	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Air Half-life (CORAL) - v. 1.0.1**
	Printing Date: **April 2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Air Half-life (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:

April 2022

2.2.QMRF author(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] Erika Colombo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy (erika.colombo@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:

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

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

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

2.6.Date of model development and/or publication:

July 9, 2019

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

[1] Pathan Mohsin Khan,Diego Baderna,Anna Lombardo,Kunal Roy,Emilio Benfenati, Chemometric modeling to predict air half-life of persistent organic pollutants (POPs). (2020) Journal of Hazardous Materials https://doi.org/10.1016/j.jhazmat.2019.121035

[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:

NA

3.2. Endpoint:

ENV FATE 5.1.1. Phototransformation in Air

3.3. Comment on endpoint:

NA

3.4. Endpoint units:

Half life = hours

3.5. Dependent variable:

Log(hours)

3.6. Experimental protocol:

NA

3.7. Endpoint data quality and variability:

The experimental air half-life data of mono-constituent organic chemicals were extracted from Gouin et al. 2004 [2] and Gramatica and Papa 2007 [3] according to the Organization of Economic Cooperation and Development (OECD) principle one [4].

The first source of data, Gouin et al. (2004) [1], contains categorical data and no details on the test. As mentioned in Gouin et al 2004: "*half-lives were assigned on a semi-decade logarithmic scale to one of nine classes as follows: (1) 5 h (range: 0– 10 h), (2) 17 h (10–30), (3) 55 h (30–100), (4) 170 h (100– 300), (5) 550 h (300–1000), (6) 1700 h (1000–3000), (7) 5500 h (3000–10 000), (8) 17 000 h (10 000–30 000) and (9) 55 000 h (30 000–100 000)… To the extent possible, these allocations were made by careful analysis of experimental degradation rate data, but inevitably a high degree of scientific judgment was involved. It is recognized that by allocating a chemical to a half-life class, there is likely to be an estimation error of ±1 to 2 classes (Mackay et al., 1999)".*The original data source is a handbook reporting an assessment based on scientific judgment of the available experimental and estimated data [6].

The extracted data of air half-life were carefully curated following the second OECD principle for QSAR modeling, to remove the duplicates and inorganic molecules including hydrates to avoid any systematic errors. All the chemical structures were drawn manually using JChem excel plugin from the SMILE notation (JChem, 2019 [9]), a ChemAxon tool (https://chemaxon.com/products/jchem-for-office [10]) and cross-verified using Chemical Abstracts Service (CAS) numbers from the Chemical book (available from https://www.chemicalbook.com/ [11]) and PubChem database (available from https://pubchem.ncbi.nlm.nih.gov/ [12]) to rectify any error in chemical structure.

The experimental data of air half-life were transformed into the logarithmic scale (LOG10) for getting easily manageable numbers. The final dataset was composed of 302 mono-constituent organic chemicals. The data set was distributed in the active training set (≈25%), invisible training set (≈25%), calibration set (≈25%), and external validation set (≈25%) which are categories which are sued in the specific CORAL modeling approach (http://www.insilico.eu/coral/) c.f. also point 6.5 [6]

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

One-variable model based on 2D descriptor.

4.2.Explicit algorithm:

Monte Carlo method based on CORAL descriptors.

Endpoint = -2.2271129+ 0.0900047 × DCW(1,15)

Monte Carlo optimisation is based on two kinds of target functions TF_1 and TF_2:

$$TF_1 = r_{TRN} + r_{iTRN} - |r_{TRN} - r_{iTRN}| * 0.3$$

$$TF_2 = TF_1 + IIC_{CLB} * 0.1$$

The r_{TRN} and r_{iTRN} are the correlation coefficient between the experimental and predicted values of the endpoint for active training and passive training sets. The IIC_{CLB} is the index of ideality of correlation which is calculated as:

