Leadscope Enterprise pathway model for mammalian Androgen Receptor (AR) activation *in vitro* (U.S. EPA CoMPARA data)

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for the U.S. EPA Collaborative Modeling Project for Androgen Receptor Activity (U.S. EPA CoMPARA data) binding in mammalian cells *in vitro*, model made by the Danish QSAR Group at DTU Food.

1.2 Other related models

No

1.3. Software coding the model

Leadscope Predictive Data Miner (LPDM), a component of Leadscope Enterprise Server version 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

Ana Caroline Vasconcelos Martins
National Food Institute at the Technical University of Denmark

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

Nikolai Georgiev Nikolov
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

Eva Wedebye
National Food Institute at the Technical University of Denmark

Nikolai Nikolov
National Food Institute at the Technical University of Denmark

2.6 Date of model development and/or publication

Development finalized in 2016 and published in 2020

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

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

Cross KP, Myatt G, Yang C, Fligner MA, Verducci JS, Blower PE Jr. (2003) Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.*, 46, 4770-4775. doi: 10.1021/jm0302703

Valerio LG, Yang C, Arvidson KB, Kruhlak NL. (2010) A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, 6:4, 505-518. doi: 10.1517/17425250903499286

2.8 Availability of information about the model

The training set was kindly provided by the U.S. Environmental Protection Agency (EPA) and is non-proprietary. The model algorithm is proprietary from commercial software. This model was made for the U.S. EPA CoMPARA project.

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

Chimpanzee, mouse and human cell lines (11 biochemical and cell-based *in vitro* assays, of which 9 based on human cell lines, 1 based in chimpanzee cell line and 1 based on mouse cell line).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.18.b. Receptor binding and gene expression (Androgen Receptor)

3.3 Comment on endpoint

From Lynch et al 2016:

"The androgen receptor (AR, NR3C4) is a transcription factor which regulates male sexual development, while also maintaining accessory sexual organ function. The structure of AR includes an N-terminal region which contains the activation function-1 (AF-1), a DNA-binding domain (DBD), a hinge region, and a ligand binding domain (LBD) which contains the ligand-regulated AF-2. AR is an evolutionarily conserved receptor and is closely related to the human glucocorticoid and progesterone receptors including even recognizing analogous DNA response elements. However, these receptors have a different hormone ligand specificity. AR is the main transcription factor implicated in transmitting hormone signals inside the prostate gland. As a key transcription factor regulating male sexual development, altering regulation of this nuclear receptor causes abnormal development of the prostate."

"AR activation can occur through direct or indirect pathways. Direct AR activation occurs through a multi-step process. First, the unliganded receptor sequestered by heat shock proteins and immunophilins in the cytoplasm of the cell binds to a ligand through the LBD. This causes a conformational change allowing for the dissociation from the complex anchoring AR in the cytoplasm. Once free, AR homodimerizes and the nuclear localization signal amino acid sequence becomes exposed. The nuclear localization signal subsequently binds to importins, which then transport AR into the nucleus. Once inside the nucleus, the ligand-receptor complex and its co-activators accumulate at sequence-specific nuclear foci. However, like its other nuclear receptor counterparts, AR can also be activated through multiple other pathways in an indirect manner; direct ligand binding is not necessary."

"Xenobiotic perturbation of AR has many possible adverse outcomes in humans. This includes multiple types of endocrine disruption such as changes in spermatogenesis and the synthesis of sex hormones. AR is a key driver of prostate cancer growth and AR expression and sensitivity have also been shown to increase in the androgen-responsive human prostatic carcinoma (LNCaP) cell line when grown in androgen-depleted medium. Other studies have shown that AR also has an important role in the modulation of multiple additional cancer types including liver, kidney, and bladder, and is linked to hepatocellular hypertrophy. Therefore, recognizing exogenous compounds and environmental chemicals which activate AR is critical in detecting endocrine disrupters and possible cancer modulators."

From Kleinstreuer *et al*. 2017:

"*A multiassay AR pathway model was developed based on the results of 11 ToxCast and Tox21 in vitro HTS assays covering the androgen signaling pathway including receptor-binding, coregulator recruitment, chromatin-binding of the mature transcription factor, and gene transcription and combining the in vitro results into an AUC score representing the whole AR activity to mimic the in vivo results. A certain number of chemicals could be expected to act as true AR agonists or antagonists, but there are also chemicals that are known to interfere with these various assay technologies through false signals such as autofluorescence or cytostatic mechanisms."*

"*Here, the data from 11 AR pathway assays were supplemented with an additional antagonist confirmation assay using a higher concentration of the activating ligand to characterize competitive binding. This battery of in vitro AR assays was used to screen a library of 1855 chemicals. Observed patterns of assay activity included no assays activated, all agonist assays activated, all antagonist assays activated, specific subsets of assays across technologies activated, and technology-specific assay activation. To navigate this complexity in the results, we developed a computational network model to infer whether chemicals that activate specific patterns of in vitro assays were more likely to be AR agonists, AR antagonists, false positives due to specific types of assay interference, or true negatives.*"

