups have published datasets or performed modeling on ER and AR and these data are readily available for further evaluation with new modeling methods [34-38].

While the crystal structures of human PXR [39-43] have led to a greater understanding of the ligand binding domain (LBD) and ligand-receptor interactions [39-45], ligand-based computational models possess the key features for predicting binding [46-49]. PXR pharmacophores, have been used to predict interactions for antibiotics [50] verified in vitro, and machine learning methods have also been evaluated [25, 38, 51-53]. Several protein-based docking studies have also been used to predict PXR agonists [25, 54-56], although machine learning methods appear to be advantageous to date.

We have recently described troubleshooting various computational methods [57] and specifically compared different methods for PXR [25]. There is a continuous search for new methods that might offer advantages for computational modeling to overcome some of these limitations and specifically for NRs [58]. A new ligand based software called LASSO (Ligand Activity in Surface Similarity Order) has recently been described that is focused on similarities in biomolecular activity rather than structural similarity [59]. The key components describing LASSO are the 23 kinds of Interacting Surface Point Type (ISPT) molecular descriptors (see Supplemental Table 1), which capture the
essence of the surface point information in a feature vector containing the counts of each surface point type, and creates the feature vector for that ligand. This vector serves as the descriptor of that molecule with the assumption that ligands with similar feature vectors will have similar activity. As only the 3D orientation of the Interacting Surface Point depends on the conformation, LASSO is conformation independent. LASSO has been shown to be able to readily screen over 1 million structures/minute, identify active molecules by enriching screened databases and provides a means for scaffold hopping [59]. The current study applies LASSO to various NR datasets to generate models, validate them and make the models available on a public website to illustrate how the method can be used. This work can be considered an extension of our previous troubleshooting studies [25, 57]

Materials and Methods

I) Training and Test Set Molecular Dataset Selection

One of the goals of this study was to determine what level of enrichment for binders (at weak or strong binding threshold) can be afforded using the ChemSpider LASSO descriptors (ligand-based approach) and compared with enrichment from a structure-based docking approach (eHiTS). For AR, the dataset consisting of 203 molecules with relative binding affinities and activity threshold classes of (a) strong (b) moderate (c) weak and (d) inactive/non-binding ligands, we evaluated the ability of LASSO to differentiate both (a) strong and (a+b) strong and moderate binders (all others considered to be non-binding). The training set for AR, derived with the LASSO descriptors were obtained from the DUD set [60] and differ considerably from the test
set. To evaluate the LASSO-descriptors for the ER dataset consisting of 50 molecules with 15 “hits” (i.e. considerably weak binders) and 35 “non-hits” for the estrogen binding that differ considerably from the training set obtained again we used the DUD ER specific (default or agonist and antagonist) as a training set [60].

In addition to the ChemSpider-LASSO approach for the AR test set we used the eHiTS structure-based (molecular docking) screening strategy on two conformations of the AR (using PDB structures 2AMA and 1XNN) and reported the minimum score across the two-conformations examined (this approach was used to add flexibility to the receptor). Similarly, for the ER dataset we docked against two functionally distinct conformations of the Estrogen receptor (3ERT and 1GWR).

Datasets for LASSO modeling: Structure File preparation

The rat ER Binding dataset (Kᵢ values for 50 compounds of environmental relevance [35]) was obtained from EPA’s DSSTox database (http://epa.gov/ncct/dsstox/ [34]). This dataset contains 15 industrial chemical “binders” (i.e. non-therapeutic) with significantly weaker binding affinities than would be desired for drug lead candidates (i.e. 3-5 fold weaker binding affinity than the natural ligand 17β-estradiol). Similarly, the NCTR’s rat AR activity dataset (competitive inhibition assays) also used in this study, contains 146 AR binders and 56 non-binders (http://www.fda.gov/nctr/science/centers/toxicoinformatics/edkb/index.htm [37, 38]). All structures were imported into MOE and geometry optimized using the MMFFx force-field in MOE (Chemical Computing Group, Montreal, Canada).

