

	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: InVitroHCL_HumanMic_AliphaticHydroxylation_d_rank2
	Printing Date: 17-set-2021

1. QSAR identifier

1.1. QSAR identifier (title):

InVitroHCL_HumanMic_AliphaticHydroxylation_d_rank2

1.2. Other related models:

None

1.3. Software coding the model:

QSARINS

Software for QSAR MLR models development

nicola.chirico@uninsubria.it; paola.gramatica@uninsubria.it; ester.papa@uninsubria.it

<http://dunant.dista.uninsubria.it/qsar/>

2. General information

2.1. Date of QMRF:

11/10/2020

2.2. QMRF author(s) and contact details:

Ilaria Casartelli University of Insubria Ilaria Casartelli; Linda Bertato; Ester Papa

icasartelli@studenti.uninsubria.it; l.bertato@uninsubria.it; ester.papa@uninsubria.it

<http://dunant.dista.uninsubria.it/qsar/>

2.3. Date of QMRF update(s):

-

2.4. QMRF update(s):

-

2.5. Model developer(s) and contact details:

Linda Bertato; Ester Papa University of Insubria Linda Bertato; Ester Papa l.bertato@uninsubria.it;

ester.papa@uninsubria.it <http://dunant.dista.uninsubria.it/qsar/>

2.6. Date of model development and/or publication:

2018-2020

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

[1]Toxtree: toxic hazard estimation software (Module SMARTCyp - Cytochrome P450-Mediated Drug Metabolism and metabolites prediction) <https://sourceforge.net/projects/toxtree/>

[2]QSARINS: A new software for the development, analysis, and validation of QSAR MLR models <http://dunant.dista.uninsubria.it/qsar/>

[3]QSARINS-Chem standalone version software: Insubria datasets and models <http://dunant.dista.uninsubria.it/qsar/>

[4]PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints v 2.21 <http://www.yapcsoft.com/dd/padeldescriptor/>

2.8. Availability of information about the model:

Model developed as output of the project CEFIC LRI-ECO 44 Available in

QSARINS-Chem (<http://dunant.dista.uninsubria.it/qsar/>)

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

no

3. Defining the endpoint - OECD Principle 1

3.1. Species:

Human

3.2. Endpoint:

QMRF 5. Toxicokinetics QMRF 5. 8. Toxicokinetics. Metabolism (including metabolic clearance)

3.3. Comment on endpoint:

This QSAR has been developed to model the *in vitro* intrinsic hepatic clearance quantified in human microsomes. The *in vitro* intrinsic clearance represents the liver's ability to transform the substance independently of the blood flow and the availability of the substance. The *in vitro* intrinsic clearance is defined as the ratio between the maximum speed of the reaction and the Michaelis constant.

3.4. Endpoint units:

mL/h/mg proteins

3.5. Dependent variable:

Log_CL in vitro, int

3.6. Experimental protocol:

Reference Protocol OECD 319 A and B *in vitro* guidance documents.

3.7. Endpoint data quality and variability:

Data curation has been performed as a task of the CEFIC-LRI ECO44 Project by evaluating consistency of information reported in literature and coherence with the OECD 319 A and B *in vitro* guidance documents.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Multiple Linear Regression by means of Ordinary Least Squares

4.2. Explicit algorithm:

MLR - QSAR model

Multiple Linear Regression by means of Ordinary Least Squares

$$\text{Log_CL in vitro, int} = 0.32^* (\pm 0.26) + 0.57^{***} (\pm 0.23) \text{ PubchemFP377} + 0.15^{***} (\pm 0.06) \text{ maxHBint4} + 0.27^{***} (\pm 0.13) \text{ minHBint5} + 0.44^{**} (\pm 0.24) \text{ PubchemFP805} - 0.93^{***} (\pm 0.49) \text{ minHBd} - 0.37^{**} (\pm 0.23) \text{ PubchemFP143} + 0.04^{***} (\pm 0.02) \text{ C2SP2} - 0.04^{***} (\pm 0.02) \text{ SHBint7} + 0.38^{**} (\pm 0.03) \text{ mindssC}$$

Significance (P values): ^{***}, 0.001; ^{**}, 0.01; ^{*}, 0.0

4.3.Descriptors in the model:

[1]PubchemFP377 C(~N)(:C)(:C)

[2]maxHBint4 Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 4

[3]minHBint5 Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 5

[4]PubchemFP805 OC1CC(N)CCC1

[5]minHBd Minimum E-States for (strong) Hydrogen Bond donors

[6]PubchemFP143 ≥ 1 any ring size 5

[7]C2SP2 Doubly bound carbon bound to two other carbons

[8]SHBint7 Sum of E-State descriptors of strength for potential hydrogen bonds of path length 7

[9]mindssC Minimum atom-type E-State: =C<

4.4.Descriptor selection:

An input file including more than 700 molecular descriptors of different types (0D, 1D, 2D) were calculated in PaDEL-Descriptor v. 2.21.

Constant, semi-constant and highly correlated descriptors were excluded in a pre-reduction step. Models were initially developed by the all-subset procedure up to 2 variables, then model's complexity was increased using a Genetic Algorithm (GA) based selection procedure. The cost function used by the GA was Q2LOO (leave-one-out).

4.5.Algorithm and descriptor generation:

Molecular descriptors were calculated using the software Padel-Descriptor v. 2.21 using canonicalized SMILES as input. SMILES were canonicalized using the software OpenBabel v. 2.3.2.

