Leadscope Enterprise model for human CYP3A4 induction *in vitro* (U.S. NIH data)

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for human CYP3A4 induction *in vitro* (U.S. NIH data), Danish QSAR Group at DTU Food

1.2 Other related models

None

1.3. Software coding the model

Leadscope Predictive Data Miner, a component of Leadscope Enterprise Server version 3.5.3-5.

2. General information

2.1 Date of QMRF

December 2018

2.2 QMRF author(s) and contact details

QSAR Group at DTU Food;

Danish National Food Institute at the Technical University of Denmark;

http://qsar.food.dtu.dk/;

qsar@food.dtu.dk

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Maria Nektaria Triantafillopoulos;

National Food Institute at the Technical University of Denmark;

2.3 Date of QMRF update(s)

None

2.4 QMRF update(s)

None

2.5 Model developer(s) and contact details

Nikolai Nikolov

National Food Institute at the Technical University of Denmark

Eva Wedebye

National Food Institute at the Technical University of Denmark

Sine Rosenberg

PhD student 2013-2017 at the National Food Institute at the Technical University of Denmark

Marianne Dybdahl

National Food Institute at the Technical University of Denmark

2.6 Date of model development and/or publication

August 2018

2.7 Reference(s) to main scientific papers and/or software package

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., and Blower, P. E. J. (2000) LeadScope: Software for Exploring Large Sets of Screening Data. *Chem. Inf. Comput. Sci.*, 40, 1302-1314.

Cross, K.P., Myatt, G., Yang, C., Fligner, M.A., Verducci, J.S., and Blower, P.E. Jr. (2003) Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.*, *46,* 4770-4775.

Valerio, L. G., Yang, C., Arvidson, K. B., and Kruhlak, N. L. (2010) A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, 6:4, 505-518.

2.8 Availability of information about the model

The training data set is non-proprietary. Experimental data which was applied to make the training set was kindly provided by U.S. NIH NCATS (National Center for Advancing Translational Sciences) with chemical structure information and qHTS screening results for binding to the Ligand Binding Domain (LBD) of human PXR at the protein level (hPXR-LBD). Generation of the experimental data is described in Shukla et al. 2011. The model algorithm is proprietary from commercial software.

2.9 Availability of another QMRF for exactly the same model
No other available QMRF for this model

3. Defining the endpoint

3.1 Species

Human (full-length human PXR, DPX-2 cell line constructed by using HepG2 cells derived from human hepatocellular carcinoma)

3.2 Endpoint
QMRF 4. Human Health Effects

QMRF 4.18.a. Endocrine Activity. CYP3A4 induction .

3.3 Comment on endpoint

From Rosenberg et al. 2017:

The nuclear receptor (NR) superfamily is a large group of transcription factors that control expression of multiple genes involved in a broad range of biological processes, such as development, homeostasis and metabolism. The transcriptional activity of NRs is primarily regulated through ligand binding [1]. The Pregnane X Receptor (PXR), first described by Kliewer and colleagues in 1998, is a member of the NR superfamily [2,3]. PXR is mainly expressed in the liver, intestine and kidneys, and plays a key role in the regulation of genes involved in the metabolism and efflux of endogenous hormones and xenobiotic molecules [3–5]. The genes regulated by PXR include genes encoding enzymes, such as cytochrome P450s (CYPs), glucuronyltransferases and sulfotransferases, as well as transporters, such as P-glycoprotein and multidrug resistance proteins [2,3,6–8]. The ligand-binding domain (LBD) of PXR is large and flexible, and can change its shape to accommodate structurally diverse molecules including steroids, bile acids, antibiotics, statins, and pesticides [9,10]. A considerable amount of inter-species variation has been observed in the PXR LBD with human, rabbit and rat sharing roughly 75-80% amino acid identity [11,12]. There are numerous examples of differences in ligand binding to PXR and resulting downstream transcription of enzymes and transporters between species, which complicates the extrapolation of results from in vivo animal studies to humans [11,13–15].

Among the many PXR target genes is the gene encoding CYP3A4, an oxidizing enzyme involved in phase I metabolism of various compounds [4,31]. CYP3A4 is considered the main drug-metabolizing CYP isoform in the human liver and is involved in the metabolism of more than 50% of drugs on the market [2,5]. In most cases, CYP3A4 causes chemicals to become less biologically active and promotes their excretion; but in other cases it has the opposite effect causing bioactivation by converting them to metabolites that are more toxic than the parent molecule.

