	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **QSARs for predicting up regulation of pregnane X receptor (PXR) as MIEs of hepatic steatosis**
	Printing Date: **June 1st, 2022**

1.QSAR identifier

1.1.QSAR identifier (title):

[1] BRF for predicting up regulation of pregnane X receptor (PXR)

1.2.Other related models:

[2] BRF RF for predicting down regulation of pregnane X receptor (PXR)

[3] BRF for predicting up regulation of liver X receptor (LXR)

[4] BRF for predicting down regulation of liver X receptor (LXR)

[5] BRF for predicting up regulation of Aryl hydrocarbon receptor (AhR)

[6] BRF for predicting down regulation of Aryl hydrocarbon receptor (AhR)

[7] BRF for predicting up regulation of Nuclear factor (erythroid-derived 2)-like 2 (Nrf2)

[8] BRF for predicting down regulation of Peroxisome proliferator-activated receptors alpha (PPARα)

[9] BRF for predicting down regulation of Peroxisome proliferator-activated receptors gamma (PPARγ)

1.3.Software coding the model:

randomForest (R package) (v4.6-12).

KNIME (v3.4)

2.General information

2.1.Date of QMRF:

1 June 2022

2.2.QMRF author(s) and contact details:

[1] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

[2] Erika Colombo; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; erika.colombo@marionegri.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

[2] Serena Manganelli; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; serena.manganelli@marionegri.it

[3] Cosimo Toma; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; cosimo.toma@marionegri.it

[4] Alessandra Roncaglioni; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; alessandra.roncaglioni@marionegri.it

[5] Emilio Benfenati; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; emilio.benfenati@marionegri.it

[6] Enrico Mombelli; Institut National de l'Environnement Industriel et des Risques (INERIS); enrico.mombelli@ineris.fr

2.6.Date of model development and/or publication:

2018

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

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modeling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modeling, accepted manuscript.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In Data Analysis, Machine Learning and Applications, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319-326.

[4] Genuer, R.; Poggi, J. M.; Tuleau-Malot, C., VSURF: An R Package for Variable Selection Using Random Forests. The R Journal 2015, 7, 19-33.

2.8.Availability of information about the model:

All the information about the model are reported in the reference publication (see 2.7).

2.9.Availability of another QMRF for exactly the same model:

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Human

3.2.Endpoint:

QMRF 6. Other QMRF 6. 6. Other

3.3.Comment on endpoint:

Up regulation of Pregnane X receptor (PXR) [GeneSymbol:NR1I2 | GeneID:8856 | Uniprot_SwissProt_Accession: O75469]

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

Categorical (1 for positive, 0 for negative). pAC50 values greater than zero are actives (1), pAC50 values equal to zero are inactives (0). If at least one of the assays considered for each endpoint were active, the sample was flagged as active.

3.6.Experimental protocol:

Data referred to ToxCast assays ATG_PXR_TRANS_up (AEID: 135) and ATG_PXRE_CIS_up (AEID: 103). Attagene (ATG) assays are cell-based, multiplexed-redout assays that uses HepG2, a human liver cell line, with measurements taken at 24 hour after chemical dosing in 24-well plate.

These assays are designed to make measurements of mRNA induction, a form of inducible reporter, as detected with fluorescence intensity signals by Reverse transcription polymerase chain reaction (RT-PCR) and Capillary electrophoresis technology.

Changes to fluorescence intensity signals are indicative of inducible changes in transcription factor activity. This is quantified by the level of mRNA reporter sequence unique to:
The cis-acting reporter gene response element which are responsive of an endogenous human receptor subfamily (CIS assays);
The transfected trans-acting reporter gene and exogenous transcription factor GAL4, which are responsive of a given human receptor isoform (TRANS assays).
Further info on the assays: http://www.attagene.com/technology.php

3.7.Endpoint data quality and variability:

Experimental data used in this work were isolated from a collection of 24 in vitro HTS assays from the ToxCast program, executed by Attagene Inc. (RTP, NC), under contract to the U.S. EPA (Contract Number EP-W-07-049). During this program, several experiments evaluated the impact of more than 8,000 chemicals on the previously described TFs involved in the MIE of steatosis AOP.
For approximately half the chemicals tested during the ToxCast project, cytotoxicity was observed in the range of concentrations tested. Thus, a significant proportion of measured activities may represent a false positive response caused by assay interference process linked to a cytotoxicity-related 'burst' of activities (Judson et al., 2006, Toxicol. Sci. 2016, 152, 323-339).
The presence of possible false negatives was also reported. The volatility of particular chemical categories (e.g., solvent chemicals) included in ToxCast or the low solubility may explain their general lack of significant effect. For data curation see section 6.6 of QMRF

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Consensus of four single models based on 1) Random Forest (RF) and Balanced Random Forest (BRF)

4.2.Explicit algorithm:

Consensus was the combination of four different QSAR models. Two different algorithms were applied, with and without a prior feature selection (FS):

1) Balanced Random Forest (BRF) is a combination of under-sampling and the ensemble idea. This technique artificially alters the class distribution so that classes are represented equally in each tree. The randomForest R package (version 4.6-12) was used for the BRF approach. The *mtry* value was the one provided by default in R.
2) Random Forest was derived based on undersampling of the training set, i.e. random deletion of the most represented class (i.e. negative chemicals) until both classes were equal in number. RF implemented in KNIME was used to derive the undersampling based model. The mtry value was the one provided by default in KNIME.

The number of trees was selected in the range 25-251, based on the lowest prediction error returned in 10-fold internal cross validation. Table indicates the number of trees for each model:

BRF		Undersampling	
w/o FS	w/ FS	w/ FS	w/ FS
501	501	101	101

4.3.Descriptors in the model:

BRFw/oFS: 1095

US w/o FS: 313

BRF w/ FS: ChiA_Dz(Z), Eta_alpha_A, GGI3, MAXDN, MLOGP2, P_VSA_ppp_D, SdssC, P_VSA_ppp_L, SpMax_B(s), SpMax4_Bh(m), SpMax_B(m), SpMin2_Bh(m), SpPosA_B(v), Wap, X5v

US w/ FS: ChiA_Dz(Z), Eta_alpha_A, GGI3, MAXDN, MLOGP2, P_VSA_ppp_D, SdssC, P_VSA_ppp_L, SpMax_B(s), SpMax4_Bh(m), SpMax_B(m), SpMin2_Bh(m), SpPosA_B(v), Wap, X5v

4.4.Descriptor selection:

Use of VSURF (https://cran.r-project.org/web/packages/VSURF/VSURF.pdf) R package.

4.5.Algorithm and descriptor generation:

Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterized by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed. RF automatically identified descriptors most relevant for describing the endpoint.

For two of the four derived models, optimal subsets of descriptors for modeling were obtained with the R package VSURF. The algorithm consists in a three step variable selection based on the logic underpinning the random forest (RF) algorithm (i.e. permutation importance and out-of-bag error). The first step eliminates irrelevant descriptors according to the permutation-based RF score of importance and a user-defined threshold. The second step finds important descriptors closely related to the response variable (interpretation step) and the third step (prediction step) identifies a sufficient parsimonious set of important descriptors leading to a good prediction of the response variables. The VSURF selection procedure was carried out as a function of a number of trees ranging from 25 to 251, then the pool of descriptors returning the lowest internal error was retained.

4.6.Software name and version for descriptor generation:

Dragon v7.0.8

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kodesolutions.net https://chm.kode-solutions.net/products_dragon.php

4.7.Chemicals/Descriptors ratio:

Not relevant for Random Forests

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets.

5.2.Method used to assess the applicability domain:

The Applicability domain chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [6]. The VEGA AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency

between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments.

