QMRF identifier (JRC Inventory): **To be entered by JRC**
QMRF Title: **kMHalf-Life Model version 1.0.0 Arnot/episuite 1.0.0**
Printing Date: **Mar 9, 2020**

1.QSAR identifier

1.1.QSAR identifier (title):

kMHalf-Life Model version 1.0.0 Arnot/episuite 1.0.0

1.2.Other related models:

1.3.Software coding the model:

VEGAHUB

https://www.vegahub.eu/contacts/

https://www.vegahub.eu/

2.General information

2.1.Date of QMRF:

2.2.QMRF author(s) and contact details:

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

2.6.Date of model development and/or publication:

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

Arnot JA, Mackay D, Parkerton TF, Bonnell M., “A database of fish biotransformation rates for organic chemicals.” Environmental Toxicology and Chemistry (2008), 27, 2263-2270.

2.8.Availability of information about the model:

Guide to kM/Half-Life Model version 1.0.0 i VEGAHUB - VEGA

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

The model estimates screening level whole body primary biotransformation
half-lives (HL; log days) and rate constants (kM; /log days) for
discrete organic chemicals in fish, based on the work of Arnot and as
implemented in the BCFBAF module of the Epi Suite software:
Arnot JA, Mackay D, Bonnell M., “Estimating metabolic biotransformation
rates in fish from laboratory data.” Environmental Toxicology and
Chemistry (2008), 27, 341-351.
Arnot JA, Mackay D, Parkerton TF, Bonnell M., “A database of fish
biotransformation rates for organic chemicals.” Environmental Toxicology
and Chemistry (2008), 27, 2263-2270.
The model is based on a dataset of 632 experimental kM biotransformation
rates in fish, and consists of a linear regression based on the LogP
prediction (here calculated with the Meylan LogP model implemented in
VEGA), on the Molecular Weight and on the contribution of a set of
correction fragments.

3.2.Endpoint:

whole body primary biotransformation half-lives (HL; day) and rate constants (kM; /day) for discrete organic chemicals in fish.

3.3.Comment on endpoint:

kM biotransformation rates in fish

3.4.Endpoint units:

biotransformation half-lives (HL; day) and rate constants (kM; /day)

3.5.Dependent variable:

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model is a re-implementation of the original model developed by based on the work of Arnot and as implemented in the BCFBAF module of the Epi Suite software (Arnot et al., 2008 a and b) [REF: a) Arnot JA, Mackay D, Bonnell M., “Estimating metabolic biotransformation rates in fish from laboratory data.” Environmental Toxicology and Chemistry (2008), 27, 341-351; b) Arnot JA, Mackay D, Parkerton TF, Bonnell M., “A database of fish biotransformation rates for organic chemicals.” Environmental Toxicology and Chemistry (2008), 27, 2263-2270.]

4.2.Explicit algorithm:

Whole-body biotransformation rate constants were calculated from the data set using the kinetic mass balance model estimation method. Three estimates of central tendency calculated by this method include a deterministic value for which some negative values are possible, an MC median for which some negative values are also possible, and an MC geometric mean that is calculated from positive kM values only.When all three estimates of central tendency yielded positive results for a given set of experimental data inputs, the average of these three values was used to provide a representative individual value (kM,i). When deterministic or MC median values were negative for a set of data inputs, kM,i was

assumed equal to the adjusted MC geometric mean. Each kM,i value was normalized to a mass- and temperature specific rate constant (kM,N) for a 10-g ?sh at 15 C as 0.25k k (W /W) exp[0.01(T T)] M,N M,i N i N i where WN is the normalized mass of the organism (0.01 kg or 10 g), Wi is the original study-speci?c mass of the organism (kg), TN is the normalized water temperature (15 C), and Ti is the original study-specific water temperature (C). Whole-body biotransformation rate constant values can be converted to other mass- and temperature-specific conditions using the preceding equation. Weight and temperature values used for normalization were selected to represent the approximate median values of these parameters in the database.

