QMRF identifier (JRC Inventory): **To be entered by JRC**
QMRF Title: **Persistence (soil) quantitative model (IRFMN) - v. 1.0.1**
Printing Date: **20-10-2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Persistence (soil) quantitative model (IRFMN) - 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:

20-10-2022

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] Elena Boriani boriani.elena@gmail.com

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] Anna Lombardo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy anna.lombado@marionegri.it https://www.marionegri.it/

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

https://www.vegahub.eu/contacts/ benfenati.emilio@gmail.com https://www.vegahub.eu

2.6.Date of model development and/or publication:

2018

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

[1] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, and Screening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology; Springer; 2019. p. 365-81.

[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:

Persistence in soil measured as half-life in days

3.2.Endpoint:

ENV FATE 5.2.3. Biodegradation in soil. OECD. Test No. 307: Aerobic and Anaerobic Transformation in Soil (2002)

3.3.Comment on endpoint:

The model is based on half-life ultimate biodegradation test data and provide evaluation of persistence property in the water compartment.

3.4.Endpoint units:

Days.

3.5.Dependent variable:

Log(days)

3.6.Experimental protocol:

We cleaned the data following strict eligibility criteria based on the OECD 307 test guideline [3]. For the procedure, see point 3.7 (data quality).

3.7.Endpoint data quality and variability:

Dataset cleaning consists of two different parts: a data and a structure cleaning. The first is endpoint and goal dependent. The data have been cleaned considering the testing guidelines, the regulatory thresholds and the goals of the work. Considering the low quality of the data reported in the other sources, we maintained only the ECHA and Gouin et al., 2004 data. ECHA were selected applying the OECD testing guideline 309.

The first source of data, Gouin et al. (2004), 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)".*[8] For this dataset, we eliminated the substances whose ranges are between two persistence classes ((not persistent, or nP; persistent, or P[1]; very persistent, or vP[2])) and the compounds with a range greater than 30 days.

The other source, the ECHA registration data (extracted from the ECHEMPORTAL website (https://www.echemportal.org/echemportal/) in October 2016 and February 2017) [2], contains continuous values and some details on the test. For this source, we eliminated data with reliability 3, not tested in lab, without a guideline or based on guidelines different from the OECD 307. After the manual check, we maintained all the data obtained in aerobic conditions and tested on the identity substance. The mixture of substances or the data not found on the web site of the ECHA were eliminated. We identified the outliers in four ways:

- Concordance (maximum - minimum): we considered outliers all the values with a difference greater than 30 days (arbitrary chosen).

[1] P: Degr.half-life (of parent compound or any of its degr. products) > 40 days in fresh surface water

[2] vP: Degr.half-life (of parent compound or any of its degr. products) > 60 days in fresh surface water

- Concordance (maximum/minimum): we considered outliers all the values with a value greater than 3.
- Concordance (class range): we considered outliers all the values with the minimum and the maximum reported in different classes (nP, P and vP).

Log half-life distribution: we verified the normal distribution (with the R tool) and we considered outliers all the values outside the range: average ± 3SD (standard deviation).

We performed the second part, the structure cleaning, using an in-house KNIME workflow that searches for the structures in the DSStox data base (US EPA.c, 2019) and through the CIR (2019) node both from the name and the CAS no [6]. Once neutralised and normalised with the VEGA node, the software compares the structures and asks for a manual check, using online databases like ChemIDplus (NIH, 2019), PubChem compound (https://www.ncbi.nlm.nih.gov/pccompound), Sigma-Aldrich (Merck KGaA, 2019) and GuideChem (http://www.guidechem.com/) in case of no agreement. The workflow checks for duplicates too. We eliminated all doubtful compounds, the inorganic atoms, the mixtures of isomers (excluded stereoisomers), the mixtures of compounds, the UVCB (Chemical Substances of Unknown or Variable Composition, Complex Reaction Products and Biological Materials), the metalorganic compounds and the macromolecules. We finally obtained a dataset of 226 molecules for persistency in soil. We split the dataset into a training (181 mono constituent organic compounds) and a test sets (45 mono constituent organic compounds) using a stratified sampling and forcing ECHA substances in the test set.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Counter-propagation neural network (CPANN) present at NIC (https://www.ki.si/en/departments/).[5]

4.2.Explicit algorithm:

We calculated a wide set (1022) of 2D molecular descriptors with the VEGA core libraries. Then we used an in-house tool developed in the R software to select the best descriptors set to be used in each of the three models.

