	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: HYDROLYSIS (IRFMN/CORAL) - v. 1.0.1
	Printing Date: November 2022

1.QSAR identifier

1.1.QSAR identifier (title):

Hydrolysis (IRFMN/CORAL) v.1.0.1

1.2.Other related models:

No

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:

[1] 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] Alla P. Toropova Laboratory of environmental chemistry and toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS (alla.toropova@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:

[1] Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy (andrey.toropov@marionegri.it) https://www.marionegri.it/

[2] Alla Toropova Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy (alla.toropova@merionegri.it) https://www.marionegri.it/

[3] Giovanna J. Lavado Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy (giovanna.lavado@merionegri.it) https://www.marionegri.it/

2.6.Date of model development and/or publication:

April 22, 2020

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

[1] A. A. Toropov, A. P. Toropova, A. Lombardo, A. Roncaglioni, G. J. Lavado, and E. Benfenati, ‘The Monte Carlo method to build up models of the hydrolysis half-lives of organic compounds’, SAR and QSAR in Environmental Research, vol. 32, no. 6, pp. 463–471, Jun. 2021, doi: 10.1080/1062936X.2021.1914156.

[2] Toropov A.A., Toropova, A.P., Roncaglioni, A., Benfenati, E. Prediction of biochemical endpoints by the CORAL software: Prejudices, Paradoxes, and Results. (2018) Methods in Molecular Biology, 1800, pp. 573-583.

[3] 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:

NA

3.2.Endpoint:

ENV FATE 5.1.2. Hydrolysis. OECD, Test No. 111: Hydrolysis as a Function of pH

3.3.Comment on endpoint:

According to OECD, the test should be performed at three different pH (4, 7, and 9) at two different temperatures (50 and 25°C). This model is based on the hydrolysis data at 25°C and pH7. All the data were half-life (HL) in days converted in logarithm units.

3.4.Endpoint units:

logarithmic scale for hydrolysis half-life

3.5.Dependent variable:

Log(HL)

3.6.Experimental protocol:

Test Guideline No. 111

3.7.Endpoint data quality and variability:

Data were retrived from [6]. As in Khan et al. [6], the HHL values at the temperature of 25°C, pH 7, were selected according to the OECD 111 guideline. We used data for 70 substances. The structure was curated using KNIME software with an in-house workflow that searches for the SMILES on the web using the CompTox Chemicals Dashboard (https://comptox. epa.gov/dashboard) and the Chemical Identifier Resolver (CIR; https://cactus.nci.nih.gov/ chemical/structure) databases, starting from CAS and name. Then, the structure available from AMBIT (http://cefic-lri.org/toolbox/ambit/) and the ones retrieved from the web was processed to get the correct Kekulé structure (RDKit Kekulizer node), transform in the canonical format (RDKit Canon SMILES node) and, finally, standardize them (Lychi Standardizer node). All the SMILES structures were compared for consistency (chemicals with the doubtful structure were eliminated). These compounds were randomly distributed into a training set (≈60%), calibration set (≈20%), and validation set (≈20%). We studied three random splits (prepared in these percentages).

After the implementation in VEGA, the dataset of 70 substances was split in training (58 substances) and test (12 substances)

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The Hydrolysis (IRFMN/CORAL) v.1.0.1 is a one-variable model based on 2D descriptor. This model is based on the hydrolysis data at 25°C and pH7

4.2.Explicit algorithm:

The Monte Carlo method

Endpoint = -1.2867 (± 0.0419) + 0.2545 (± 0.0033) * DCW(1,35)

4.3.Descriptors in the model:

NA

4.4.Descriptor selection:

2D optimal descriptor

$$DCW(T^*, N^*) = CW(C5) + CW(C6) + APP(N,O,S) + \sum_{k=1}^{NA} CW(S_k) + \sum_{k=1}^{NA-1} CW(SS_k) + \sum_{k=1}^{NA-2} CW(SSS_k)$$

4.5.Algorithm and descriptor generation:

The Monte Carlo Method

4.6.Software name and version for descriptor generation:

CORAL 2019 (modified)

Istituto di Ricerche Farmacologiche Mario Negri IRCCS - 20124 Milano, Italy

http://www.insilico.eu/coral/

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 > 0.7, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index ≤ 0.7, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

