	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Log P model (Meylan/Kowwin) v. 1.1.5**
	Printing Date: 30/05/2022

1.QSAR identifier

1.1.QSAR identifier (title):

Log P model (Meylan/Kowwin) v. 1.1.5

1.2.Other related models:

Version 1.0.1

First official release published in the VEGA platform.

Version 1.0.2

Dataset has been revised, several duplicate compounds were removed. New dataset consists of 2524 molecules. Statistics in the current document have been updated.

Version 1.0.3

This version is updated with the new calculation core (1.0.26) where the VEGA similarity algorithm is slightly changed. The new version considers halogen atoms are really similar, especially Chlorine and Bromine atoms are considered almost the same. The main difference with previous algorithm can be thus seen just for halogenated compounds.

A more precise check for similarity has been introduced for the extraction of experimental values, in order to avoid mismatches (as the similarity index is based on fingerprints, there are some rare cases in which a value equal to 1 does not points to a exactly isomorph compound).

The final assessment has been fixed, in previous version a bug occurred (the final assessment was not consistent with the AD assessment reported in the following sections)

There are NO changes in prediction values, but as similarity is changed and a bug fixed, some differences in AD assessment can be found.

Version 1.0.4

This version is updated with the new calculation core (1.0.27), that generates a graphically renewed PDF report. In this version, the propositions for prediction and assessment are changed, but there are NO changes in their values.

Version 1.1.0

This version is a full update, the logP prediction method is changed from previous version. The logP calculation is based on Meylan method, but ALogP and MLogP values, as calculated in previous versions, are still provided.

Version 1.1.2

This version is updated with the new calculation core (1.1.1) based on a new release of the CDK libraries (1.4.9). These updates can influence the calculation, so there could be some changes in the predictions produced.

The new calculation core implements a new version of the algorithm used for calculating the similarity index. This means that the list of similar molecules given as part of the applicability domain evaluation will often be different from the ones produced by older releases of the model. Furthermore, the applicability domain index (ADI) itself and the final assessment could often be different.

Model statistics in the current guide have been updated with the new values.

Some thresholds for the applicability domain sub-indices have been revised to obtain better performances.

A lower bound of -5.0 log units for calculated logP values has been introduced.

Version 1.1.3

This version is updated with the new calculation core (1.2.0). This update can influence some calculation, in particular similarity evaluation, so there could be some changes in the applicability domain values produced.

A further check of structures and experimental data has been performed, resulting in the removal of some compounds from the original dataset (10,005 compounds) which had inconsistent experimental data.

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:

May 2022

2.2.QMRF author(s) and contact details:

[1]Domenico Gadaleta Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche "Mario Negri", IRCCS domenico.gadaleta@marionegri.it

[2]Emilio Benfenati Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche "Mario Negri", IRCCS emilio.benfenati@marionegri.it

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1]William M. Meylan Syracuse Research Corporation, Merrill Lane, Syracuse, NY 13210.

[2]Philip H. Howard Syracuse Research Corporation, Merrill Lane, Syracuse, NY 13210

[3]Alberto Manganaro Kode srl info@kode-solutions.net

2.6.Date of model development and/or publication:

1995

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

[1] Meylan, W.M. and P.H. Howard, Atom/fragment contribution method for estimating octanol/water partition coefficients. 1995, J. Pharm. Sci. 84: 83-92

[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:

NA

3.Defining the endpoint - OECD Principle 1

3.1.Species:

NA

3.2.Endpoint:

QMRF 1. 6. Octanol-water partition coefficient (Kow) EC A.8 Partition Coefficient (EU methodincludes both shake flask and HPLC)

3.3.Comment on endpoint:

NA

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

Logarithm of octanol/water partition coefficient (log P)

3.6.Experimental protocol:

EC A.8 Partition Coefficient OECD 123 Partition Coefficient (nOctanol/Water): Slow-Stirring Method (2006)

OECD 117 Partition Coefficient (n-octanol/water) HPLC Method (1989)

OECD 107 Partition Coefficient (n-octanol/water); Shake Flask Method(1981 & 1995)

3.7.Endpoint data quality and variability:

The training set available in VEGA was collected mainly from EPI Suite KowWin module (9961 compounds)

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Linear regression based on fragments/atom contribution and correction factors

4.2.Explicit algorithm:

Regression equation based on the hydrophobicity contribution of 120 atom types

The Meylan/Kowwin model in VEGA 1.4.4 is an implementation of the atom fragment contribution (AFC) method described by Meylan et al., 1995. It is a “reductionist” approach and was developed via multiple linear regressions of reliable, experimental log P values. The regressions were performed in two separate and sequential stages. The first regression correlated “atom/fragment values” with log P.

Each nonhydrogen atom (e.g. carbon, nitrogen, oxygen, sulfur) in a structure is a “core” for a fragment; the exact fragment is determined by what is connected to the atom. In some cases, entire functional groups are treated as core “atoms”. A contribution in terms of log P was assigned to each fragment. A second regression correlated “correction factors” to log P. In general, the correction factors are values for various steric interactions, hydrogen bondings, and effects from polar functional substructures. Their values are derived from the differences between the log P estimates from fragments alone and the measured log P values. The log P of a compound is then estimated by simply summing all atoms/fragment values and correction factors contained in a structure. The equation reported in Meylan et al., 1995 that describes the log P is the following:

$\text{logP} = \text{Sum}_i\,(f_i\,n_i) + \text{Sum}_i\,(c_i\,n_i) + 0.229$

Where f_i are coefficients for various atom/fragments identified and c_j are correction factors.

