Leadscope Enterprise model for *in vitro* human constitutive androstane receptor (hCAR) inhibition at concentrations up to 50 µM.

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for *in vitro* human constitutive androstane receptor (hCAR) inhibition at concentrations up to 50 µM.

1.2 Other related models

Leadscope Enterprise model for *in vitro* human constitutive androstane receptor (hCAR) inhibition at concentrations up to 20 µM.

1.3. Software coding the model

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

2. General information

2.1 Date of QMRF

July 2020

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

Ana Caroline Vasconcelos Martins
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

Kazue Kelly Chinen
University of California, Los Angeles, CA 90095, USA

Eva Bay Wedebye
National Food Institute at the Technical University of Denmark

Nikolai Georgiev Nikolov
National Food Institute at the Technical University of Denmark

Kyrylo Klimenko
Postdoc 2017-2018 at DTU Food

2.6 Date of model development and/or publication

Development finalized in 2019 and published in 2020

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. *J. 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 and the experimental data which was applied to make the training set originates from the Tox21 Program's high-throughput in vitro assay and is available at https://tripod.nih.gov/tox21/assays/.

The model algorithm is proprietary from commercial software.

2.9 Availability of another QMRF for exactly the same model

No other QMRFs are available for this model

3. Defining the endpoint

3.1 Species

Human hepatoma (HepG2) cells transfected with a double-stable human CAR and CYP2B6-2.2kb.

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.18.b. Receptor binding and gene expression (*in vitro* human constitutive androstane receptor (hCAR) agonism)

3.3 Comment on endpoint

The constitutive androstane receptor (CAR) belongs to the human nuclear receptor (NR) superfamily, a 48-member group of “orphan” and “adopted-orphan” NRs. In humans, the CAR protein is encoded by the NR1I3 gene from the NR subfamily 1, group I, member 3. The NR subfamily 1 group I also includes the Vitamin D Receptor (VDR) and the Pregnane X Receptor (PXR). CAR displays so-called constitutive activity, meaning that it is active also in the absence of a ligand. Many known CAR agonists are also species-specific. CAR is expressed mainly in the liver and small intestine and mediates the induction of metabolizing enzymes, such as cytochrome P450 3A (CYP3A) isoenzymes, conjugation enzymes such as UDP glucuronosyltransferase family 1 member A1, and transporters such as P-glycoprotein. Along with the NR PXR, CAR is a principal regulator of the metabolism of xenobiotic compounds. PXR and CAR cross-regulate their target genes cytochrome P450 (CYP) CYP2B and CYP3A. CAR also plays an important role in the metabolism of a number of endogenous substances such as thyroid and steroid hormones, cholesterol, bile acids, bilirubin, glucose, and lipids. CAR inhibition may have negative consequences, namely, decreased metabolizing potential in the body, which leads to decreased turnover of endogenous hormones as well as decreased detoxification and excretion of xenobiotics.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives

3.5 Dependent variable

Human CAR inhibition: positive or negative

3.6 Experimental protocol

The hCAR activation U.S. Tox21 qHTS in vitro assay applied in AID 1224838 is a luminescence-based assay using human hepatoma (HepG2) cell line transfected with a double-stable human CAR and CYP2B6-2.2kb.

Substances that activate hCAR result in expression of the luciferase reporter gene and the level of luciferase activity is an indirect measure of hCAR activation.

For the PubChem AID 1224838 assay, compounds were tested in triplicate at 16 different concentrations with varying concentration ranges among the different substances. The assay is multiplexed with a cell viability assay to differentiate true hCAR agonists from cytotoxic substances AID 1224837.

Some substances can stabilize luciferase and increase its half-life resulting in its accumulation and a measured increase in luminescence signal. Such substances may be incorrectly interpreted as hCAR activators in the applied hCAR agonism qHTS assay and we therefore applied AID 1224835 to identify luciferase stabilizers to exclude them from QSAR modeling.

3.7 Endpoint data quality and variability

The assay results used in the development of the model were provided by the U.S. Tox21 Program. These datasets were used as a basis for our study as well as computer-readable structure-data files (SDF) on the

tested chemicals substances structures from PubChem AID 1224838 on small molecule agonists of the hCAR signaling pathway. All chemicals have been screened in the same testing protocol and undergone the same data processing which may have led to a decrease of the experimental variability.