Three human PXR datasets were used, namely: Dataset 1 represented 80 actives EC₅₀<100 μM and 64 inactives EC₅₀>100 μM that were drug-like molecules. The
SMILES string for each molecule named or CAS number provided was obtained by downloading from either PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChemSpider (http://www.chemspider.com/) or sketched using the BUILDER module of SYBYL [56]. Dataset 2 represented 93 actives and 75 inactives that were drug-like molecules from a previous study [61]. The molecular structures encoded as SMILES strings [62] were downloaded from the supplementary information tables in the original publication [61]. Dataset 3 represented 30 actives and 89 inactives from a dataset of steroidal compounds, (namely androstanes, estratrienes, pregnanes and bile salts) as well as the ligands used in the crystal structures with hPXR activation determined by a luciferase-based reporter assay [25]. Human PXR activation was determined by a luciferase-based reporter assay as has been previously described in these and other publications.

**LASSO models for ER and AR**

The methodology of LASSO has already been previously described in detail, [59] and the method performance in terms of diversity of test set and % enrichment of a database has also already been evaluated for the DUD set in paper just mentioned and also for other targets published elsewhere. (http://www.simbiosys.com/ehits_lasso/ehits_lasso_table.html) To examine the performance of eHiTS LASSO, with this endocrine panel subset of target proteins of the total ~48 nuclear receptors. We used the newly assembled directory of useful decoys (DUD) [60] dataset to augment both the KIERBL and NCTR AR datasets.

**PXR models - method I**
The previously mentioned three PXR datasets were received from three different sources described earlier. Set 1, called: “hpxr_test”, contained 80 actives and 64 inactives or decoys; set 2, called: “hpxr_train”, contained 93 actives and 75 inactives, finally set 3, called: “PXR119-class”, contained 30 actives with 89 inactives. Out of these three data sets, 7 screening prediction models were built using only the actives (the inactives were automatically generated by the software).

The following models were developed: Model 1 was trained on the 1st dataset (hpxr_test, 80 ligands) and tested with the other two sets (123 ligands). Model 2 was trained on the 2nd dataset (hpxr_train, 93 ligands) and tested with the other two sets (110 ligands). Model 3 was trained on the 3rd dataset (PXR119-class, 30 ligands); tested with the other two sets (173 ligands). Model 4-6 were trained on sets 1 and 2, i.e. 173 ligands, sets 1 and 3 , i.e. 110 ligands, and sets 2 and 3 , i.e. 123 ligands); and tested with the remaining one set of actives (i.e. 30, 93 and 80 ligands respectively). Model 7 was trained on all actives (1, 2 and 3) and tested on the same. This was done as an extreme case, to see the maximum potential training effect.

**PXR models - method II**

A second method for creating LASSO prediction models for the PXR test case was also investigated. Actives from the 3 datasets were all merged, resulting in an SDF file with 203 ligands with relative binding affinities and activity threshold classes (a) strong, (b) moderate, (c) weak and (d) inactive/non-binding ligands. To determine how many actives are needed to be selected for a good LASSO prediction model and also to see if the source of the actives is important, 25 LASSO models were developed
(Supplemental Table 2). Prediction models were built by selecting 8, 16, 32, 64, and 128 actives, starting from positions: 1st, 9th, 17th, 33rd and 64th in the merged actives file.

The above 25 models were then tested for enrichment factor using the actives from the total active set and leaving out the ones used for training (this was 8, 16, 32, 64 or 128 ligands respectively) mixed with drug-like decoys, that were obtained from another recent screening study [63]. To assess the effect of the size of the decoy set upon the prediction model, random 3000 (3k), 8000 (8k) and the whole 24,000 (24k) decoy sets were used. In each case the decoys from all three sets received (228 structures in total) were added into the decoy test set.