5.2.Method used to assess the applicability domain:

The AD and 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 [7]. The 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 < 0.6, accuracy of prediction for similar molecules found in the training set is good

If 1.2 > index ≥ 0.6, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1.2, 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 < 0.6, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1.2 > index ≥ 0.6, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1.2, 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 < 0.6, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.2 > index ≥ 0.6, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1.2, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

Model descriptors range check:

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

If index = True, descriptors for this compound have values inside the descriptor range of the compounds of the training set

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

Atom Centered Fragments similarity check:

This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 1 > index ≥ 0.7, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If index < 0.7, a prominent number of atoms centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

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). 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: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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

NA

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

NA

6.5.Other information about the training set:

The initial dataset was divided into three sets (training, calibration, and validation) using a random distribution.

6.6.Pre-processing of data before modelling:

The SMILES were pre-processed using the KNIME software to kekulize (RDKit Kekulizer node), canonicalize (RDKit Canon SMILES node) and standardize (Lychi Standardizer node) the structure retrieved from the web (using Comptox and Cactus form both CAS and name) and from AMBIT. All the SMILES structures obtained were compared for consistency.

6.7.Statistics for goodness-of-fit:

After the implementation in VEGA:

Training n=58, R2=0.71, RMSE= 0.84

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:

Form the original model:

	Training set n= 44	Calibration set n= 14
R^2	0.7170	0.8542
1	0.0424	0.0009
2	0.0032	0.0922
3	0.0239	0.0061
4	0.1430	0.0859
5	0.0001	0.0509
6	0.0004	0.0868
7	0.0001	0.0014
8	0.0028	0.0163
9	0.0002	0.0482
10	0.0717	0.0000
Y-scrambling	0.0288	0.0389

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:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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:

All

7.5.Other information about the external validation set:

NA

7.6.Experimental design of test set:

Available

7.7.Predictivity - Statistics obtained by external validation:

After the implementation in VEGA:

Test set n=12, R2=0.53, RMSE= 0.61

Test set could be out of AD: n 1

Test set out of AD: n 11, R2 0.56, RMSE 0.56

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:

Analysis of results on several runs of the Monte Carlo optimization

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori only.

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] Toropova, A.P., Toropov, A.A., Benfenati, E., Castiglioni, S., Bagnati, R., Passoni, A., Zuccato, E., Fanelli, R. Quasi-SMILES as a tool to predict removal rates of pharmaceuticals and dyes in sewage
(2018) Process Safety and Environmental Protection, 118, pp. 227-233.

[2] Toropova, A.P., Toropov, A.A., Benfenati, E., Leszczynska, D., Leszczynski, J. Prediction of antimicrobial activity of large pool of peptides using quasi-SMILES (2018) BioSystems, 169-170, pp. 5-12.

[3] Andrey A. Toropov, Alla P. Toropova, Alessandra Roncaglioni, Emilio Benfenati. Prediction of biochemical endpoints by the CORAL software: Prejudices, Paradoxes, and Results. (2018) Methods in Molecular Biology, 1800, pp. 573-583. https://link.springer.com/protocol/10.1007%2F978-1-4939-7899-1_27

[4] Toropova, A.P., Toropov, A.A., Marzo, M., Escher, S.E., Dorne, J.L., Georgiadis, N., Benfenati, E.
The application of new HARD-descriptor available from the CORAL software to building up NOAEL models
(2018) Food and Chemical Toxicology, 112, pp. 544-550.

[5] OECD, *Test No. 111: Hydrolysis as a Function of pH*. Paris: Organisation for Economic Co-operation and Development, 2004. Accessed: Nov. 09, 2022. [Online]. Available: https://www.oecd-ilibrary.org/environment/test-no-111-hydrolysis-as-a-function-of-ph_9789264069701-en

[6] P.M. Khan, A. Lombardo, E. Benfenati, and K. Roy, First report on chemometric modeling of hydrolysis half-lives of organic chemicals, Environ. Sci. Pollut. Res. 28 (2021), pp. 1627–1642. doi:10.1007/s11356-020-10500-0.

[7] Floris et al. "A generalizable definition of chemical similarity for read-across." Journal of cheminformatics 6.1 (2014): 39

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