We undertook further QSAR-targeted processing of the Tox21 hCAR data by setting criteria for absolute activity for actives, and just as importantly for QSAR models development purposes, by setting criteria to only select the most robust inactives. We apply the 20 µM potency cut-off with a 25% absolute effect. For each substance, our QSAR-targeted process led to the assignment of one of the following outcomes: “active”, “inactive”, or “inconclusive”. Only actives and inactives were used for QSAR development and validation. For the data processing, we filtered each test CRS through in-house tools, specifically developed for the purpose of determining active responses with non-cytotoxic concentrations showing at least 25% effect (in absolute value), accepting only the best Tox21 Hill curve classes. For inactives, we required Tox21 Hill curve class 4 (i.e. inactive) and that the substance exhibited no cytotoxicity up to a 10 µM concentration.

Dataset for the training set originates from the U.S. Tox21 Program and is presented in PubChem databases AID 1224838 and AID 1224837 that both are luminescence-based assays (see 3.6 for protocol description). Additionally to this, data from AID 1224835, with luciferase as endpoint, was incorporated in the data curation, justified by the potential of false positives resulting from an initial inhibition of the enzyme leading to an increase in half-life and accumulation within the cell that can be measured as an increase in luminescence signal that can be interpreted as identification of a compound activating hCAR. Thus compounds classified as actives in the AID 1224835 were excluded from the training set.

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 LPDM composite model consisting of 10 sub-models, using all the positives (170 chemicals) in each of these and different subsets of the negatives (1700 chemicals) (see 4.4), i.e. the cocktail composite model contains 11 sub-models. The specific implementation is proprietary within the LPDM 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

LPDM is a software program for systematic sub-structural analysis of a substance 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. LPDM 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 LPDM 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 LPDM 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 substance molecular weight, polar surface area (PSA) and rotatable bonds. These eight molecular descriptors are also included in the descriptor selection process.

LPDM has a default automatic descriptor pre-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. LPDM 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).

After pre-selection of descriptors the LPDM program performs partial least squares (PLS) regression for a continuous response variable, or 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 using the built-in LPDM functionality, 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 the categorical outcome in the response variable PLR was used to develop the predictive model. Development of a PLR predictive model starts with the pre-selected descriptors with further selection of descriptors in an iterative procedure, and selection of the optimum number of factors based on minimizing the predictive residual sum of squares.

Composite models were developed with creation of a number of sub-models and by using three QSAR modelling approaches in which all underwent a 10 times 20 % - out LPDM cross-validation:

1. A single model, i.e. a non-composite model using the full training set.
2. A composite model, with a number of sub-models of equal weight based on balanced training subsets.
3. A composite 'cocktail' model, combining the single model from 1) with the sub-models of the composite model from 2).

The descriptors for each of the sub-models are automatically selected from the LPDM feature library based solely on the training set substances used to build the individual sub-models and was not affected by the full training set substances. Therefore, a different number of descriptors (structural features and molecular descriptors) are selected and distributed on varying number of PLS factors for each sub-model.

Because of the unbalanced training set (i.e. 170 positives and 1700 negatives) 10 sub-models for smaller individual training sets were made in the composite approach (point 2), and a single model was also developed (point 1) and integrated with the composite model in a 'cocktail' model (point 3).

Based on model performance as measured by a LPDM cross-validation the model developed using approach 3 integrating number 1 and 2 into a cocktail composite model was chosen.

4.5 Algorithm and descriptor generation

Algorithm and descriptor generation takes place in LPDM in a process integrated with descriptor selection and therefore the whole subject is described in section 4.4.

4.6 Software name and version for descriptor generation

LPDM, a component of Leadscope Enterprise 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 4.4), an overall descriptor/chemical ratio for this model cannot be calculated. The data for individual models as follows:

Name of the model	Substances	Descriptors	PLS factors
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-1	340	193	4
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-2	340	204	4
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-3	340	188	3
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-4	340	189	4
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-5	340	188	4
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-6	340	184	5
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-7	340	194	5
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-8	340	190	3
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-9	340	192	3
CAR_Ant_50uM_TOTALTrain_1870_Multiple_Scaffolds_Model-10	340	207	1
CAR_Ant_50uM_TOTALTrain_1870_Scaffolds_1_Model	1870	301	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 LPDM and an in-house further probability refinement algorithm on the output from LPDM to reach the final applicability domain call.

1. LPDM

For assessing if a test compound is within the structural applicability domain of a given model LPDM 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 Jaccard / Tanimoto method and using the LPDM predefined library of 27,000 features. 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. Furthermore, LPDM requires that the test compound contains at least one model feature or scaffold from the model. 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 LPDM.