**Results**

The enrichment plots shown for AR (Figure 1) and ER (Figure 2) with the percent actives recovered versus percent of dataset reveals an enrichment of ~2 fold at 20% of the dataset coverage regardless of whether a ligand or structure based approach was used. For the AR dataset if the interaction threshold is specified as strong, or strong+moderate different levels of enrichment are incurred by either ChemSpider-LASSO or eHiTS results. This translates into an improved performance of either ligand or structure-based screening approaches to bin molecules with stronger interactions (cyan and purple) than those substantially altered through the addition of weaker binding classes (magenta and yellow). Interestingly, in terms of the early-recognition problem, eHiTS is more sensitive (4 fold at 20%) than LASSO (1.5 fold at 20%), however, all 15 actives are captured by the LASSO descriptor within the first 37% of the dataset (with a minimum value of
LASSO=0.07) at considerably lower computational cost. A means of incorporating this into a real scenario would be to screen ChemSpider for AR with a descriptor above a threshold (in this case 0.07) from a specific dataset on ChemSpider and follow up these “hits” only with a more costly structure-based approach.

For the ER dataset where all 15 binders are in fact weak binders (i.e. 3 to 5 orders of magnitude weaker binders than the natural ligand 17β-estradiol) the default LASSO descriptors outperform (3 fold enrichment) the structure-based approach (2 fold enrichment) and the agonist-trained LASSO method outperforms the antagonist LASSO method (most likely due to a large diversity among antagonists than agonists). Here we can see that even for weakly interacting partners (i.e. low affinity binders for ER) we can still obtain enrichment that is substantially better than random.

These tandem virtual screening approaches combine computationally-efficient ligand-based ChemSpider LASSO descriptors\(^1\) prior to more costly structure-based virtual screening strategies, dramatically improving virtual screening and “early-recognition problem” workflow efficiency.

Piggy-backing more costly structure-based virtual screening strategies on top of an initial screen with could dramatically assist in virtual screening endeavors and the early recognition problem.

\(^1\) Since ChemSpider is at its core a rich and diverse collection of chemical structures, these were used in order to produce LASSO predictions for over 14 million compounds against a series of 40 targets including AR and ER. A LASSO search feature was added to ChemSpider to allow users to search the database by LASSO value (see Figure 3A). Scientists can readily search for the top 1000 compounds (or less) with the highest LASSO value for a particular target of interest. An advanced search in ChemSpider can combine LASSO value searches with other parameters such as molecular weight, Rule of 5 values and specific data sources (for example, selecting molecules from commercial data sources only).
We have also shown an example of a molecule, mibolerone, a strong AR and ER binder based on LASSO (Figure 3B) which is known as a potent AR binder [64]. The LASSO surface point type values are shown in Supplemental Table 3 and more visually in Supplemental Figure 2.

When we used LASSO models with hPXR in method I (Figure 4) we found the best results with Model 1 which suggested 40% of the ligands can be pushed into the top 10% of the screened database resulting with an enrichment factor 4 fold better than random (Figure 5). In Method II we found the same enrichment factor using 64 actives in a 24 thousand compound decoy set (Figure 6). Another way to evaluate the models is to present the statistics for using dataset 1 to predict dataset 2 (N = 168) for which we obtained Sensitivity 12%, Specificity 99%, Accuracy 51%, Matthew’s correlation 0.2. Using dataset 2 to predict dataset 1 gives similar results. These results suggest that the models could identify potential human PXR agonists in databases similar to other target proteins [59].

Discussion

For both the AR and ER ligands the main objective was to see how ligand-based screening tools, such as LASSO’s ChemSpider implementation, perform such that they could be used for prioritizing chemicals for testing. The AR dataset contained a mixture of drug-like and environmental receptor modulators, whereas the ER dataset contained primarily environmental chemicals. Even in light of the relatively weak binding affinity of the “actives”, i.e. Kᵢ of 10⁻⁴-10⁻⁶, while these would be poor candidates for lead optimization into drugs, they still pose an interaction potential with biological systems.
such as NRs if they bioaccumulate. Using these leads from LASSO screening with other methods such as molecular docking or free-energy perturbation simulations may also be useful. The validation of the approaches outlined above was pursued by examining two real datasets. These were the FDA’s NCTR AR [38] and a 50 chemical dataset containing environmental molecules evaluated for ER binding affinity [35]. In addition we have used multiple sets of PXR agonists described previously. Our results show enrichments of between 2 fold and 4 fold depending on the NR. For PXR there have been numerous recent studies using different machine learning methods and descriptors [25, 54, 56], and while the Matthews correlation coefficient in this study is lower than previous studies, the level of enrichment from between 4 fold (40% of actives in top 10% of compounds screened) to 10 fold (10% actives in the top 1% of compounds screened) was very encouraging.