4.6.Software name and version for descriptor generation:

Padel Descriptor v. 2.21

Software to Calculate Molecular Descriptors and Fingerprints

-

<http://www.yapcsoft.com/dd/padeldescriptor/>

Open Babel v. 2.3.2

Open Babel: The Open Source Chemistry Toolbox

-

<http://openbabel.org>

4.7.Chemicals/Descriptors ratio:

10.1

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Statistical AD:

The applicability domain of the model was verified by the leverage approach and fixed thresholds has been used to define both structural and response outliers (see section 5.4). The plot of leverages (HAT diagonals) versus standardised residuals, i.e. the Williams plot, verified the presence of response outliers (i.e. compounds with

cross-validated standardized residuals greater than 2.5 standard deviation units) and chemicals structurally very influential in determining the model's coefficients (i.e. compounds with a leverage value (h) greater than $3p'/n$ (h^*), where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value ($h > h^*$), which are structural outliers, predictions should be considered less reliable.

Mechanistic AD: The applicability domain of the model is related to the most probable site of reaction and the related reactivity, identified by the Toxtree module SMARTCyp.

5.2. Method used to assess the applicability domain:

The structural applicability domain of the model was assessed by the leverage approach, on the bases of a cut-off hat value $h^*=0.33$. HAT values for each compound are calculated as the diagonal elements of the HAT matrix ($H = X(X^T X)^{-1} X^T$).

The response applicability domain can be verified by the standardized residuals (cut off values 2.5 standard units).

5.3. Software name and version for applicability domain assessment:

QSARINS

Software for QSAR MLR models development

nicola.chirico@uninsubria.it; paola.gramatica@uninsubria.it

<http://dunant.dista.uninsubria.it/qsar/>

Module SMARTCyp - Cytochrome P450-Mediated Drug Metabolism and metabolites prediction

Module included in the software Toxtree: toxic hazard estimation software

-

<https://sourceforge.net/projects/toxtree/>

5.4. Limits of applicability:

HAT i/i ($h^*=0.33$)

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

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

All

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

All

6.5.Other information about the training set:

91 chemicals were included in the training set.

6.6.Pre-processing of data before modelling:

The endpoint was log transformed prior to modelling.

6.7.Statistics for goodness-of-fit:

R^2 : 0.75 RMSE_{TR}:0.32

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

Q^2_{LOO} : 0.69

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

Q^2_{LMO} (30%): 0.68

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2_{YSCR} : 0.10

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: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3.Data for each descriptor variable for the external validation set:

All

7.4.Data for the dependent variable for the external validation set:

All

7.5.Other information about the external validation set:

To verify the predictive capability of the proposed model, the dataset was split, before model development, into a training set used for model development and a prediction set used for external validation.

7.6.Experimental design of test set:

One every three chemicals were included in the prediction set. Chemicals were sorted by response using the automatic procedure available in QSARINS, the first and last chemicals were selected as training. Chemicals were selected on the basis of response order using the automatic procedure available in QSARINS. (29 chemicals in the prediction set)

7.7.Predictivity - Statistics obtained by external validation:

Q^2_{ext} F3: 0.58 RMSE: 0.41 R^2 : 0.58

7.8. Predictivity - Assessment of the external validation set:

The splitting performed in the software QSARINS allowed for the selection of meaningful training sets and representative prediction sets. Training and prediction sets are balanced according to both structure and the response.

7.9. Comments on the external validation of the model:

The full model, calibrated on the complete dataset (thus ensuring a wider applicability domain), is implemented in the software QSARINS-Chem for predictive purposes. The model equation used for the external validation (reported also in section 4.2) and the statistics are the following: $\text{Log_CL in vitro, int} = 0.32^* (\pm 0.26) + 0.57^{***} (\pm 0.23) \text{ PubchemFP377} + 0.15^{***} (\pm 0.06) \text{ maxHBint4} + 0.27^{***} (\pm 0.13) \text{ minHBint5} + 0.44^{**} (\pm 0.24) \text{ PubchemFP805} - 0.93^{***} (\pm 0.49) \text{ minHBd} - 0.37^{**} (\pm 0.23) \text{ PubchemFP143} + 0.04^{***} (\pm 0.02) \text{ C2SP2} - 0.04^{***} (\pm 0.02) \text{ SHBint7} + 0.38^{**} (\pm 0.03) \text{ mindssC}$
Significance (P values): *** , 0.001; ** , 0.01; * , 0.05
Domain of applicability: $h^* = 0.33$

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

This model predicts biotransformation for chemicals which have been identified as reactive by Aliphatic Hydroxylation (rank 2) using the SMARTCyp module of the Toxtree software.

The mechanistic basis of this model is defined by the most probable reaction sites identified by SMARTCyp and the related reaction.

8.2. A priori or a posteriori mechanistic interpretation:

a priori and *a posteriori*.

8.3. Other information about the mechanistic interpretation:

The model was developed by statistical selection of the molecular descriptors. The interpretation of these descriptors, listed in section 4.3, is provided *a posteriori*.

9. Miscellaneous information

9.1. Comments:

This model has been developed as output of the project CEFIC-LRI ECO44.

9.2. Bibliography:

9.3. Supporting information:

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

10. Summary (JRC QSAR Model Database)

10.1. QMRF number:

To be entered by JRC

10.2. Publication date:

To be entered by JRC

10.3. Keywords:

To be entered by JRC

10.4. Comments:

To be entered by JRC