Leadscope Enterprise model for activation of mammalian Estrogen Receptor (ER) *in vitro* (U.S. EPA CERAPP data)

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for activation of mammalian Estrogen Receptor (ER) activation *in vitro* (U.S. EPA CERAPP data), 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

Sine Abildgaard Rosenberg
National Food Institute at the Technical University of Denmark when the version for CERAPP was developed in LPDM v.3.1.1-10

2.3 Date of QMRF update(s)

None

2.4 QMRF update(s)

None

2.5 Model developer(s) and contact details

Sine Abildgaard Rosenberg
National Food Institute at the Technical University of Denmark when the version for CERAPP was developed in LPDM v.3.1.1-10

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

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

2.6 Date of model development and/or publication

Development of CERAPP version in LPDM v.3.1.1-10 finalized in June 2014 and the version developed in LPDM v.3.5 reported here is published in 2020

2.7 Reference(s) to 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 originally made for the U.S. EPA CERAPP 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

Bovine, mouse and human cell lines (18 biochemical and cell-based *in vitro* assays, of which 16 based on human cell lines, 1 based on bovine 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 (Estrogen Receptor)

3.3 Comment on endpoint

There is increasing evidence that a variety of environmental substances have the potential to disrupt the endocrine system by mimicking or inhibiting endogenous hormones such as estrogens and androgens. These "endocrine disrupting chemicals" (EDCs) may adversely affect development and/or reproductive function.

Endogenous estrogens are involved in the development and adult function of organs of the female genital tract, neuroendocrine tissues and the mammary glands; their role in reproduction spans from maintenance of the menstrual cycle to pregnancy and lactation. These effects are primarily mediated through the estrogen receptors (ERs), members of the nuclear receptor superfamily. When estrogen binds to the ER in the cytoplasm a receptor-hormone complex dimer is formed. This dimer translocates to the nucleus, where it recruits co-factors to form the active transcription factor (TF) complex. The active TF binds to the estrogen response element upstream to the target gene. This binding activates transcription of mRNA and subsequent translation to proteins that exert the hormone effects. Two isoforms of the ER exists in humans, alpha and beta, and both are widely expressed in different tissue types although there are some differences in their expression pattern.

Exogenous substances able to bind to and activate the ERs (i.e. ER agonists) have the ability to mimic natural estrogens and cause adverse effects to the reproductive system. Likewise, exogenous substances that bind to the ERs without subsequent activation (i.e. ER inhibitors) can potentially disturb the effect of the endogenous estrogens by blocking the receptors.

Results from 18 *in vitro* high-throughput screening assays that probe the ER signaling pathway in a mammalian system were integrated in a computational network model (Judson et al. 2014). The assays were a combination of biochemical and cell-based *in vitro* assays and probe perturbations of the ER pathway at multiple sites: receptor binding, receptor dimerization, DNA binding of the active transcription factor, gene transcription and changes in ER-induced cell growth kinetics. The network model uses activity patterns across the 18 *in vitro* assays to predict whether the substance is an ER activator, an ER inhibitor, or instead is causing activity through narrow (technology-specific) or broad assay interference. For example, if a substance is active in all of the assays in the ER activation pathway of the network model a score for activation is calculated as the AUC for the accumulated Hill model (based on the AC50 from the assays). If none or only parts of the assays in the ER activation pathway are active, the substance is a clear negative or is causing some form of assay interference (narrow or broad depending on which assays in the pathway that are active), respectively. These substances have an ER activation score of 0 and are all assumed to be negative (Judson et al. 2014).

In order to make a classification model, substances with an ER activation score of 0 were defined as inactives and substances with an AUC score of 0.1 or above were defined as an ER activator.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives

3.5 Dependent variable

Mammalian Estrogen Receptor activation: positive or negative

3.6 Experimental protocol

See S1, Appendix 1 in Judson *et* al. 2015

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 substances have been tested in the same laboratory and the process of assigning an ER activation score using the network model (see 3.2) has been equal for all substances. However, we do not have measures of reproducibility of the results.

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 (80 substances) in each of these and different subsets of the negatives (1340 substances) (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. 80 positives and 1340 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.

4.7 Descriptors/chemicals ratio

The model system uses molecular descriptors and structural features specific to a group of structurally related substances from the global training set. Therefore estimations of the number of used descriptors may be difficult. In general, we estimate that the models effectively use an order of magnitude less descriptors than numbers of substances in the training set when we set our domain definition where we weed out low probability active and inactive predictions (see 5.1).