The molecular descriptors used in eHiTS LASSO are independent of ligand conformation and have been shown to successfully enrich screened databases across a wide range of target families [59]. Lying somewhere between a 2D and a 3D descriptor the ISPT descriptor does not contain any shape or 2D connectivity information. There may however be some molecular size information implicit in the descriptor due to capturing the counts of surface points and larger molecules will have more surface points than smaller molecules (and eHiTS LASSO may be somewhat sensitive to this).

The relatively high speed of eHiTS LASSO on a single CPU [59], makes it an ideal tool to be used as a pre-docking screen. From a troubleshooting perspective, eHiTS LASSO will return a high percentage of false positives, due to not considering 3D relationships of surface properties. Because of this, it will also return a higher percentage
of different scaffolds, enabling scaffold hopping. It is also important to note that LASSO would not be able to differentiate stereoisomerism apart from perhaps diastereomeric pairs which have structurally (configurationally) different features rather than conformationally different features, for which this method is conformation invariant.

Taking the results of eHiTS LASSO and feeding the top N% into a docking program would allow the docking program to weed out many of the false positives binders. For this reason, eHiTS LASSO is currently available integrated with the commercially available eHiTS docking tool, and can be readily used as a pre-docking screening tool for large virtual screens.

From the current study we have shown significant enrichments when testing computational models for AR, ER and PXR. While AR and ER predictions are currently already implemented in ChemSpider, it is clear that adding predicted values for PXR and other NRs as they become available would be beneficial to the community in terms of accessing an open source of chemical structures with pre-generated descriptors. It should be noted however that the generation of model data for a database as large as that hosted by ChemSpider (now well over 25 million compounds) is not a small undertaking and consumes a significant amount of compute time, data preparation and handling in order to deliver the models to the community for consumption.

The use of such ligand-based computational methods as exemplified by LASSO in this study could also be useful for the design and selection of chemical products that are less hazardous to human health and the environment. This may make them useful in green chemistry [65] (http://www.epa.gov/gcc/pubs/about_gc.html) as well as in biomedical research. The ready accessibility of such NR binding predictions from
computational models like LASSO will be key in future for both pharmaceutical and environmental applications and databases like ChemSpider can have an important role in providing them to the public as a pre-docking criteria, as we have demonstrated in this study. This study used published rat ER, AR as well as human PXR data. LASSO could also be applied to build models for the same NRs across multiple species, such that they could be used to estimate interspecies variation in ligand binding.

Conflicts of Interest
AJW is employed by the Royal Society of Chemistry which owns ChemSpider and associated technologies. SE and MRG was on the advisory board for ChemSpider from June 2007 until May 2011. AS, ZZ and OR are employed by SimBioSys, Inc which owns LASSO and eHiTs.

Supporting Information Available: Supplemental material is available online. All endocrine disruptor datasets are available upon request from the authors.

ABBREVIATIONS
AR, Androgen receptor; DUD, Directory of useful decoys; ER, Estrogen receptor; ISPT, Interacting Surface Point Type; LASSO, Ligand Activity by Surface Similarity Order; LBD, Ligand binding domain; PXR, Pregnane X receptor; QSAR, Quantitative Structure Activity Relationship; SNNS, Stuttgart Neural Network Simulator.
References

1. Smith, A.G. and G.E. Muscat, *Skeletal muscle and nuclear hormone receptors: implications for cardiovascular and metabolic disease*. Int J Biochem Cell Biol, 2005. 37(10): p. 2047-63.

2. Hotchkiss, A.K., et al., *Fifteen Years after "Wingspread"--Environmental Endocrine Disrupters and Human and Wildlife Health: Where We are Today and Where We Need to Go*. Toxicol. Sci., 2008. 105(2): p. 235-259.