Theoretical maximum whole-body kM,MAX values Nichols, Fitzsimmons, and Burkhard estimated maximum kM values based on biotransformation in the liver only as a result of blood ?ow limitations to the liver and protein binding.

The maximum kM values based on hepatic rates ranged from 1.5 (log KOW 4) to 7.5 per day (log KOW 0) for a 1-kg ?sh at 10 C and from 9.5 (log KOW 4) to 48.9 (log KOW 0) per day for a 1-g ?sh at 25 C [11]. The possibility of extrahepatic biotransformation, particularly for phase II pathways, and the uncertainty of protein-binding estimates were also discussed. In vitro studies have shown that enzymatic activity in extrahepatic tissues (kidney, gill, blood, and muscle) can approximate the enzymatic activity of the liver in some cases; however, this is variable and uncertain. Biotransformation rates in vivo depend on tissue specific affinity constants and ?ow rates. The total cardiac output to the liver in fish is estimated to be approximately 20% [10,11]. Using this information as preliminary guidance, screening-level theoretical maximum whole-body kM,MAX values were assumed to be up to a factor of five greater than the suggested hepatic values to account for possible extrahepatic biotransformation. Thus, whole-body kM,MAX values for a 10-g ?sh at 15 C were estimated as 125 (log KOW 1), 100 (1 log KOW 2), 75 (2 log KOW 3), 50 (3 log KOW 4), and 25 (log KOW 4) per day.

These criteria were used to ag kM values that were possibly too high and assign these values as having greater uncertainty while recognizing that whole-body rates will be chemical specific.

4.3.Descriptors in the model:

4.4.Descriptor selection:

4.5.Algorithm and descriptor generation:

4.6.Software name and version for descriptor generation:

4.7.Chemicals/Descriptors ratio:

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

For each predicted compound there is the final assessment of the prediction (i.e. the prediction made together with the analysis of the applicability domain).

– Applicability Domain: Similar compounds, with predicted and experimental values Here it is reported the list of the six most similar compounds found in the training and test set of the model, along with their depiction and relevant information (mainly experimental value and predicted value).

– Applicability Domain: Measured Applicability Domain scores Here it is reported the list of all Applicability Domain scores, starting with the global Applicability Domain Index (ADI). Note that the final assessment on prediction reliability is given on the basis of the value of the ADI. For each index, it is reported its value and a brief explanation of the meaning of that value.

– Reasoning: Relevant chemical fragments and moieties If some rare and/or missing Atom Centered Fragments are found, they are reported here with a depiction of each fragment.</body></html>

5.2.Method used to assess the applicability domain:

The applicability domain of predictions is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model, calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments). For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a negative evaluation.

Following, all applicability domain components are reported along with their explanation and the intervals used. - Similar molecules with known experimental value. This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation.

Defined intervals are: 1 >= index > 0.85 strongly similar compounds with known experimental value in the training set have been found 0.85 >= index > 0.7 only moderately similar compounds with known experimental value in the training set have been found index <= 0.7 no similar compounds with known experimental value in the training set have been found

- Accuracy (average error) of prediction for similar molecules. This index takes into account the error in prediction for the two most similar compounds found. Values near 0 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model behaves. Defined intervals are: index < 0.5 accuracy of prediction for similar molecules found in the training set is good 0.5 <= index < 1.0 accuracy of prediction for similar molecules found in the training set is not optimal index > 1.0 accuracy

of prediction for similar molecules found in the training set is not adequate
- Concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules) . This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made agrees with the experimental values found in the model's space, thus the prediction is reliable. Defined intervals are: index < 0.5 similar molecules found in the training set have experimental values that agree with the target compound predicted value 0.5 <= index < 1.0 similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value index > 1.0 similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value
- Maximum error of prediction among similar molecules. This index takes into account the maximum error in prediction among the two most similar compounds. Values near 0 means that the predicted compounds falls in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are: index < 0.5 the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability 0.5 <= index < 1.0 the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability index >= 1.0 the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability
- Global AD Index. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are: 1 >= index > 0.85 predicted substance is into the Applicability Domain of the model 0.85 >= index > 0.75 predicted substance could be out of the Applicability Domain of the model index
<= 0.75 predicted substance is out of the the Applicability Domain of the model