Name of the model	Substances	Descriptors	PLS factors
CERAPP_35_0.1_0_Multiple_ Model-1	214	157	3
CERAPP_35_0.1_0_Multiple_ Model-2	214	159	1
CERAPP_35_0.1_0_Multiple_ Model-3	214	115	1
CERAPP_35_0.1_0_Multiple_ Model-4	214	162	2
CERAPP_35_0.1_0_Multiple_ Model-5	215	150	1
CERAPP_35_0.1_0_Multiple_ Model-6	215	133	2
CERAPP_35_0.1_0_Multiple_ Model-7	214	115	2
CERAPP_35_0.1_0_Multiple_ Model-8	214	157	3
CERAPP_35_0.1_0_Multiple_ Model-9	213	116	3
CERAPP_35_0.1_0_Multiple_ Model-10	213	132	3
CERAPP_35_0.1_0_1_Model 1.0	1420	142	1

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 substance is within the structural applicability domain of a given model, LPDM examines whether the test substance bears enough structural resemblance to the training set substances used for building the model (i.e. a structural domain analysis). This is done by calculating the distance between the test substance and all substances 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 substances in the training set that have a distance equal to or smaller than 0.7 with respect to the test substance. The higher the number of neighbours, the more reliable the prediction for the test substance. Statistics of the distances are also calculated. Furthermore, LPDM requires

that the test substance contains at least one model feature or scaffold from the model. Effectively no predictions are made for test substances 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

LPDM does not generate predictions for test substances 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

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 substances 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 analyzed 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 substances 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

6.2 Available information for the training set

SMILES

6.3 Data for each descriptor variable for the training set

No

6.4 Data for the dependent variable for the training set

All

6.5 Other information about the training set

1420 substances are in the training set: 80 positives and 1340 negatives.

6.6 Pre-processing of data before modeling

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

The original CERAPP model developed in LPDM v.3.1.1-10 underwent a five times two-fold cross-validation. This was done by randomly removing 50% of the full training set used to make the "mother model", where the 50% contains the same ratio of positive and negatives as the full training set. A new model (validation sub-model) was created on the remaining 50% using the same settings in LPDM but with no information from the "mother model" regarding descriptor selection etc. The validation sub-model was applied to predict the removed 50% (within the defined applicability domain). Likewise, a validation sub-model was made on the removed 50% of the training set and this model was used to predict the other 50% (within the defined applicability domain). This was repeated five times.

Predictions from the ten sub-models were pooled and Coopers statistics were calculated. This gave the following results for the predictions which were within the applicability domains of the respective sub-models:

- Sensitivity (100% * true positives / (true positives + false negatives)): 100%*270/(270+65) = 80.60%
- Specificity (100% * true negatives / (true negatives + false positives)): 100%*4650/(4650+278) = 94.36%
- Balanced accuracy ((sensitivity + specificity) / 2): (80.6% + 94.36%)/2 = 87.5%
- Coverage (100% * in-domain predictions / all predictions): 100% * 5263 / (5 * 1420) = 74.1%

The final version of the CERAPP activation model, reported in this QMRF, developed in LPDM v.3.5 underwent a two times five-fold (i.e. 20 % out in total 10 times) cross-validation by DTU Food cross-validation procedure.

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 (100% * true positives / (true positives + false negatives)): 78.5± 10.3%
- Specificity (100% * true negatives / (true negatives + false positives)): 96.7±1.1%
- Balanced accuracy ((sensitivity + specificity) / 2): 87.6± 5.3%
- Coverage (100% * in-domain predictions / all predictions): 77.3±2.5%

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 training set

Results for the DTU LPDM model contributed to the CERAPP project reported in Mansouri et al. 2016 are given in this section (LPDM v. 3.1.1-10 is used for CERAPP is).

Please note that the LPDM model reported in this QMRF is developed in a newer version of LPDM (the LPDM v.3.5 and an optimized DTU modeling approach is used for the DTU model reported in this QMRF).

An "external validation" set with results using the CERAPP 18 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 was done in Mansouri et al. 2016. The evaluation set is available from the U.S. EPA CERAPP site (URL not given here as it has changed since the publication of Mansouri et al. 2016 publication and may change also in the future, so it is better to search the site given at any point in time);

	Sensitivity	Specificity
ToxCast	0.875	0.904
All literature	0.673	0.898
multSrc Literature	0.738	0.898
no_VW literature	0.689	0.898
in_AD Literature	0.678	0.932
allPar Literature	0.764	0.932

7.2 Available information for the external training 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 training set

No

7.4 Data for the dependent variable for the external training set

Yes

7.5 Other information about the training set

The evaluation set comprised 6,319 substances, of which 350 were positive and 5,969 were negative. Of these, 84.14% were in AD according to Mansouri et al. 2016 Table S3, however the numbers for TP, TN, FP and FN are not given.