"*Evaluating and validating the AR pathway model requires high-quality reference data for AR agonist and antagonist activity. Unlike the ER pathway, which has a well-characterized set of in vitro and in vivo reference chemicals, the reference chemical set for the AR pathway is much less developed. Previous work focused on identifying chemicals that were positive or negative for (anti)androgenicity, without a specific emphasis on potency, and often included compounds that were "presumed" active or inactive. Using a comprehensive list*

of putative AR-active or -inactive chemicals from past and present international validation studies, we performed a literature search to compile high-quality published in vitro AR binding and transactivation (TA) assay data. To facilitate external validation of the AR pathway model results, no ToxCast or Tox21 assay data were included in the literature search. We identified a set of chemicals with reliable and reproducible in vitro results from the literature and binned the chemicals into defined potency categories. The list of proposed reference chemicals and the supporting data are provided and were used to evaluate the current computational model of AR pathway activity based on the Tox21 and ToxCast assays."

3.4 Endpoint units

No units, 1 for positives and 0 for negatives

3.5 Dependent variable

Positives: At least three TA experiments of which at least 70% yielded positive TA results and at least one positive binding result.

Negatives: At least three TA experiments yielding negative results and no TA experiments yielding positive results

3.6 Experimental protocol

See Kleinstreuer et al. 2017

3.7 Endpoint data quality and variability

The data is expected to be of high quality because of the integration of several assays to exclude false positives caused by narrow (technology-specific) or broad assay interference. Also, the variability in the data is expected to be low as for each assay all chemicals have been tested in the same laboratory and the process of assigning an AR binding score using the network model (see 3.2) has been equal for all chemicals.

According to Kleinstreuer et al. 2017, 29 reference chemicals for AR agonism were identified with high-quality in vitro AR results in the literature and were among the 1855 chemicals tested in ToxCast / Tox21. They could therefore be used for performance-based external validation of the AR pathway model results.

Activation (agonism) performance based on 8 positive and 21 negative reference substances, based on text and Table 2 in Kleinstreuer et al. 2017: If inconclusive scores were considered positive, the AR pathway model had a balanced accuracy of 95.2% (**100% sensitivity (8/8) and 90.5% specificity (19/21)**) against the agonist reference substances, and if CoMPARA inconclusive results were excluded, the balanced accuracy was 97.5% (**100% sensitivity (8/8) and 95% specificity (19/20)**).

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 'Multiple model', see 4.4, that integrates a LPDM composite model consisting of 10 sub-models, using all the positives (43 chemicals) in each of these and different subsets of the negatives (1616 chemicals) (see 4.4). 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. 43 positives and 1616 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 2 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.

4.7 Descriptors/chemicals ratio

As this model is a composite model consisting of 10 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
COMPARA2_Multiple_Scaffolds_Model-1	205	119	1
COMPARA2_Multiple_Scaffolds_Model-2	205	154	3
COMPARA2_Multiple_Scaffolds_Model-3	205	118	1
COMPARA2_Multiple_Scaffolds_Model-4	205	173	3
COMPARA2_Multiple_Scaffolds_Model-5	205	175	3
COMPARA2_Multiple_Scaffolds_Model-6	205	118	3
COMPARA2_Multiple_Scaffolds_Model-7	204	172	2
COMPARA2_Multiple_Scaffolds_Model-8	204	167	1
COMPARA2_Multiple_Scaffolds_Model-9	204	119	1
COMPARA2_Multiple_Scaffolds_Model-10	204	153	2

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 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 0.6 and negative predictions with probability equal to or less than 0.3 are accepted. Predictions within the structural applicability domain but with probability between 0.3 to 0.6 or 0.5 to 0.6 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.

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.

5.3 Software name and version for applicability domain assessment

LPDM, a component of Leadscope Enterprise version 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

Yes, from Mansouri et al. 2020 here: https://ehp.niehs.nih.gov/doi/full/10.1289/EHP5580

6.2 Available information for the training set

The file contains identifier, structure information in SDF format 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

1659 compounds are in the training set: 43 positives and 1616 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

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:

- Sensitivity (true positives / (true positives + false negatives)): 76.3±16.1%
- Specificity (true negatives / (true negatives + false positives)): 99.3±0.5%
- Balanced Accuracy ((Sensitivity + Specificity) /2): 87.8±8.0%
- Coverage ((In-Domain predictions) / (All predictions)): 84.5±1.9%

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

An external validation set with results using the CoMPARA 11 tests integrated by a network model is not available. Rather, an evaluation set with results from other types of assays and gathered from the literature is available from Mansouri et al. 2020 here: https://ehp.niehs.nih.gov/doi/full/10.1289/EHP5580. Results for the DTU Leadscope model contributed to the CoMPARA project are given here, please note that the Leadscope model reported in this QMRF developed in a newer version of Leadscope (the Leadscope version

used for CoMPARA is v.3.4 and in the model reported here is developed in Leadscope v.3.5, however using the exact same settings).