The approach employed is based on a forward selection technique. It starts from the descriptor that has the highest correlation with the experimental values; then, at each iteration, all the remaining descriptors are tested and the one leading to the best model is added. The models in this process are simple linear regressions, applied with a bootstrap cross-validation approach: 500 iterations — for each one the regression model is built on a random subset (60% of the original training set) and tested on the remaining 40% with the calculation of a fitness function (a linear combination of the R squared coefficient and the root mean square error (RMSE)).

The progressive addition of descriptors to the model increases the cross-validation performance up to a plateau, where the optimal number of descriptors is reached, and adding further descriptors causes over-fitting.

Once the descriptors were selected, the final models were developed using counter-propagation neural network (CPANN) software developed by the National Institute of Chemistry, the CPANNatNIC [5].

4.3.Descriptors in the model:

[1] MLogP MLogP Descriptor Moriguchi LogP

[2] SpPosA_p Burden Eigenvalue Descriptors Normalized spectral positive sum

[3] CATS2D_3_DL CATS 2D Descriptors CATS 2D D-L at topological distance 3

[4] EEig7ri Edge Adjacency Descriptors

[5] P_VSA_i_3 PVSA Descriptors

[6] B1(C..O) Topological Distances Descriptors Presence/absence of C-O at topological distance 1

[7] B1(C..Cl) Topological Distances Descriptors Presence/absence of C-Cl at topological distance 1

[8] B3(O..O) Topological Distances Descriptors Presence/absence of O-O at topological distance 3

4.4.Descriptor selection:

Descriptors are selected using an in-house tool developed in R (with a bootstrap cross-validation approach) and then implemented in java using cdk libraries.

4.5.Algorithm and descriptor generation:

We calculated the descriptors using the pool of descriptors implemented in VEGA and we selected the most relevant ones using an in-house validated forward selection in R (with a bootstrap cross-validation approach). The selected number of descriptors is 8.

4.6.Software name and version for descriptor generation:

VEGA implemented descriptors available in-house.

4.7.Chemicals/Descriptors ratio:

181/8 = 23

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:

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

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 Applicability Domain and the 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 [4]. 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 < 0.6, 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 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 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

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 atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

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:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

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:

Training set n = 181.

6.6.Pre-processing of data before modelling:

The procedure is written in 3.7 (data quality).

6.7.Statistics for goodness-of-fit:

Training set: n = 181; R2 = 0.96; RMSE = 0.17

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

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, part of the test set is proprietary (ECHA data)**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 = 45; R2 = 0.67; RMSE = 0.82

Test set in AD: n = 32; R2 = 0.83; RMSE = 0.34

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

The test set is not available because contains proprietary data.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

No assumption on the mechanism is done.

8.2.A priori or a posteriori mechanistic interpretation:

NA

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

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

[2] ‘EChemPortal Provides Free Public Access to Information on Properties of Chemicals’: Accessed 1 March 2022. https://www.echemportal.org/echemportal/.

[3] OECD. Test No. 307: Aerobic and Anaerobic Transformation in Soil. Paris: Organisation for Economic Co-operation and Development, 2002. https://www.oecd-ilibrary.org/environment/test-no-307-aerobic-and-anaerobic-transformation-in-soil_9789264070509-en.

.

[4] Floris, Matteo, Alberto Manganaro, Orazio Nicolotti, Ricardo Medda, Giuseppe Felice Mangiatordi, e Emilio Benfenati. «A generalizable definition of chemical similarity for read-across». Journal of Cheminformatics 6, n. 1 (18 october 2014): 39. https://doi.org/10.1186/s13321-014-0039-1.

[5] Drgan, Viktor, Špela Župerl, Marjan Vračko, Claudia Ileana Cappelli, and Marjana Novič. ‘CPANNatNIC Software for Counter-Propagation Neural Network to Assist in Read-Across’. Journal of Cheminformatics 9, no. 1 (22 May 2017): 30. https://doi.org/10.1186/s13321-017-0218-y.

[6] Gadaleta, Domenico, Anna Lombardo, Cosimo Toma, and Emilio Benfenati. ‘A New Semi-Automated Workflow for Chemical Data Retrieval and Quality Checking for Modeling Applications’. Journal of Cheminformatics 10, no. 1 (10 December 2018): 60. https://doi.org/10.1186/s13321-018-0315-6.

[7] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, and Screening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology; Springer; 2019. p. 365-81.

[8] 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