Findings from previous studies indicate that there is an association between PXR activation by environmental chemicals and adverse health effects [15,18,20,21]. The importance of PXR activation is also reflected in a number of suggested adverse outcome pathways (AOPs) available from the online AOP-Wiki [22], for example an AOP describing how activation of PXR and other related NRs upregulate thyroid hormone catabolism resulting in hypothyroidism and subsequent adverse neurodevelopmental outcomes [23]. The AOPs are envisioned to promote the industry's and regulators' use of results from alternative methods such as in vitro tests and computational models in chemical risk assessments to reduce, refine or replace traditional animal tests [24–26], for example by applying the AOP in an Integrated Approaches to Testing Assessment (IATA) context to support regulatory decisions [27].

Compound collection, qHTS assays and the classification of the qHTS results into actives, inconclusives and inactives have been described previously by Shukla et al. 2009, Shukla et al. 2011 and Huang et al. 2011.

The training set was created by combining the training set and the external validation set for CYP3A4 induction reported in Rosenberg et al. 2017.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Human CYP3A4 induction, positive or negative.

3.6 Experimental protocol

The experimental protocol is described in detail in Shukla et al. 2011. A cell-based *in vitro* assay was used together with qHTS assay to screen drugs and bioactive compounds for their ability to induce human CYP3A4. Induction of CYP3A4 was measured by a luciferase reporter readout separately from PXR activation in DPX-2 cells using P450-Glo Screening Systems. Luciferin assay substrate was added after incubation and P450-Glo reagent was added to the plates followed by further incubation before reading on ViewLux plate reader. P450-Glo is a robust reagent and has previously been used for CYP3A4 profiling with little to no fluorescence interference. Rifampicin was used as positive control to assess maximal response in CYP3A4 induction assay.

3.7 Endpoint data quality and variability

The primary qHTS measuring CYP3A4 induction assay performed well and revealed 212 active compounds (8%) with high quality CRCs. Less than 50% of compounds displayed potency values <10 µM. The activities of 70 compounds out of 72 retested were confirmed, yielding a 97% confirmation rate. (Shukla et al. 2011).

The datasets originate from the same source, i.e. NIH NCATS. All chemicals have been screened in the same testing protocols and undergone the same data processing, and this has likely contributed to decrease the experimental variability.

No measures of the reproducibility of the overall positive and negative end calls as used for QSAR model was available. Still, the data in training set x and the test set were assessed to be of high quality and expected to be a good basis for QSAR model development.

4. Defining the algorithm

4.1 Type of model

A categorical QSAR model based on structural features and numeric molecular descriptors.

4.2 Explicit algorithm

This is a categorical QSAR model made by use of partial logistic regression (PLR). The model is a 'Cocktail model', see 4.4, that integrates a so-called single model and a Leadscope composite model with 10 sub-models, i.e. the cocktail composite model contains 11 sub-models. The specific implementation is proprietary within the Leadscope software.

4.3 Descriptors in the model

ALogP,

Hydrogen Bonds Acceptors and Donors,

Lipinski Score,

Molecular Weight,

Parent Atom Count,

Parent Molecular Weight,

Polar Surface Area,

Number of rotational bonds,

Structural features.

4.4 Descriptor selection

Leadscope Predictive Data Miner is a software program for systematic sub-structural analysis of a chemical using predefined structural features stored in a template library, training set-dependent generated structural features (scaffolds) and calculated molecular descriptors. The feature library contains approximately 27,000 pre-defined structural features and the structural features chosen for the library are motivated by those typically found in small molecules: aromatics, heterocycles, spacer groups, simple substituents. Leadscope allows for the generation of training set-dependent structural features (scaffold generation), and these features can be added to the pre-defined structural features from the library and be included in the descriptor selection process. It is possible in Leadscope to

remove redundant structural features before the descriptor selection process and only use the remaining features in the descriptor selection process. Besides the structural features Leadscope also calculates eight molecular descriptors for each training set structure: the octanol/water partition coefficient (alogP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), Lipinski score, atom count, parent compound molecular weight, polar surface area (PSA) and rotatable bonds. These eight molecular descriptors are also included in the descriptor selection process.