3. Tenbaum, S. and A. Baniahmad, *Nuclear receptors: Structure, function and involvement in disease*. The International Journal of Biochemistry & Cell Biology, 1997. 29(12): p. 1325-1341.

4. Zimber, A. and C. Gespach, *Bile acids and derivatives, their nuclear receptors FXR, PXR and ligands: role in health and disease and their therapeutic potential*. Anticancer Agents Med Chem, 2008. 8(5): p. 540-63.

5. Feldman, P.L., M.H. Lambert, and B.R. Henke, *PPAR modulators and PPAR pan agonists for metabolic diseases: the next generation of drugs targeting peroxisome proliferator-activated receptors?* Curr Top Med Chem, 2008. 8(9): p. 728-49.
6. Ma, X., J.R. Idle, and F.J. Gonzalez, *The pregnane X receptor: from bench to bedside*. Expert Opinion on Drug Metabolism & Toxicology, 2008. 4(7): p. 895-908.

7. Bertolotti, M., et al., *Nuclear receptors as potential molecular targets in cholesterol accumulation conditions: insights from evidence on hepatic cholesterol degradation and gallstone disease in humans*. Curr Med Chem, 2008. 15(22): p. 2271-84.

8. Baulieu, E.E., et al., *Steroid hormone receptors*. Vitam Horm, 1975. 33: p. 649-736.

9. Henley, D.V. and K.S. Korach, *Endocrine-Disrupting Chemicals Use Distinct Mechanisms of Action to Modulate Endocrine System Function*. Endocrinology, 2006. 147(6): p. s25-32.

10. Krasowski, M.D., E.J. Reschly, and S. Ekins, *Intrinsic disorder in nuclear hormone receptors*. J Proteome Res, 2008. 7(10): p. 4359-72.

11. Pike, A.C., et al., *Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist*. Embo J, 1999. 18(17): p. 4608-18.

12. Brzozowski, A.M., et al., *Molecular basis of agonism and antagonism in the oestrogen receptor*. Nature, 1997. 389(6652): p. 753-8.
13. Schwabe, J.W., et al., *DNA recognition by the oestrogen receptor: from solution to the crystal.* Structure, 1993. 1(3): p. 187-204.

14. Sack, J.S., et al., *Crystallographic structures of the ligand-binding domains of the androgen receptor and its T877A mutant complexed with the natural agonist dihydrotestosterone.* Proc Natl Acad Sci U S A, 2001. 98(9): p. 4904-9.

15. Bohl, C.E., et al., *A ligand-based approach to identify quantitative structure-activity relationships for the androgen receptor.* Journal of Medicinal Chemistry, 2004. 47(15): p. 3765-76.

16. Shaffer, P.L., et al., *Structural basis of androgen receptor binding to selective androgen response elements.* Proc Natl Acad Sci U S A, 2004. 101(14): p. 4758-63.

17. Bertilsson, G., et al., *Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction.* Proc Natl Acad Sci U S A, 1998. 95(21): p. 12208-13.

18. Blumberg, B., et al., *SXR, a novel steroid and xenobiotic-sensing nuclear receptor.* Genes Dev, 1998. 12(20): p. 3195-205.

19. Kliewer, S.A., et al., *An orphan nuclear receptor activated by pregnanes defines a novel steroid signalling pathway.* Cell, 1998. 92: p. 73-82.
20. Verma, S., M.M. Tabb, and B. Blumberg, *Activation of the steroid and xenobiotic receptor, SXR, induces apoptosis in breast cancer cells*. BMC Cancer, 2009. 9: p. 3.

21. Gupta, D., et al., *Expanding the roles for pregnane X receptor in cancer: proliferation and drug resistance in ovarian cancer*. Clin Cancer Res, 2008. 14: p. 5332-5340.

22. Zhou, J., et al., *The antiapoptotic role of pregnane X receptor in human colon cancer cells*. Mol Endocrinol, 2008. 22: p. 868-880.

