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

1.QSAR identifier

1.1.QSAR identifier (title):

Persistence in soil 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:

31-03-2020

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:

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

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

2.6.Date of model development and/or publication:

2015

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] A. Manganaro, F. Pizzo, A. Lombardo, A. Pogliaghi, E. Benfenati, “Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm”, Chemosphere (2015)

[3] Pizzo, Fabiola, Anna Lombardo, Marc Brandt, Alberto Manganaro, and Emilio Benfenati. ‘A New Integrated in Silico Strategy for the Assessment and Prioritization of Persistence of Chemicals under

REACH'. Environment International 88 (1 March 2016): 250–60.
https://doi.org/10.1016/j.envint.2015.12.019.

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 compartment 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 test data and provide qualitative evaluation (four classes) of persistence property in the soil compartment.

3.4.Endpoint units:

Days.

3.5.Dependent variable:

Classification in four class, P, P/vP, nP/P, nP,**3.6.Experimental protocol:**

OECD. Test No. 307: Aerobic and Anaerobic Transformation in Soil (2002)

3.7.Endpoint data quality and variability:

HL data were collected from several sources. Gouin et al. (2004) [5] give information on HL, expressed in hours, for 233 organic compounds in nine classes (on a semi-decade log scale basis). It covers four environmental media: water (not specified whether marine or fresh water), sediment (not specified whether marine or fresh water), soil and air.

An average HL is assigned to each class, which is the only value available for each chemical. Gramatica and Papa (2007) [6] HL provide data for 250 organic compounds, referring to the same compartments and classified as in Gouin et al. (2004).

For all the compounds present in both datasets we double-checked the chemical structures and their correspondence with CAS number with ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus/) and Pubchem Compound (https://pubchem.ncbi.nlm.nih.gov/). Salts, mixtures, doubtful compounds and duplicates were eliminated, as well as duplicate compounds with conflicting experimental values. We obtained datasets of 298 mono costituent organic compounds. Another source was available from United States Geological Survey (USGS)[7] only for soil, containing 318 HL. These compounds were checked as above and the continuous values were classified following the same criteria as in Gouin et al. (2004), and then added to the soil dataset, obtaining a dataset of 537 compounds.

In the RIVM Report (Linders et al., 1994) [8] disappearance time 50 (DT50) for soil compartments were also available. DT50 differs from HL because HL refers to first or pseudo-first order reactions.

However, we assumed DT50s as similar to reported HLs. These compounds too were checked, classified and added to the dataset as above, obtaining final datasets of 568 data.

Differently from the model described in the cited paper ([3] in sec. 2.7) the k-NN has been rebuilt using all the available experimental data, indeed all compounds used as test set have been added to the training set, while the optimal configuration parameters for the k-NN model has been preserved. With this modification, the k-NN model exploits all the available information and performs slightly better than the statistics reported in the original paper

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model is based on the half-lives test data and provides a qualitative evaluation (four categories) of persistence property in the soil compartment. It has been developed using an ensemble of k-NN modelling and a set of alerts extracted with Sarpy software, by Istituto di Ricerche Farmacologiche Mario Negri.

4.2.Explicit algorithm:

This model has been built with the istKNN application (developed by Kode srl, http://chm.kode-solutions.net) and it is based on the similarity index developed inside the VEGA platform; the index takes into account several structural aspects of the compounds, such as their fingerprint, the number of atoms, of cycles, of heteroatoms, of halogen atoms, and of particular fragments (such as nitro groups). The index value ranges from 1 (maximum similarity) to 0. On the basis of this structural similarity index, the four compounds from the dataset resulting most similar to the chemical to be predicted are taken into account; compounds with a similarity value lower than 0.75 are discarded, and if only one compound remains available for prediction, it is kept only if it has a similarity value higher than 0.8. If no compounds fall under these conditions, no prediction is provided. The prediction is calculated as the most representative class in the compound selected with the above mentioned procedure, considering their similarity index values as a weight so that the most similar compounds have an higher influence in the prediction.