Leadscope has a default automatic descriptor selection procedure. This procedure selects the top 30% of the descriptors (structural features and molecular descriptors) according to X^2-test for a binary variable, or the top and bottom 15% descriptors according to t-test for a continuous variable. Leadscope treats numeric property data as ordinal categorical data. If the input data is continuous such as IC_{50} or cLogP data, the user can determine how values are assigned to categories: the number of categories and the cut-off values between categories. (Roberts *et al.*2000).

Five predictive models where build for each of the four training sets where the following five different modelling approaches in LPDM where used:

1. 'Single model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors for descriptor selection.
2. 'Single model' using both the Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.
3. 'Composite model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors in the descriptor selection.
4. 'Composite model' using both the Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.
5. 'Cocktail model' integrating a 'Single model' and a 'Composite model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors for descriptor selection.
6. 'Cocktail model' integrating a 'Single model' and a 'Composite model' using both the Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection

Based on model performance as measured by preliminary Leadscope Predictive Data Miner cross-validation the model developed using approach 5, integrating number 1 and 3 into a cocktail composite model, was chosen.

Descriptors were automatically selected among the structural features and the eight molecular descriptors.

4.5 Algorithm and descriptor generation

For descriptor generation see 4.4.

After selection of descriptors the Leadscope Predictive Data Miner program performs partial least squares (PLS) regression for a continuous response variable, or partial logistic regression (PLR) for a binary response variable, to build a predictive model. By default the Predictive Data Miner performs leave-one-out or leave-groups-out (in the latter case, the user can specify any number of repetitions and percentage of structures left out) cross-validation on the training set depending on the size of the training set. In the cross-validation made by Leadscope the descriptors selected for the 'mother model' are used when building the validation sub-models and they may therefore have a tendency to give overoptimistic validation results.

In this model because of the categorical outcome in the response variable PLR was used to build the predictive model. Because of the unbalanced training set (i.e. 180 positives and 2091 negatives) 7-10 sub-models for smaller individual training sets were made. The descriptors for each of the sub-models were automatically selected from the Leadscope feature library based solely on the training set compounds used to build the individual sub-model and was not affected by the full training set chemicals. Therefore, a different number of descriptors (structural features and molecular descriptors) were selected and distributed on varying number of PLS factors for each sub-model.

4.6 Software name and version for descriptor generation

Leadscope Predictive Data Miner, a component of Leadscope Enterprise, Server version 3.5.3-5.

4.7 Descriptors/chemicals ratio

As this model is a composite model consisting of 11 sub-models with varying training set size and using different descriptors and number of PLS factors (see 4.5), an overall descriptor/chemical ratio for this model cannot be calculated. The data for individual models are as follows:

Model	Substances	Descriptors	PLS factors
1	2,271	221	2
2	389	271	3
3	389	267	4
4	389	276	4
5	389	266	2
6	389	268	3
7	389	256	4
8	389	268	2
9	389	267	4
10	390	270	5
11	389	290	4

5. Defining Applicability Domain

5.1 Description of the applicability domain of the model

The definition of the applicability domain consists of two components; the definition of a structural domain in Leadscope and the in-house further probability refinement algorithm on the output from Leadscope to reach the final applicability domain call.

1. Leadscope

For assessing if a test compound is within the structural applicability domain of a given model Leadscope examines whether the test compound bears enough structural resemblance to the training set compounds used for building the model (i.e. a structural domain analysis). This is done by calculating the distance between the test compound and all compounds in the training set (distance = 1 - similarity). The similarity score is based on the Tanimoto method. The number of neighbours is defined as the number of compounds in the training set that have a distance equal to or smaller than 0.7 with respect to the test compound. The higher the number of neighbours, the more reliable the prediction for the test compound. Statistics of the distances are also calculated. Effectively no predictions are made for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated in Leadscope.

2. The Danish QSAR group

In addition to the general Leadscope structural applicability domain definition the Danish QSAR group has applied a further requirement to the applicability domain of the model. That is only positive predictions with a probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 are accepted. Predictions within the structural applicability domain but with probability equal to or greater than 0.5 and less than 0.7, or greater than 0.3 and less than 0.5 are defined as positives out of applicability domain and negatives out of applicability domain, respectively. When these predictions are weeded out the performance of the model in general increases at the expense of reduced model coverage.