2. The Danish QSAR group

In addition to the general LPDM 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 between 0.5 to 0.7 or 0.3 to 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

DTU-developed in-house post-treatment procedure to assign domain flags according to the description in 5.1.

LPDM 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.

5.3 Software name and version for applicability domain assessment

LPDM, a component of Leadscope Enterprise 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 LPDM. 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

It will be available in the Danish QSAR Data Base.

6.2 Available information for the training set

PUBCHEM SID and activity call according to the DTU in-house processing 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

For the final model, 1870 compounds are in the training set: 170 positives and 1700 negatives. The initial model has in the training set 136 positives and 1360 negatives.

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

These results were reported in Chinen et al. 2020.

A two times five-fold (i.e. 20 % out) cross-validation by DTU Food cross-validation procedure was performed.

Cooper's statistics were calculated for each of the left-out sets for predictions within the defined applicability domain from the ten validation sub-models and used to calculate average values and standard deviations. This gave the following results for the predictions which were within the applicability domains of the respective sub-models:

Final model:

- Sensitivity (true positives / (true positives + false negatives)): 72.4±10.2%
- Specificity (true negatives / (true negatives + false positives)): 91.6±1.5%
- Balanced Accuracy ((Sensitivity + Specificity) /2): 82.0±5.0%

Initial model (leaving 20% out for external validation):

- Sensitivity (true positives / (true positives + false negatives)): 72.4±14.3%
- Specificity (true negatives / (true negatives + false positives)): 92.6±2.4%
- Balanced Accuracy ((Sensitivity + Specificity) /2): 82.5±7.9%

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

7.1 Availability of the external validation set

Yes

7.2 Available information for the external validation set

PUBCHEM SID and activity call according to the DTU in-house processing for each substance.

7.3 Data for each descriptor variable for the external validation set

None

7.4 Data for the dependent variable for the external validation set

Yes

7.5 Other information about the validation set

The test set for the final QSAR model is composed of 2,444 inactive substances. The test set for the initial model is composed of 34 active substances and 2,784 inactive substances.

7.6 Experimental design of test set

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

7.7 Predictivity – Statistics obtained by external validation

These results were reported in Chinen et al. 2020.

Final model:

- Sensitivity (true positives / (true positives + false negatives)): N/A
- Specificity (true negatives / (true negatives + false positives)): 92.4%
- Balanced Accuracy ((Sensitivity + Specificity) /2): N/A
- Coverage for negatives ((In-Domain predictions) / (All predictions)): 59.8%

Initial model:

- Sensitivity (true positives / (true positives + false negatives)): 55.0%
- Specificity (true negatives / (true negatives + false positives)): 92.5%
- Balanced Accuracy ((Sensitivity + Specificity) /2): 73.8%
- Coverage ((In-Domain predictions) / (All predictions)): 59.9%

7.8 Predictivity – Assessment of the external validation set

Not performed

7.9 Comments on the external validation of the model

None

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.

CAR inhibition is a mechanistic endpoint related to a number of health outcomes, see section 3.3.

8.3 Other information about the mechanistic interpretation

None

9. Miscellaneous information

9.1 Comments

None

9.2 Bibliography

Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE. (2000) LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.*, 40, 1302–1314. doi: 10.1021/ci0000631.

U.S. Tox21 Program. 2019. Tox21 Data Browser: Assays. Available: https://tripod.nih.gov/tox21/assays/ (accessed September 19 2019).

NIH. 2019. PubChem AID 1224838: qHTS assay to identify small molecule antagonists of the constitutive androstane receptor (CAR) signaling pathway. Available: https://pubchem.ncbi.nlm.nih.gov/bioassay/1224838 [accessed September 12 2019].

NIH. 2019. PubChem AID 1224893: qHTS assay to identify small molecule antagonists of the constitutive androstane receptor (CAR) signaling pathway: Summary. Available: https://pubchem.ncbi.nlm.nih.gov/bioassay/1224893 [accessed September 12 2019].

Chinen KK, Klimenko K, Nikolov NG, Wedebye, EB (2020) QSAR modeling of different minimum potency levels for in vitro human CAR activation and inhibition and screening of 80,086 REACH and 54,971 U.S. substances. *Computational Toxicology*, 14, 100121(1-14). Doi: 10.1016/j.comtox.2020.100121.