4.3.Descriptors in the model:

See 4.2

4.4.Descriptor selection:

See 4.2

4.5.Algorithm and descriptor generation:

See 4.2

4.6.Software name and version for descriptor generation:

NA

4.7.Chemicals/Descriptors ratio:

8364/120 = 69.7

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 and 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 ≥ index > 0.75, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index ≤ 0.75, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

No ADI threshold is used to provide performance calculations of the validation set

5.2.Method used to assess the applicability domain:

The Applicability Domain 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 [2]. 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 > 0.9, strongly similar compounds with known experimental value in the training set have been found

If 0.9 ≥ index > 0.75, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.75, 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 < 0.5, accuracy of prediction for similar molecules found in the training set is good

If 1 > index ≥ 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

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

If 1 > index ≥ 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1, 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 < 0.5, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1 > index ≥ 0.5, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

5.3.Software name and version for applicability domain assessment:

VEGA

Included in the VEGA software and automatically displayed when running the model

emilio.benfenati@marionegri.it

https://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: No

Formula: Yes

INChI: No

MOL file: 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:

Two different validation statistics are reported here:

1) Validation performance on a training set as reported in the reference publication (Meylan & Howard, 1995) (2351 compounds)

2) Validation performance of EPISuite implementation (KOWWIN module) referring to a training set of 2447 compounds

6.6.Pre-processing of data before modelling:

1) Performance reported by Meylan & Howard, 1995 (see 6.7) refers to a dataset of 2351 compounds. The dataset was built from an initial dataset of 8406 unique, mono-constituent organic chemicals with measured logP values retrieved from reliable sources. This dataset was split into a training set (2351 compounds) and a test set (6055). Compounds with simpler structures were put in the training set while other were included in the test set. The dataset is currently not available.

2) The KOWWIN training set is made of 2447 compounds and it can be downloaded from the Internet at: http://esc.syrres.com/interkow/KowwinData.htm. Substructure searchable formats of the data can be downloaded at: http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm

6.7.Statistics for goodness-of-fit:

The following statistics are to be referred to 1) the original model [1] not implemented in VEGA; 2) The models implemented in VEGA using the dataset not more available:

1) Meylan et al., 1995 reported an R^2= 0.982 on their training set of 2351 compounds.

2) KOWWIN User's Guide reports the following statistics on the KOWWIN training set (2447 compounds): number in dataset = 2447 correlation coef (r2) = 0.982 standard deviation = 0.217

absolute deviation = 0.159

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 RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: Yes

INChI: No

MOL file: 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:

Three external validations have been performed: 1) Validation performed on a test set of 6055 compounds as reported by Meylan& Howards, 1995. 2) Validation of the KOWWIN implementation performed on a test set of 10,946 compounds. 3) Validation of VEGA implementation of the model, performed on a dataset of 9,961 compounds

7.6.Experimental design of test set:

1) The dataset from Meylan & Howard 1995 was built from an initial dataset of 8406 unique organic chemicals with mesured logP values retrieved from reliable sources. This dataset was split into a training set (2351 compounds) and a test set (6055). Compounds with simpler structures were put in the training set while other were included in the test set. The dataset is currently not available.

2) The KOWWIN training set is made of 10,946 compounds and it can be downloaded from the Internet at: http://esc.syrres.com/interkow/KowwinData.htm.

Substructure searchable formats of the data can be downloaded at: http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm

3) The dataset of compounds used for validate VEGA implementation has been built starting from the original dataset provided in EPI suite Meylan/Kowwin model. The set has been processed and cleared from compounds that were replicated or that had problems with the provided molecule structure. The final dataset has 9,961 compounds and is freely available for download from model's documentation included in VEGA v. 1.1.4

7.7.Predictivity - Statistics obtained by external validation:

1) Meylan et al., 1995 reported an R2 = 0.943 on a validation set of 6055 compounds.

2) KOWWIN User's Guide reports the following statistics on the KOWWIN training set (10,946 compounds) number in dataset = 10946 correlation coef (r2) = 0.943 standard deviation = 0.479 absolute deviation = 0.356 3)

3) Being the training set no more available, the dataset from EPI Suite KowWin module (9,961 compounds) has been uploaded in VEGA documentation as “training set”. The performances of models to predict these substances are the following:

Dataset set: n = 9961; R^2= 0.86; RMSE = 0.76**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:

Fragments were selected based on their correlation to log P

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:

Meylan's rules have been defined on an old dataset that is no longer available in the literature [1], so IRFMN decided for the implementation of VEGA to insert as "training" a more recent dataset used for the validation of the rules themselves (9961 substances)

9.2.Bibliography:

[1] Meylan, W.M. and P.H. Howard, Atom/fragment contribution method for estimating octanol/waterpartition coefficients. 1995, J. Pharm. Sci. 84: 83-92

[2] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39. doi: 10.1186/s13321-014-0039-1. PMID: 25383097; PMCID: PMC4212147.

[3] OECD. Test No. 123 (2006): Partition Coefficient (1-Octanol/Water): Slow-Stirring Method; Organisation for Economic Co-operation and Development: Paris

[4] OECD. Test No. 117 (1989): Partition Coefficient (n-Octanol/Water), HPLC Method; Organisation for Economic Co-operation and Development: Paris

[5] OECD. Test No. 107 (1981 & 1995): Partition Coefficient (n-Octanol/Water): Shake Flask Method; Organisation for Economic Co-operation and Development: Paris

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