	*QMRF identifier (JRC Inventory):* **To be entered by JRC**
	QMRF Title: **Water solubility model (IRFMN) - v. 1.0.1**
	Printing Date: **15-04-2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Water solubility 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:

April 2022

2.2.QMRF author(s) and contact details:

Edoardo Carnesecchi, 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:

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.

2.8.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:

P-CHEM, 4.8 water solubility

3.3.Comment on endpoint:

The modelled endpoint is a physical chemical property.

3.4.Endpoint units:

-Log(mol/L)

3.5.Dependent variable:

Results are valid for a temperature within 10°C-25°C. Prediction is expressed both mg/l and -Log10(mol/L)

3.6.Experimental protocol:

The protocol for the test varied through the decades, including limit test and up-down procedure. No details about the protocol are associated to each experimental data point used for the modelling. However for a general reference to often employed test methods c.f. OECD Test Guideline No. 105 [4] .

3.7.Endpoint data quality and variability:

The model is based on the dataset retrieved from TEST software that consists of 5020 chemicals from the EPISuiteTM database (US EPA. EPI Suite, Version 4.0. US EPA.a, 2019 (accessed 5/21/09)). Chemicals with water solubilities exceeding 1,000,000 mg/L were omitted from the overall dataset. In addition, data were limited to data points that are within 10°C of 25°C. The dataset was split in training (4014 mono constituent organic compounds) and test set (1004 mono constituent organic compounds).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Artificial Neural Network (ANN).

4.2.Explicit algorithm:

The ANN has been trained with the 'nnet' package in the software platform R and exported as a PMML.

4.3.Descriptors in the model:

15 molecular descriptors were calculated within the VEGA descriptors and based on the Dragon software calculation and definition:

ALogP

piPC8

SEigm

MATS1v

PW6

P_VSA_i_4

LogP

nR10

CATS2D_4_DL

BEH2p

BIC3

GATS2m

H-050

D/Dr3

CATS2D_7_DL

4.4.Descriptor selection:

An in-house tool developed in the R statistical platform has been used to select the best descriptors set and size to be employed for the final model. The approach was based on a forward selection technique: starting from the descriptor most correlated with the experimental data, at each iteration the descriptor leading to

the best model (among all the available descriptors) was added, until the size of 25 descriptors. Models have been built, with a linear regression modelling, applied with a bootstrap cross-validation approach (n = 100). For each model, the fitness function has been calculated as the R squared coefficient obtained from the models built in each bootstrap iteration.

This fitness function has been used to select the best descriptor to be added to proceed to the next iteration. From this procedure, the set of descriptors with the best cross-validation values has been chosen as the set to be used for the final model. The "best" values have been considered taking into account their trend: by progressively adding descriptors to the model, the cross-validation performances increase until a plateau (and, following, a decrease), which means that the optimal number of descriptors have been reached, and adding further descriptors would lead to over-fitting.

4.5.Algorithm and descriptor generation:

ANN with DRAGON descriptor.

4.6.Software name and version for descriptor generation:

DRAGON (v. 7.0)

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Prof. R.Todeschini - distributed by Talete srl, via Pisani 13, 20124 Milano, Italy

http://www.disat.unimib.it/chm

The selected descriptors have been implemented in VEGA.

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

4.7.Chemicals/Descriptors ratio:

4014/15 = 268

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. below under point 5.2).

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 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 [2]. 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:

No

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:

From EPI Suite database chemicals with water solubilities data were extracted. Chemicals that exceeding 1,000,000 mg/L were omitted from the overall dataset. In addition, data were limited to data points that were measured within 10°C of 25°C.

6.7.Statistics for goodness-of-fit:

Training set: 4014 compounds

R^2: 0.86

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:

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:

YES

7.2.Available information for the external validation set:

CAS N: Yes

Chemical Name: NO

Smiles: YES

Formula: NO

INChI: No

MOL file: No

Source: No

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

No

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:

The predictive ability of each of the QSAR methodologies was evaluated using statistical external validation.

For all methods, the whole dataset was divided randomly into a training set (80% of the overall set) and a test set (20% of the overall set).

The prediction accuracy was evaluated in terms of the standard performance measures: R^2 and RMSE (root mean square error), for the test set.

For mathematical details, see the reference in section 7.7.

7.7.Predictivity - Statistics obtained by external validation:

Test set: 1004 compounds

R^2: 0.83

RMSE: 0.93

Test set in AD: n = 341; R^2 = 0.91; RMSE = 0.69

Test set could be out of AD: n = 360; R^2 = 0.85; RMSE = 0.83

Test set out of AD: n = 303; R^2 = 0.55; RMSE = 1.24

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:

A posteriori

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] Martin, T; Toxicity Estimation Software Tool (TEST); U.S. Environmental Protection Agency, Washington, DC, (2016) https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

[2] 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.

[3] 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.

[4] OECD GUIDELIN E FOR THE TESTIN G OF CHEMICALS: OECD TG 105 “Water Solubility” (Column or Shake Flask Method, ver. from 1981 or from 1995, no significant difference).

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