7.6 Experimental design of test set

From Mansouri et al. 2016:

"*Experimental evaluation set. A large volume of estrogen-related experimental data has accumulated in the literature over the past two decades. The information on the estrogenic activity of chemicals was mined and curated to serve as a validation set for predictions of the different models. For this purpose, in vitro experimental data were collected from different overlapping sources, including the U.S. EPA's HTS assays, online databases, and other data sets used by participants to train models:*

- *HTS data from Tox21 project consisting of ~ 8,000 chemicals evaluated in four assays (Attene-Ramos et al. 2013; Collins et al. 2008; Huang et al. 2014; Shukla et al. 2010; Tice et al. 2013), extending beyond the 1,677 used in the training set.*
- *The U.S. FDA Estrogenic Activity Database (EADB), which consists of literature derived ER data for ~ 8,000 chemicals (Shen et al. 2013).*
- *Estrogenic data for ~ 2,000 chemicals from the METI (Ministry of Economy, Trade and Industry, Japan) database (METI 2002).*
- *Estrogenic data for ~ 2,000 chemicals from ChEMBL database (Gaulton et al. 2012).*

The full data set consisted of > 60,000 entries, including binding, agonist, and antagonist information for ~ 15,000 unique chemical structures. For the purpose of this project, this data set was cleaned and made more consistent by removing in vivo data, cytotoxicity information, and all ambiguous entries (missing values, undefined/nonstandard end points, and unclear units). Only 7,547 chemical structures from the experimental evaluation set that overlapped with the CERAPP prediction set, for a total of 44,641 entries, were kept and made available for download on the U.S. EPA ToxCast™ Data web site (https://www3.epa.gov/research/COMPTOX/CERAPP_files.html, EvaluationSet.zip) (U.S. EPA 2016). The non-CERAPP chemicals were excluded from the evaluation set (see "Chemical Structure Curation" section). Then, all data entries were categorized into three assay classes: (a) binding, (b) reporter gene/transactivation, or (c) cell proliferation. The training set end point to model is the ER model AUC that parallels the corresponding individual assay AC_{50} values, and therefore all units for activities in the experimental data set were converted to μM to have approximately equivalent concentration–response values for the evaluation set. Chemicals with cell proliferation assays were considered as actives if they exceeded an arbitrary threshold of 125% proliferation. For entries where testing concentrations were reported in the assay name field, those values were converted to μM and considered as the AC_{50} value if the compound was reported as active. All inactive compounds were arbitrarily assigned an AC_{50} value of 1 M."

7.7 Predictivity – Statistics obtained by external validation

According to Mansouri et al. supplementary material S3 (Sensitivity taken from "SN allPar Literature", and specificity taken from "SP SllPar Literature", numbers for TP, TN, FP and FN not given), the following statistics were found for the original DTU LPDM v.3.1.1-10 CERAPP model:

- Sensitivity: 76.4%
- Specificity: 93.2%
- BA: 84.8%

7.8 Predictivity – Assessment of the external validation set

From Mansouri et al. 2016:

"*The performance of most models showed a clear improvement of 0.05 to 0.1 on the BA after applying all the filters on the literature data to keep only the unambiguous chemicals. We believe that this effectively reduced the uncertainty of the literature sources. This step also highlighted differences between ToxCast™ and the literature data and confirmed the existence of uncertainty in the literature data. Uncertainty and data discordance was also reported in literature review of in vivo uterotrophic bioassays (Kleinstreuer et al. 2015)*."

7.9 Comments on the external validation of the model

As noted above, the results reported in section 7 origin from the evaluation (not validation) of the DTU original CERAPP LPDM model developed in version 3.1.1-10. External validation was not performed on the LPDM v.3.5 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

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

ER 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 ER activator (i.e. has an ER activation score equal to or above 0.1) according to the network model based on the 18 ER pathway *in vitro* assays.

9.2 Bibliography

Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, Xia M, Huang R, Rotroff DM, Filer DL, Houck KA, Martin MT, Sipes N, Richard AM, Mansouri K, Setzer RW, Knudsen TB, Crofton KM, Thomas RS. (2015) Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro Thigh-Throughput Screening Assays for the Estrogen Receptor. *Toxicol.Sci.,* 148, 137-154. doi: 10.1093/toxsci/kfv168

Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL, Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, Judson RS (2016). CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environmental Health Perspectives 124, 7, 1023-1033, free access at http://dx.doi.org/10.1289/ehp.1510267.