5.3.Software name and version for applicability domain assessment:

Global AD Index

5.4.Limits of applicability:

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: Yes

INChI: Yes

MOL file: Yes

NanoMaterial: 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:

6.6.Pre-processing of data before modelling:

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES. All molecules found as input are preprocessed before the calculation of molecular descriptors, in order to obtain a standardized representation of compound. For this reason, some cautions should be taken. - Hydrogen atoms. In SDF files, hydrogen atoms should be explicit. As some times SDF file store only skeleton atoms, and hydrogen atoms are implicit, during the processing of the molecule the system tries to add implicit hydrogens on the basis of the known standard valence of each atom (for example, if a carbon atoms has three single bonds, an hydrogen atom will be added such to reach a valence of four). In SMILES molecules, the default notation uses implicit hydrogen. Anyway please note that in some cases it is necessary to explicitly report an hydrogen; this happens when the conformation is not unambiguous. For example, when a nitrogen atom is into an aromatic ring with a notation like "cnc" it is not clear whether it corresponds to C-N=C or to C-[NH]-C, thus if the situation is the latter, it should be explicitly reported as "c[nH]c". - Aromaticity. The system calculates aromaticity using the basic Hueckel rule. Note that each software for drawing and storing of molecules can use different approaches to aromaticity (for instance, commonly the user can choose between the basic Hueckel rule and a loose approach that lead to considering aromatic a greater number of rings). As in the input files aromaticity can be set explicitly (for instance, in SMILES format by using lowercase letters), during the processing of the molecule the system removes aromaticity from rings that don't satisfy the Hueckel rule. Please note that when aromaticity is removed from a ring, it is not always possible to rebuild the original structure in Kekule form (i.e. with an alternation of single and double bonds, like in the SMILES for benzene, C=1C=CC=CC1), in this case all bonds are set to single. Furthermore, please note that aromaticity detection is a really relevant issue, some molecular descriptors can have significantly different values whether a ring is perceived as aromatic or not. For this reason it is strongly recommended: - Always use explicit hydrogens in SDF file. - Avoid explicit aromaticity notation in original files; in this way, the perception of aromaticity is left to the preprocessing step and there is no chance of mistakes due to the transformation of rings that were set to aromatic in the original format but not recognized as aromatic in VEGA.

Note that when some modification of the molecule are performed during the preprocessing (e.g. adding of lacking hydrogens, correction of aromaticity), a warning is given in the remark field of the results.

6.7.Statistics for goodness-of-fit:

Tot RMSE 0.58 R2 0.75 mean obs 0.50 n 632 Training RMSE 0.87 R2 0.77 n 252 mean obs 0.13 Test RMSE 0.51 R2 0.79 n 63 mean obs 0.30

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

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

6.10.Robustness - Statistics obtained by Y-scrambling:

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:

No

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: Yes

INChI: Yes

MOL file: Yes

NanoMaterial: 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:

7.6.Experimental design of test set:

7.7.Predictivity - Statistics obtained by external validation:

7.8.Predictivity - Assessment of the external validation set:

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:

8.2.A priori or a posteriori mechanistic interpretation:

The mechanistic interpretation of the model is provided a posteriori, i.e. by interpretation of the final set of the selected descriptors.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

9.3.Supporting information:

Training set(s)

dataset_KM_ARNOT_training.csv	file:///C:\Users\Lenovo\Documents\lavoro_QMRF \kMHalf-Life Model version 1.0.0\dataset_KM_ARNOT_training.csv

Test set(s)

dataset_KM_ARNOT_test.csv	file:///C:\Users\Lenovo\Documents\lavoro_QMRF \kMHalf-Life Model version 1.0.0\dataset_KM_ARNOT_test.csv

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