$$IIC_{CLB} = r_{CLB} \frac{\min(^{-}MAE_{CLB}, {}^{+}MAE_{CLB})}{\max(^{-}MAE_{CLB}, {}^{+}MAE_{CLB})}$$

$$^{-}MAE_{CLB} = \frac{1}{^{-}N} \sum_{k=1}^{^{-}N} |\Delta_k|, \quad ^{-}N \text{ is the number of } \Delta_k < 0$$

$$^{+}MAE_{CLB} = \frac{1}{^{+}N} \sum_{k=1}^{^{+}N} |\Delta_k|, \quad ^{+}N \text{ is the number of } \Delta_k \leq 0$$

$$\Delta_k = \text{observed}_k - \text{calculated}_k$$

The r_{CLB} is the correlation coefficient for the experimental and calculated values of the endpoint for the calibration set. MAE is the mean absolute error. The $+MAE_{CLB}$ is calculated for positive Δ_k; the $-MAE_{CLB}$ is calculated for negative Δ_k.

For this model, the target function TF_2 gives better models in comparison with the Monte Carlo optimisation based on theTF_1.

4.3.Descriptors in the model:

Applied the correlation weights for standard SMILES-attributes S, SS, and SSS together with the correlation weights of HARD codes.

4.4.Descriptor selection:

2D optimal descriptor
DCW(T^*,N^*)=CW(NOSP)+CW(HALO)+CW(BOND)+_(k=1)^NACW(S_k)+_(k=1)^NvCW(EC0_k)

4.5.Algorithm and descriptor generation:

The Monte Carlo Method for the optimization, and SMILES attributes as descriptors.

The optimal descriptor of correlation weights (DCW) used here is calculated as the following

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

where Sk is SMILES atom; SSk is pairs of SMILES atoms; SSSk is compositions of three SMILES atoms; the CW (Sk), CW(SSk), and CW(SSSk) are correlation weights of the above SMILES attributes; the T is the threshold to separate SMILES attributes in two classes, i.e. rare and non-rare. The N is the number of epochs of the Monte Carlo optimisation. The T* and N* are values which provide better statistics for the calibration set. The NA is the number of Sk in the SMILES.

The correlation weights are calculated with the Monte Carlo method (http://www.insilico.eu/coral/).

The model for Y calculated with numerical data on the correlation weights, is the following:

$$Y = C_0 + C_1 * DCW(T^*, N^*)$$

The C0 and C1 are regression coefficients. The DCW (T*,N*) is calculated with Eq. 1. The Y is the endpoint under consideration.

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.

Indices are calculated on the first k = 2 most similar molecules, each having S_k similarity value with the target molecule.

Similarity index (*IdxSimilarity*) is calculated as:

$$\frac{\sum_k S_k}{k} \times (1 - Diam^2)$$

where *Diam* is the difference in similarity values between the most similar molecule and the *k*-th molecule.

Accuracy index (*IdxAccuracy*) is calculated as:

$$\frac{\sum_c^k |exp_c - pred_c|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_c$ is the c-*th* molecule predicted value by the model.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c^k |exp_c - pred_{target}|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule.

Max Error index (*IdxMaxError*) is calculated as:

$$max(|exp_c - pred_c|)$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule, evaluated over the k molecules.

ACF contribution (*IdxACF*) index is calculated as

$$ACF = rare \times missing$$

where: *rare* is calculated on the number of fragments found in the molecule and found in the training set in less than 3 occurences as following: if the number is 0, *rare* is set to 1.0; if the number is 1, *rare* is set to 0.6; otherwise *rare* is set to 0.4

missing is calculated on the number of fragments found in the molecule and never found in the training set as following: if the number is 0, *missing* is set to 1.0; if the number is 1, *missing* is set to 0.6; otherwise *missing* is set to 0.4

Descriptors Range (*IdxDescRange*) index is calculated as 1.0 if all molecular descriptors used in the prediction fall within the range of descriptors used in the whole training set, 0.0 otherwise.