7.2 Available information for the external validation set

The file contains identifier, structure information in MOL format and activity call for each substance.

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

Yes

7.5 Other information about the validation set

4839 compounds are in the evaluation set: 167 positives and 4672 negatives.

7.6 Experimental design of test set

From Mansouri et al. 2020:

"The evaluation set comprised data extracted from the literature to be used for evaluating the predictive ability of the models (mostly for verification purposes and not to compare models), performed in parallel with the model building efforts."

7.7 Predictivity – Statistics obtained by external validation

According to Mansouri et al. supplementary material S7 (underlying numbers not given), the following statistics were found for the DTU Leadscope v.3.4 model:

- Sensitivity: 77.3%
- Specificity: 97.8%
- Balance Accuracy: 87.5%

7.8 Predictivity – Assessment of the external validation set

From Mansouri et al. 2020:

"*The EPA's NCCT collected and curated PubChem data (64 sources), restructured it, and mapped the bioactivity values to related biological targets. In this effort, we started with ~80,000 experimental values for AR activity, which mapped to about ~11,000 chemicals that we grouped by modality (agonist, antagonist) and hit call (active, inactive). To improve the consistency between the different PubChem entries and to add binding modality, three rules were applied:*

- *In the case of multiple records for a test chemical, a minimum concordance of two out of three assay results was required to assign a positive activity score.*
- *An active agonist or antagonist was considered a binder.*

- *Inactive agonists and antagonists were considered nonbinders.*

The KNIME standardization workflow referenced earlier was applied to the chemical structures (Mansouri et al. 2016a; McEachran et al. 2018). After removing ToxCast™ chemicals (used for the training set), the generated standard InChI codes matched 7,281 chemicals from the CoMPARA list (prediction set). This list of 7,281 chemicals, with associated data extracted from the literature, was used as the evaluation set. The removed ToxCast™ chemicals were mostly associated with ToxCast™ data only. The evaluation set chemicals were split into three data sets based on the available experimental data. The resulting lists included 4,839 structures for agonist, 4,040 for antagonist, and 3,882 for binding"

7.9 Comments on the external validation of the model

A true external validation was not performed, but rather an evaluation using other types of data was used so the results from the robust leave-many-out cross-validation may be more representative of the predictivity of the model.

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.

AR activation is a mechanistic endpoint related to a number of health outcomes.

8.3 Other information about the mechanistic interpretation

None

9. Miscellaneous information

9.1 Comments

The model can be used to predict if a substance is an AR activator (i.e. has an AR activator score equal to or above 0.1) according to the network model based on the 11 AR pathway *in vitro* assays.

9.2 Bibliography

Mansouri K, Kleinstreuer N, Abdelaziz AM, Alberga D, Alves VM, Andersson PL, Andrade CH, Bai F, Balabin I, Ballabio D, Benfenati E, Bhhatarai B, Boyer S, Chen J, Consonni V, Farag S, Fourches D, García-Sosa AT, Gramatica P, Grisoni F, Grulke CM, Hong H, Horvath D, Hu X, Huang R, Jeliazkova N, Li J, Li X, Liu H, Manganelli S, Mangiatordi GF, Maran U, Marcou G, Martin T, Muratov E, Nguyen D-T, Nicolotti O, **Nikolov NG**, Norinder

U, Papa E, Petitjean M, Piir G, Pogodin P, Poroikov V, Qiao X, Richard AM, Roncaglioni A, Ruiz P, Rupakheti C, Sakkiah S, Sangion A, Schramm K-W, Selvaraj C, Shah I, Sild S, Sun L, Taboureau O, Tang Y, Tetko IV, Todeschini R, Tong W, Trisciuzzi D, Tropsha A, Van Den Driessche G, Varnek A, Wang Z, **Wedebye EB**, Williams AJ, Xie H, Zakharov AV, Zheng Z, Judson RS (2020), CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity, *Environmental Health Perspectives*, 128(2): 0270021-17. doi: 10.1289/EHP5580.

Kleinstreuer NC, Ceger P, Watt ED, Martin M, Houck K, Browne P, Thomas RS, Casey WM, Dix DJ, Allen D, Sakamuru S, Xia M, Huang R, Judson R (2017), Development and validation of a computational model for androgen receptor activity. *Chem Res Toxicol*, 30(4): 946–964, PMID:27933809. doi: 10.1021/acs.chemrestox.6b00347.

Lynch C, Sakamuru S, Huang R, Stavreva D, Varticovski L, Hager G, Judson R, Houck K, Kleinstreuer NC, Casey WM, Paules RS, Simeonov A, Xia M. (2016) Identifying Environmental Chemicals as Agonists of the Androgen Receptor by Applying a Quantitative High-throughput Screening Platform. *Toxicology*, 385: 48–58. doi: 10.1016/j.tox.2017.05.001.