	***QMRF identifier (JRC Inventory):*** **To be entered by JRC**
	QMRF Title: **kMHalf-Life Model version 1.0.0 Arnot/episuite v 1.0.1**
	Printing Date: November 2022

1.QSAR identifier

1.1.QSAR identifier (title):

kMHalf-Life Model version 1.0.0 Arnot/episuite v 1.0.1

1.2.Other related models:

NA

1.3.Software coding the model:

VEGA (https://www.vegahub.eu/)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

2.General information

2.1.Date of QMRF:

November 2022

2.2.QMRF author(s) and contact details:

Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy (Emilio.benfenati@marionegri.it) https://www.marionegri.it/

2.3.Date of QMRF update(s):

No update

2.4.QMRF update(s):

No update

2.5.Model developer(s) and contact details:

Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

2.6.Date of model development and/or publication:

NA

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

[1] 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] Benfenati E, Manganaro A, Gini G. VEGA-QSAR: AI inside a platform for predictive toxicology Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy Published on CEUR Workshop Proceedings Vol-1107

2.8.Availability of information about the model:

The model is non-proprietary and the training set is available.

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

Another QMRF is not available.

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

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:

NA

3.4.Endpoint units:

Biotransformation half-life (log kM/Half-Life (days)and rate constants (kM; day)

3.5.Dependent variable:

(log kM/Half-Life (days)

3.6.Experimental protocol:

NA

3.7.Endpoint data quality and variability:

The dataset was retrieved from BCFBAF model of EPISUITE. The dataset of 631 experimental kM biotransformation rates in fish (compiled in units of log biotransformation half-lives in days) was divided into a training set of 421 compounds for model derivation and validation set of 210 compounds for model testing.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model is a re-implementation of EPISUITE Biotransformation (kM) BCFBAF model. The original model is described in Arnot [sec 9.2]

The model is based on a dataset of 631 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.

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 centra 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).

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 Kow limitations to the liver and protein binding.

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 Kow rates. The total cardiac output to the liver in fish is estimated to be approximately 20%. 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.

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:

NA.

4.4.Descriptor selection:

NA

4.5.Algorithm and descriptor generation:

NA.

4.6.Software name and version for descriptor generation:

NA

4.7.Chemicals/Descriptors ratio:

NA

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.

ADI is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.85, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.85 ≥ AD index > 0.75, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index ≤ 0.75, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

No ADI threshold was used to provide performance calculations of the validation sets

5.2.Method used to assess the applicability domain:

The Applicability domain and chemical similarity are measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [3]. 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.

These indices are defined in this way for this QSAR model:

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:

If 1 ≥ index > 0.85, strongly similar compounds with known experimental value in the training set have been found

If 0.85 ≥ index > 0.7, only moderately similar compounds with known experimental value in the training set have been found

If 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 classification accuracy in prediction for the two most similar compounds found. Values near 1 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

If index < 0.5, accuracy of prediction for similar molecules found in the training set is good

If 1.0 ≥ index ≥ 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index > 1.0, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for 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 disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If index < 0.5, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1.0 ≥ index ≥ 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If 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 between similar molecules:

This index takes into account the maximum error in prediction between the two most similar compounds. Values near 0 means that the predicted compounds fall in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

If index < 0.5, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.0 ≥ index ≥ 0.5, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index > 1.0, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

5.3. Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4. Limits of applicability:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if 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: 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:

NA

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

Training: RMSE 0.54; R^2 =0.79; n = 421

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

NA

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

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

NA

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

NA

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

NA

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

NA

7.5.Other information about the external validation set:

NA

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

Test set: n 210, RMSE 0.65, R2 0.67

Test set in AD: n 70, RMSE 0.36, R2 0.90

Test set could be out of AD: n 97, RMSE 62, R2 68

Test set out of AD: n 37, RMSE 1.08, R2 –0.05**7.8.Predictivity - Assessment of the external validation set:**

NA

7.9.Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

NA

8.2.A priori or a posteriori mechanistic interpretation:

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

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] Arnot JA, Mackay D, Bonnell M., “Estimating metabolic biotransformation rates in fish from laboratory data.” Environmental Toxicology and Chemistry (2008), 27, 341-351.

[2] 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.

[3] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

9.3.Supporting information:

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

All available dataset are present in the model inside the VEGA software.

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