AD final index is calculated as following:

$$ADI = IdxSimilarity \times IdxACF \times IdxDescRange$$

The initialADI index is the used together with the other sub-indices to calculate the final ADI, on the basis of the assessment class in which each sub-index falls:

IdxAccuracy ≥	IdxConcordance ≥	IdxMaxError ≥	InitialADI ≥	ADI
1.2	1.2	1.2	0.85	1.0
0.6	0.6	0.6	0.7	0.85
All other cases				0.7

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 [5]. 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, 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: 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:

All

6.5.Other information about the training set:

The original total set of compounds has been randomly split into the training (75%) and validation (25%) sets. The training set is composed of “Training”, “Invisible Training” and “Calibration”. Invisible training and calibration sets were used during the model development for tuning model's parameters".

The training represents an *active* training set and is used to build up optimal correlation weights for the optimal descriptor; in other word it represents the set on which the model is built on. The passive (or invisible) training set is for the purpose of checkup whether current correlation weights (and the optimal descriptor) are satisfactory for chemicals, which are not involved in the calculation of the correlation weights

The task for the calibration set is to detect the moment when overtraining begins.[6]**6.6.Pre-processing of data before modelling:**

The SMILES were pre-processed using the KNIME software.

6.7.Statistics for goodness-of-fit:

Set	n	R^2	Q^2	RMSE	MAE
Training	76	0.81	0.80	0.39	0.29
Invisible Training	77	0.82	0.81	0.38	0.30
Calibration	74	0.71	0.69	0.50	0.38

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

Q^2_{LOO}=0.803

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

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

Y scrambling 0.011

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:

Set	n	R^2	Q^2	RMSE	MAE
Validation	75	0.70	0.68	0.55	0.41

Test set in AD: n = 34; R2 = 0.89; RMSE = 0.29

Test set could be out of AD: n = 23; R2 = 0.62 RMSE = 0.60

Test set out of AD: n = 18; R2 = -0.32; RMSE = 0.78

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] Toropov A., Toropova A., 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. https://link.springer.com/protocol/10.1007%2F978-1-4939-7899-1_27

[2] T. Gouin, I. Cousins, D. Mackay Comparison of two methods for obtaining degradation half-lives Chemosphere, 56 (2004), pp. 531-535

[3] P. Gramatica, E. Papa Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure

Environ. Sci. Technol., 41 (2007), pp. 2833-2839

[4] Validation of (Q)SAR Models—OECD. Retrieved November 25, 2019, from https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm

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

[6] Toropov, A. A., Toropova, A. P., Roncaglioni, A., & Benfenati, E. (2021). The system of self-consistent semi-correlations as one of the tools of cheminformatics for designing antiviral drugs. New Journal of Chemistry, 45(44), 20713–20720. https://doi.org/10.1039/D1NJ03394H

[7] Toropov A., Toropova A., 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. https://link.springer.com/protocol/10.1007%2F978-1-4939-7899-1_27

[8] A.P. Toropova, A.A. Toropov, A. Lombardo, G. Lavado, and E. Benfenati, Paradox of "ideal correlations": improved model for air half-life of persistent organic pollutants. Environmental Technology, Accepted Jan 22, 2021. DOI: 10.1080/09593330.2021.1882588

[9] JChem, https://chemaxon.com/products/jchem-for-office, (2019)

[10] JChem for Office | ChemAxon. Retrieved November 25, 2019, from https://chemaxon.com/products/jchem-for-office

[11] ChemicalBook—Chemical Search Engine. Retrieved November 25, 2019, from https://www.chemicalbook.com/

[12] PubChem. PubChem. Retrieved November 25, 2019, from https://pubchem.ncbi.nlm.nih.gov/

[13] Mackay, D., Shiu, W.Y., Ma, K.C., 1999. Physical–Chemical Properties and Environmental Fate Handbook; CRC netBASE CD-ROM. Chapman and Hall/CRC Press, Boca Raton, FL

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