5.2 Method used to assess the applicability domain

Leadscope does not generate predictions for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated.

Only positive predictions with probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 are accepted.

5.3 Software name and version for applicability domain assessment

Leadscope Predictive Data Miner (LPDM), a component of Leadscope Enterprise Server version 3.5.3-5.

5.4 Limits of applicability

The Danish QSAR group applies an overall definition of structures acceptable for QSAR processing which is applicable for all the in-house QSAR software, i.e. not only Leadscope. According to this definition accepted structures are organic substances with an unambiguous structure, i.e. so-called discrete organics defined as: organic compounds with a defined two dimensional (2D) structure containing at least two carbon atoms, only certain atoms (H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I), and not mixtures with two or more 'big components' when analysed for ionic bonds (for a number of small known organic ions assumed not to affect toxicity the 'parent molecule' is accepted). Calculation 2D structures (SMILES and/or SDF) are generated by stripping off ions (of the accepted list given above). Thus, all the training set and prediction set chemicals are used in their non-ionized form. See 5.1 for further applicability domain definition.

6. Internal validation

6.1 Availability of the training set

Yes, available upon request (see Rosenberg et al. 2017, however please note that in this model we have integrated the training set and validation set from Rosenberg et al. 2017 to constitute an expanded training set).

6.2 Available information for the training set

The file contains CAS, SMILES and activity call for each substance.

6.3 Data for each descriptor variable for the training set

No

6.4 Data for the dependent variable for the training set

Yes

6.5 Other information about the training set

2,271 compounds are in the training set: 180 positives and 2,091 negatives.

The training set was created by combining the training set and the external validation set for CYP3A4 induction, see Rosenberg et al. 2017 for more information.

6.6 Pre-processing of data before modelling

Only structures acceptable for Leadscope were used in the final training set. That is only discrete organic chemicals as described in 5.4 were used. In case of replicate structures, one of the replicates was kept if all the compounds had the same activity and all were removed if they had different activity. No further structures accepted by the software were eliminated (i.e. outliers).

6.7 Statistics for goodness-of-fit

Not performed.

6.8 Robustness – Statistics obtained by leave-one-out cross-validation

Not performed. (It is not a preferred measurement for evaluating large models).

6.9 Robustness – Statistics obtained by leave-many-out cross-validation

A two times five-fold 20 % cross-validation was performed using the Leadscope Predictive Data Miner built-in procedure. NB the implementation in LPDM may sometimes give too optimistic results.

Predictions within the defined applicability domain of the validation sub-models were pooled and Cooper's statistics calculated. This gave the following results for the predictions which were within the applicability domains of the respective sub-models:

− Sensitivity (true positives / (true positives + false negatives)): 86.7%

− Specificity (true negatives / (true negatives + false positives)): 98.2%

− Balanced Accuracy ((Sensitivity + Specificity) /2): 92.5%

− Coverage ((In-Domain predictions) / (All predictions): 85.4%

6.10 Robustness - Statistics obtained by Y-scrambling

Not performed.

6.11 Robustness - Statistics obtained by bootstrap

Not performed.

6.12 Robustness - Statistics obtained by other methods

Not performed.

7. External validation

The predecessor to the model reported here, i.e. the non-expanded QSAR model reported in Rosenberg et al. 2017, was validated by a procedure where the experimental results were blinded by U.S. NIH NCATS to the DTU QSAR developers with the results as reported here in chapter 7.

7.1 Availability of the external validation set

Yes, available upon request, see Rosenberg et al. 2017.

7.2 Available information for the external validation set

For each substance in the external validation set, CAS, SMILES and activity call is available.

7.3 Data for each descriptor variable for the external validation set

No

7.4 Data for the dependent variable for the external validation set

All

7.5 Other information about the validation set

The experimental protocol for the test set substances is identical to the one for the training set described in section 3.

7.6 Experimental design of test set

The external validation set consisted of about one third of the substances and was masked by NIH NCATS for DTU-blinded external validation after the model development was finished.