Differently from the model described in the cited paper ([3] in sec. 2.7) the k-NN has been rebuilt using all the available experimental data, indeed all compounds used as test set have been added to the training set, while the optimal configuration parameters for the k-NN model has been preserved. With this modification, the k-NN model exploits all the available information and performs slightly better than the statistics reported in the original paper.

The overall architecture provides firstly the prediction as calculated by the k-NN model. If some structural alerts are found, they do not change the prediction but modify the applicability domain value: if the alerts confirm the k-NN prediction, the applicability domain index (ADI) value increases, if the alert are in disagreement with the prediction the ADI value decreases. The alerts are anyway used to provide a prediction if the k-NN model is not able to predict the compound.

The predicted value is provided as one of the following four categories based on the categorization of non persistent (nP), persistent (P) and very persistent (vP) compounds: nP, nP/P, P/vP, vP. The nP class defines compound with HL certainly below the threshold of 120 days. The vP class includes compounds with HL certainly higher the threshold of 180 days. The remaining two categoriesindicate a not entirely certain situation: nP/P category represents compounds with HL values falling near the P HL threshold of 120 days; the P/vP caategory represents compounds with HL values falling near the vP HL threshold of 180 days.

4.3.Descriptors in the model:

Similarity index

Alerts by SarPy: fingerprints, the number of atoms, of cycles, of heteroatoms, of halogen atoms, and of particular fragments (such as nitro groups)

The following SAs have been extracted from the original dataset and are related to compounds which are nP in the soil compartment:

- nP alert no. 1, defined by the SMARTS: O=C(OCCC)
- nP alert no. 2, defined by the SMARTS: O=C(CC)C
- nP alert no. 3, defined by the SMARTS: C(O)c1ccccc1C
- nP alert no. 4, defined by the SMARTS: Cc1c(ccc(c1))OCC
- nP alert no. 5, defined by the SMARTS: COC(=O)CO
- nP alert no. 6, defined by the SMARTS: C(O)C=C
- nP alert no. 7, defined by the SMARTS: C(=NOC(=O)NC)C
- nP alert no. 8, defined by the SMARTS: N(C)CCCl
- nP alert no. 9, defined by the SMARTS: CSCc1ccccc1
- nP alert no. 10, defined by the SMARTS: O(C)CCCl
- nP alert no. 11, defined by the SMARTS: [P]

- nP alert no. 12, defined by the SMARTS: C(CN(C)C)S
- nP alert no. 13, defined by the SMARTS: C(=S)N(C)
- nP alert no. 14, multiple primary alcohols, defined by the SMARTS: C[CH2]O
- nP alert no. 15, multiple esters (aromatic), defined by the SMARTS: AC(=O)O*
- nP alert no. 16, single oximes (aliphatic), defined by the SMARTS: AC(A)=NO*
- nP alert no. 17, esters (aliphatic), defined by the SMARTS: AC(=O)O[*;!H]
- nP alert no. 18, single aldehydes (aliphatic), defined by the SMARTS: A[CH1](=O)
- nP alert no. 19, single carboxylic acids (aliphatic), defined by the SMARTS: AC(=O)O
- nP alert no. 20, single (thio-) carbamates (aliphatic), defined by the SMARTS: A[O,S]C(=[O,S])N(A)A
- nP alert no. 21, single ketones (aromatic), defined by the SMARTS: aC(=O)*
- nP alert no. 22, single phosphates/thiophosphates, defined by the SMARTS: *[O,S]=P([O,S]*)([O,S])[O,S]*

The following SAs have been extracted from the original dataset and are related to compounds which are vP in the soil compartment:

- vP alert no. 1, defined by the SMARTS: c1ccc(c(c1)c2ccccc2Cl)
- vP alert no. 2, defined by the SMARTS: c1c(Cl)cc2Oc3cc(Cl)cc(Cl)c3Oc2c1

4.4.Descriptor selection:

NA

4.5.Algorithm and descriptor generation:

istKNN application

4.6.Software name and version for descriptor generation:

istKNN application (developed by Kode srl, http://chm.kodesolutions.net)

alerts by SarPy

4.7.Chemicals/Descriptors ratio:

568/24 = 24

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.(c.f. section 5.2 below)

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 ≥ 0.65, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index < 0.65, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

The overall architecture provides firstly the prediction as calculated by the k-NN model. If some structural alerts are found, they do not change the prediction but modify the applicability domain value: if the alerts confirm the k-NN prediction, the applicability domain index (ADI) value increases, if the alert are in disagreement with the prediction the ADI value decreases. The alerts are anyway used to provide a prediction if the k-NN model is not able to predict the compound.