23. Biswas, A., et al., *Elucidating the ‘Jekyll and Hyde’ nature of PXR: The case for discovering antagonists*. Pharm Res, 2009. 26: p. 1807-1815.

24. Urquhart, B.L., R.G. Tirona, and R.B. Kim, *Nuclear receptors and the regulation of drug-metabolizing enzymes and drug transporters: implications for interindividual variability in response to drugs*. J Clin Pharmacol, 2007. 47: p. 566-578.

25. Ekins, S., et al., *Challenges Predicting Ligand-Receptor Interactions of Promiscuous Proteins: The Nuclear Receptor PXR*. PLoS Comput Biol, 2009. 5: p. e1000594.

26. Ai, N., et al., *Computational Models for Predicting the Binding Affinities of Ligands for the Wild-Type Androgen Receptor and a Mutated Variant Associated with Human Prostate Cancer*. Chem. Res. Toxicol., 2003. 16(12): p. 1652-1660.
27. Jacobs, M.N., *In silico tools to aid risk assessment of endocrine disrupting chemicals*. Toxicology, 2004. **205**(1-2): p. 43-53.

28. Tong, W., et al., *QSAR Models for Binding of Estrogenic Compounds to Estrogen Receptor \{alpha\} and \{beta\} Subtypes*. Endocrinology, 1997. **138**(9): p. 4022-4025.

29. Ekins, S., L. Mirny, and E.G. Schuetz, *A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXRa, and LXRb*. Pharmaceutical Research, 2002. **19**: p. 1788 - 1800.

30. Zhu, B.T., et al., *Quantitative Structure-Activity Relationship of Various Endogenous Estrogen Metabolites for Human Estrogen Receptor \{alpha\} and \{beta\} Subtypes: Insights into the Structural Determinants Favoring a Differential Subtype Binding*. Endocrinology, 2006. **147**(9): p. 4132-4150.

31. Plewczynski, D., et al., *Target specific compound identification using a support vector machine*. Comb Chem High Throughput Screen, 2007. **10**(3): p. 189-96.

32. Zhao, L. and R.D. Brinton, *Structure-based virtual screening for plant-based ERbeta-selective ligands as potential preventative therapy against age-related neurodegenerative diseases*. J Med Chem, 2005. **48**(10): p. 3463-6.

33. Bisson, W.H., et al., *Discovery of antiandrogen activity of nonsteroidal scaffolds of marketed drugs*. Proceedings of the National Academy of Sciences, 2007. **104**(29): p. 11927-11932.
34. Laws, S., et al. *DSSTox EPA Estrogen Receptor Ki Binding Study (Laws et al.) Database - (KIERBL): SDF file and documentation*. 2009 [cited Launch version: KIERBL_v1a_278_17Feb2009,]; Available from: www.epa.gov/ncct/dsstox/sdf_kierbl.html

35. Laws, S.C., et al., *Nature of the binding interaction for 50 structurally diverse chemicals with rat estrogen receptors*. Toxicol Sci, 2006. 94(1): p. 46-56.

36. Rabinowitz, J.R., et al., *Molecular Modeling for Screening Environmental Chemicals for Estrogenicity: Use of the Toxicant-Target Approach*. Chem Res Toxicol, 2009. 22: p. 1594-1602.

37. Fang, H., et al., *Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor*. Chem Res Toxicol, 2003. 16(10): p. 1338-58.

38. Ding, D., et al., *The EDKB: an established knowledge base for endocrine disrupting chemicals*. BMC Bioinformatics, 2010. 11 Suppl 6: p. S5.

39. Watkins, R.E., et al., *Coactivator binding promotes the specific interaction between ligand and the pregnane X receptor*. J Mol Biol, 2003. 331: p. 815-828.

40. Watkins, R.E., et al., *2.1A crystal structure of human PXR in complex with the St John's Wort compound hyperforin*. Biochemistry, 2003. 42: p. 1430-1438.