The selection of the test sets was designed and made by NIH NCATS scientists, who clustered all compounds in the dataset on structural similarity using the Euclidian distance and then, within each structure cluster and for each of the four endpoints, approximately one-third actives and one-third inactives were selected randomly.

Thus the training and test sets are structurally comparable and have similar distributions of actives and inactives. NIH NCATS sent the training sets containing structure information and experimental results and the test sets containing only structure information to the National Food Institute (Food) at the Technical University of Denmark (DTU), who performed the structure preparations, the model development and validations as well as the virtual screenings.

7.7 Predictivity – Statistics obtained by external validation

From Table 2 in Rosenberg et al. 2017:

QSAR model	Statistical parameter	Cross-validation, % (SD, %)	External validation, %
		5 times 2-fold*	Blinded test sets**
CYP3A4 Approach 5) 9 sub-models	Coverage	64.7 (3.0)	63.4 (453/715)
	Sensitivity	71.6 (7.6)	76.9 (20/26)
	Specificity	80.7 (2.7)	85.5 (365/427)
	Balanced accuracy	76.1 (3.3)	81.2

*A five times two-fold cross-validation with same active-inactive ratio as the full training set and without reusing selected descriptors from the parent model. Coverage, sensitivity and specificity are the mean from the ten cross-validation models with the standard deviation (SD) in parentheses. ** The experimental results of the test set structures were made available to DTU Food by NIH NCATS after they had been predicted in the respective models by DTU Food.

7.8 Predictivity – Assessment of the external validation set

As described under 7.6 all compounds in the total dataset were clustered on structural similarity using the Euclidian distance and within each structure cluster approximately one-third actives and one-third inactives were selected randomly for the external validation set.

From Rosenberg et al. 2017: *The distributions of experimentally active and inactive structures in these external test sets are imbalanced toward more inactives similar to the training set distributions. Although the masked test sets in total are quite large for external validation, the few actives make the calculations of sensitivity less robust.*

7.9 Comments on the external validation of the model

A special set-up between the collaborators on the development of this model was set up, where the developed QSAR model was validated by a procedure where the experimental results were blinded by U.S. NIH NCATS to the DTU QSAR developers.

8. Mechanistic interpretation

8.1 Mechanistic basis of the model

The global model identifies structural features and molecular descriptors which in the model development was found to be statistically significant associated with effect. Many predictions may indicate modes of action that are obvious for persons with expert knowledge for the endpoint.

8.2 A priori or posteriori mechanistic interpretation

A posteriori mechanistic interpretation. The identified structural features and molecular descriptors may provide basis for mechanistic interpretation.

8.3 Other information about the mechanistic interpretation

9. Miscellaneous information

9.1 Comments

None

9.2 Bibliography

Rosenberg. A., Xia M., Huang R., Nikolov N. G., Wedebye E. B., Dybdahl M., QSAR Development and Profiling of 72,524 REACH Substances for PXR Activation and CYP3A4 Induction, Computational Toxicology,.1 (2017) 39-48. doi: 10.1016/j.comtox.2017.01.001.

R. Huang, N. Southall, Y. Wang, A. Yasgar, P. Shinn, A. Jadhav, D.-T. Nguyen, C.P. Austin, The NCGC Pharmaceutical Collection: A comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics, Sci. Transl. Med. 3 (2011) 80ps16. doi:10.1126/scitranslmed.3001862.

S.J. Shukla, D.-T. Nguyen, R. MacArthur, A. Simeonov, W.J. Frazee, T.M. Hallis, B.D. Marks, U. Singh, H.C. Eliason, J. Printen, C.P. Austin, J. Inglese, D.S. Auld, Identification of Pregnane X Receptor Ligands Using Time-Resolved Fluorescence Resonance Energy Transfer and Quantitative High-Throughput Screening, Assay Drug Dev. Technol. 7 (2009) 143–169. doi:10.1089/adt.2009.193.

S.J. Shukla, S. Sakamuru, R. Huang, T.A. Moeller, P. Shinn, D. VanLeer, D.S. Auld, C.P. Austin, M. Xia, Identification of Clinically Used Drugs That Activate Pregnane X Receptors, Drug Metab. Dispos. 39 (2011) 151–159. doi:10.1124/dmd.110.035105.