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 [9]. 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 > 0.8, strongly similar compounds with known experimental value in the training set have been found

If 0.8 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.6, 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 1 ≥ index > 0.9, accuracy of prediction for similar molecules found in the training set is good

If 0.9 ≥ index > 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≤ 0.5, 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 1 ≥ index > 0.9, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.9 ≥ index > 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

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

Structural Alerts Concordance:

This index takes into account the concordance between the prediction provided by the k-NN model and the alerts found. Defined values are:

If index = 1, all alerts are related to experimental values in agreement with the prediction, thus confirming the k-NN output

If index = 0.9, no alerts have been found, thus it is not possible to confirm the k-NN output

If index = 0.85, no k-NN prediction is available and the final prediction is based only on the found alerts

If index = 0.7, one or more alerts are related to experimental values not in agreement with the prediction, thus conflicting with the k-NN output

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

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:

Leave-one-out approach (k-NN for each compound has been performed on the whole dataset without the compound itself): n = 568; Accuracy = 72%; Non predicted compounds: n = 9

Statistics are also calculated for three cases:

1. Class vP vs All other classes (nP, nP/P, vP/P)

   TP = 53, TN = 482; FP=13; FN =11, not assigned = 9, Accuracy = 96%, Specificity = 97%, Sensitivity = 83%

2. Class nP vs All other classes (vP, vP/P, nP/P)

   TP = 250, TN = 193; FP=64; FN =52, not assigned = 9, Accuracy = 79%, Specificity = 75%, Sensitivity = 83%

3. Classes vP/P + vP vs nP/P + nP

   TP = 101, TN = 404; FP=26, FN =28, not assigned = 9, Accuracy = 90%, Specificity = 64%, Sensitivity = 78%

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:

NA

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:

NA

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:

No assumption on the mechanism is done.

8.2.A priori or a posteriori mechanistic interpretation:

The model has a "a posteriori" mechanistic interpretation based on the descriptors choose

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., “Comparison of two methods for obtaining degradation half lives”, Chemosphere 56, 2004, 531-535

[2] Gramatica, P., Papa, E., “Screening and ranking of POPs for Global Half-Life: QSAR approaches for prioritization based on molecular structure”, Environ. Sci. Technol. 41, 2007, 2833-9

[3] Linders J.B.H.J., Jansma J.W., Mensink B.J.W.G., Otermann K., “Pesticedes: Beneaction or Pandora’s box? A synopsis of the environmental aspects or 351 pesticides.” RIVM Report 679101014, 1994

[4] USGS (Prioritizing Pesticide Compounds for Analytical Methods Development, 2012)

[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] A. Manganaro, F. Pizzo, A. Lombardo, A. Pogliaghi, E. Benfenati, “Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm”, Chemosphere (2015)

[3] Pizzo, Fabiola, Anna Lombardo, Marc Brandt, Alberto Manganaro, and Emilio Benfenati. ‘A New Integrated in Silico Strategy for the Assessment and Prioritization of Persistence of Chemicals under REACH’. Environment International 88 (1 March 2016): 250–60. https://doi.org/10.1016/j.envint.2015.12.019.

[4] OECD. Test Guideline 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

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

[6] Gramatica, Paola, and Ester Papa. ‘Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure’. Environmental Science & Technology 41, no. 8 (1 April 2007): 2833–39. https://doi.org/10.1021/es061773b.

[7] USGS (Prioritizing Pesticide Compounds for Analytical Methods Development, 2012)

[8] Linders J.B.H.J., Jansma J.W., Mensink B.J.W.G., Otermann K., “Pesticedes: Beneaction or Pandora’s box? A synopsis of the environmental aspects or 351 pesticides.” RIVM Report 679101014, 1994

[9] 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

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