41. Xue, Y., et al., *Crystal structure of the PXR-T1317 complex provides a scaffold to examine the potential for receptor antagonism*. Bioorg Med Chem, 2007. 15(5): p. 2156-66.
42. Chrencik, J.E., et al., Structural disorder in the complex of human pregnane X receptor and the macrolide antibiotic rifampicin. Mol Endocrinol, 2005. 19(5): p. 1125-34.

43. Teotico, D.G., et al., Structural basis of human pregnane X receptor activation by the hops constituent colupulone. Mol Pharmacol, 2008. 74(6): p. 1512-20.

44. Watkins, R.E., et al., The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity. Science, 2001. 292: p. 2329-2333.

45. Xue, Y., et al., Crystal structure of the pregnane X receptor-estradiol complex provides insights into endobiotic recognition. Mol Endocrinol, 2007. 21(5): p. 1028-38.

46. Bachmann, K., et al., PXR and the regulation of apoA1 and HDL-cholesterol in rodents. Pharmacol Res, 2004. 50(3): p. 237-46.

47. Ekins, S., et al., Human pregnane X receptor antagonists and agonists define molecular requirements for different binding sites. Mol Pharmacol, 2007. 72: p. 592-603.

48. Ekins, S. and J.A. Erickson, A pharmacophore for human pregnane-X-receptor ligands. Drug Metab Dispos, 2002. 30: p. 96-99.

49. Schuster, D. and T. Langer, The identification of ligand features essential for PXR activation by pharmacophore modeling. J Chem Inf Model, 2005. 45(2): p. 431-9.
50. Yasuda, K., et al., *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, 2008. **36**(8): p. 1689-97.

51. Khandelwal, A., et al., *Machine learning methods and docking for predicting human pregnane X receptor activation*. Chem Res Toxicol, 2008. **21**(7): p. 1457-67.

52. Kortagere, S., et al., *Hybrid scoring and classification approaches to predict human pregnane X receptor activators*. Pharm Res, 2009. **26**(4): p. 1001-11.

53. Ekins, S., et al., *Evolution of pharmacologic specificity in the pregnane X receptor*. BMC Evol Biol, 2008. **8**: p. 103.

54. Kortagere, S., et al., *Hybrid scoring and classification approaches to predict human pregane X receptor activators*. Pharm Res, 2009. **26**: p. 1001-11.

55. Kortagere, S., et al., *Evaluation of Computational Docking to Identify PXR Agonists in the ToxCastTM Database* Env Health Perspect, 2010. **118**: p. 1412-1417.

56. Khandelwal, A., et al., *Machine learning methods and docking for predicting human pregnane X receptor activation*. Chem Res Toxicol, 2008. **21**: p. 1457-1467.

57. Kortagere, S. and S. Ekins, *Troubleshooting computational methods in drug discovery*. J Pharmacol Toxicol Methods, 2010. **61**(2): p. 67-75.
58. Ai, N., et al., *Understanding nuclear receptors using computational methods*. Drug Discov Today, 2009. **14**(9-10): p. 486-94.

59. Reid, D., et al., *LASSO-ligand activity by surface similarity order: a new tool for ligand based virtual screening*. J Comput Aided Mol Des, 2008. **22**(6-7): p. 479-87.

60. Huang, N., B.K. Shoichet, and J.J. Irwin, *Benchmarking sets for molecular docking*. J Med Chem, 2006. **49**(23): p. 6789-801.

61. Ung, C.Y., et al., *In silico prediction of pregnane X receptor activators by machine learning approaches*. Mol Pharmacol, 2007. **71**(1): p. 158-68.

62. Weininger, D., *SMILES 1. Introduction and encoding rules*. J Chem Inf Comput Sci, 1988. **28**: p. 31.

63. McGaughey, G.B., et al., *Comparison of topological, shape, and docking methods in virtual screening*. J Chem Inf Model, 2007. **47**(4): p. 1504-19.

64. Murthy, L.R., et al., *Characterization of steroid receptors in human prostate using mibolerone*. Prostate, 1986. **8**(3): p. 241-53.