5.3.Software name and version for applicability domain assessment:

KNIME (Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization); Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de https://www.knime.com/

5.4.Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralized form.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3.Data for each descriptor variable for the training set:

Not available

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

The dataset described in [1] was randomly divided into a training set (TS, 80% of the original dataset) and a validation set (VS, 20% of the original dataset) comprising the same proportion of active and inactive chemicals as the original dataset. The table below reports the number of chemicals in TS and VS according to the original model:

TF	Inactives (full database)	Actives (full database)	% Actives	TS	VS
PXR_up	529	640	55	934	235

Instead, the dataset implemented in VEGA was split in training (853 chemicals) and test (216 chemicals)

6.6. Pre-processing of data before modelling:

Data were retrieved from the oldstyle_neg_log_ac50_Matrix_151020.csv file, downloaded from ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Summary_Files.

For classification purposes, for each assay chemicals with zero values for a given assay were considered as inactive, while chemicals with a continuous pAC50 value were considered active. Results from TRANS-assays were considered if specific isoforms of a TF listed among the AOPs for steatosis (e.g., PPARα, PPARγ), while CIS-assays or a combination of CIS- and TRANS-assays were used if TF isoforms were not specified. In the latter case, CIS- and TRANS- outputs were combined. A chemical was labeled as active if it was active in at least one assay and inactive if it was inactive in both types of assay.

Only chemicals exceeding 90% purity were retained, while chemicals associated to lower purity, other anomalies (e.g. withdrawn chemicals) or not yet analyzed were not included.
The structures were checked by removing inorganic chemicals and mixtures, correcting inaccurate SMILES codes with the help of chemical databases, i.e. ChemSpider and ChemIDplus and neutralizing salts.
An in-house software was used to identify and remove duplicates. For a given set of duplicated structures, if their experimental activities were identical, then only one compound was kept. If their experimental properties were different, both the chemicals were removed.
A Python script executing the MolVS standardizer (based on RDkit libraries) was written to obtain canonical tautomers. Canonical SMILES were coded using the istMolBase software based on CDK libraries.
A z-score (Eq. 1) that was assigned to each chemical-assay combination (Judson et al., 2006, Toxicol. Sci. 2016, 152, 323-339):

$$Z(chemical, assay) = \frac{-logAC50(chemical, assay) - median[-logAC50(chemical, cytotoxicity)]}{global\ cytotoxicity\ MAD} \quad (1)$$

Conversely, chemicals associated with low z-scores are more likely to be false positives confounded by cytotoxicity. A z-score threshold of 3 was considered to select only chemicals that can be considered as specifically active.

6.7.Statistics for goodness-of-fit:

After the implementation in VEGA:
Training set: n = 934, Balance Accuracy 0.99, Sensitivity 0.99, Specificity 1, MCC 0.99. TP509, TN 422, FP 0, FN 3

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

According to the original model developed in R and KNIME:
Method: 10-fold cross-validation. Results are reported above:

	BRF				US			
	w/o FS		w/ FS		w/o FS		w/ FS	
	all	AD	all	AD	all	AD	all	AD
#	934	598	934	581	934	709	934	747
P	512	341	512	322	512	392	512	407
N	422	257	422	259	422	317	422	340
ACC	0.73	0.80	0.74	0.81	0.73	0.79	0.74	0.78
SE	0.81	0.88	0.81	0.88	0.80	0.86	0.79	0.85
SP	0.64	0.70	0.65	0.71	0.64	0.70	0.68	0.69
MCC	0.46	0.59	0.47	0.61	0.45	0.58	0.47	0.55
BA	0.73	0.79	0.73	0.80	0.72	0.78	0.73	0.77
AUC	0.79	0.84	0.80	0.85	0.79	0.82	0.80	0.82
%	1.00	0.64	1.00	0.62	1.00	0.76	1.00	0.80

6.10.Robustness - Statistics obtained by Y-scrambling:

According to the original model developed in R and KNIME:

	BRF		Undersampling	
	w/o FS	w/ FS	w/o FS	w/ FS
MCC	0.00	0.00	0.00	0.00

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

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 external validation set:

See 6.5

7.6.Experimental design of test set:

See 6.5

7.7.Predictivity - Statistics obtained by external validation:

According to the original model developed in R and KNIME:

	BRF				US			
	w/o FS		w/ FS		w/o FS		w/ FS	
	all	AD	all	AD	all	AD	all	AD
#	235	151	235	144	235	171	235	177
P	128	92	128	84	128	99	128	102
N	107	59	107	60	107	72	107	75
ACC	0.77	0.87	0.77	0.85	0.74	0.82	0.78	0.84
SE	0.88	0.93	0.84	0.94	0.84	0.90	0.87	0.92
SP	0.64	0.76	0.68	0.73	0.62	0.71	0.67	0.73
MCC	0.53	0.72	0.54	0.70	0.47	0.63	0.55	0.68
BA	0.76	0.85	0.76	0.84	0.73	0.80	0.77	0.83
AUC	0.85	0.87	0.83	0.85	0.83	0.87	0.84	0.86
%	1.00	0.64	1.00	0.61	1.00	0.73	1.00	0.75

After the implementation in VEGA:

Test set: n = 235, Balanced Accuracy 0.76, Sensitivity 0.84, Specificity 0.67, MCC 0.53. TP 108, TN 72, FP 35, FN 20

Test set in AD: n = 134, Balanced Accuracy 0.80, Sensitivity 0.87, Specificity 0.73, MMC 0.60. TP 65, TN 43, FP 16, FN 10

Test set could be out of AD: n = 62, Balanced Accuracy 0.76, Sensitivity 0.84, Specificity 0.67, MCC 0.52. TP 27, TN 20, FP 10, FN 5

Test set out of AD: n = 39, Balanced Accuracy 0.63, Sensitivity 0.76, Specificity 0.50, MCC 0.27. TP 16, TN 9, FP 9, FN 5

7.8.Predictivity - Assessment of the external validation set:

According to the original model developed in R and KNIME:

MCC values were lower in external validation with respect of internal validation for endpoints with highly unbalanced datasets for chemicals in the AD. A possible explanation for this poor performance, can be found in the extreme degree of imbalance of some VS (i.e. less than 10% of active chemicals) that seriously undermines the reliability of statistical indicators.

Statistical analyses were done to identify critical MCC thresholds for reliably evaluating the performance of models on binary datasets with different degree of imbalance. These thresholds correspond to a reasonable minimum predictivity and were defined for each model by imposing a minimum percentage of correctly predicted positive and negative chemicals of 75% (i.e. SE = SP= 75%). Results demonstrated that, given the same percentage of correctly predicted active and inactive compounds, very unbalanced datasets are linked to lower MCC values. Table below shows which models overcome predictivity thresholds with respect of the degree of unbalance of datasets.

	BRF		Undersampling	
	w/o FS	w/ FS	w/o FS	w/ FS
#	598	581	709	747
P	341	322	392	407
N	257	259	317	340
MCC	0.59	0.61	0.58	0.55
MCC75	0.50	0.50	0.50	0.50
Valid?	Y	Y	Y	Y
#	151	144	171	177
P	92	84	99	102
N	59	60	72	75
MCC	0.72	0.70	0.63	0.68
MCC75	0.49	0.49	0.49	0.50
Valid?	Y	Y	Y	Y

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Not provided

8.2.A priori or a posteriori mechanistic interpretation:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modeling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modeling, *submitted manuscript*.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Romanov, S., Medvedev, A., Gambarian, M., Poltoratskaya, N., Moeser, M., Medvedeva, L., ... & Makarov, S. (2008). Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. Nature Methods, 5(3), 253.
[4] Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., ... & Gambarian, M. (2010). Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. Chemical research in toxicology, 23(3), 578-590.
[5] Liaw, A.; Wiener, M., Classification and Regression by RandomForest. R News 2002, 2, 18-22.
[6] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39. doi: 10.1186/s13321-014-0039-1. PMID: 25383097; PMCID: PMC4212147

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

Available at https://drive.google.com/open?id=1-pS273HYIVE2jz2BSPMIqam3EjcTf-3x