65. Voutchkova, A.M., T.G. Osimitz, and P.T. Anastas, *Toward a comprehensive molecular design framework for reduced hazard*. Chem Rev, 2010. **110**(10): p. 5845-82.
Legend of Figures

Figure 1. Enrichment plot for AR. Legend: blue dotted line = random hit rate, pink line = eHiTS docking score-based enrichment for strong and moderate binders (i.e. strong (LogIC50 mean = -7.7 stdev=0.7, moderate LogIC50 mean = -5.1, stdev = 0.6, probe DHT) from two crystal structures (1XNN and 2AMA), yellow line = LASSO DUD AR androgen receptor trained ligands (ChemSpider value with strong and moderate binding ligands as “hit” criteria), light blue = eHiTs minimum energy docking score from two crystal structures using strong binder hit criteria only (crystal structures 1XNN and 2AMA), Purple = DUD LASSO using strong binders hit criteria only.
Figure 2. Enrichment plot for ER. Legend: purple dotted line = random hit rate, blue line = the minimum eHiTS docking score reported for two different crystal structures (Min_eHiTS, crystal structure = 3ERT and 1GWR), Pink = using the default LASSO descriptors (LASSO-def), Yellow = using the DUD AR agonist trained descriptors as found on ChemSpider (agonist DUD LASSO), light blue = using the DUD AR antagonist trained descriptors as found on ChemSpider (antagonist DUD LASSO).
Figure 3. A. LASSO models implemented in ChemSpider. B. LASSO scores for Mibolerone, a strong AR and ER binder found with a LASSO search on ChemSpider.
The LASSO score is significant only for 4 targets: MR, AR, ER and PR while the rest of the 40 targets are all 0.10 or below for this molecule.
A. ChemSpider

Search

Find structure with LASSO Score ≥ 0.85 for ACE, angiotensin-converting enzyme and ≤ 0.10

- Nuclear Hormone Receptors
  - AR, androgen receptor
  - ER, estrogen receptor, agonist
  - ER, estrogen receptor, antagonist
  - GR, glucocorticoid receptor
  - MR, mineralocorticoid receptor
  - PPARg, peroxisome proliferator activated receptor
  - PR, progesterone receptor
  - RXRa, retinoic X receptor R

- Kinases
  - CDK2, cyclin-dependent kinase 2
  - EGF, epidermal growth factor receptor
  - FGF1, fibroblast growth factor receptor kinase
  - HSP90, human heat shock protein 90
  - P38 MAP, P38 mitogen activated protein
  - PDGFRb, platelet derived growth factor receptor kinase
  - SRC, tyrosine kinase SRC
  - TK, thymidine kinase
  - VEGFR2, vascular endothelial growth factor receptor

B. SimBioSys LASSO

| Category                     | Target                              | PDB Code | LASSO Score |
|------------------------------|-------------------------------------|----------|-------------|
| Nuclear Hormone Receptors    | MR, mineralocorticoid receptor      | 2aaa2    | 0.88        |
| Nuclear Hormone Receptors    | AR, androgen receptor               | 1xq2     | 0.42        |
| Nuclear Hormone Receptors    | ER, estrogen receptor, agonist      | 1li2     | 0.41        |
| Nuclear Hormone Receptors    | PR, progesterone receptor           | 1sr7     | 0.28        |
| Nuclear Hormone Receptors    | GR, glucocorticoid receptor         | 1m22     | 0.10        |
| Metalloenzymes                | ACE, angiotensin-converting enzyme  | 1e86     | 0.03        |
| Nuclear Hormone Receptors    | RXRa, retinoic X receptor R         | 1mwc     | 0.03        |
| Nuclear Hormone Receptors    | PPARg, peroxisome proliferator activated receptor | 1m9a | 0.02 |
| Kinases                       | SRC, tyrosine kinase SRC            | 2src     | 0.01        |
| Other Enzymes                 | HMGR, hydroxymethylglutaryl-CoA reductase | 1hw8 | 0.01 |
Figure 4. PXR LASSO 7 models derived in training method I.
Figure 5. PXR LASSO enrichment for model from set 1.
Figure 6. PXR LASSO average of 5 models over 5 runs derived in training method II.
