QMRF identifier (JRC Inventory): **To be entered by JRC**
QMRF Title: **VEGA KOC Model (IRFMN) V.1.0.1**
Printing Date: November 2022

1.QSAR identifier

1.1.QSAR identifier (title):

VEGA KOC Model (IRFMN) V.1.0.1

1.2.Other related models:

NA

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:

November 2022

2.2.QMRF author(s) and contact details:

[1] Edoardo Carnesecchi Istituto di Ricerche Farmacologiche Mario Negri IRCCS edoardo.carnesecchi@marionegri.it

[2]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri IRCCS emilio.benfenati@marionegri.it

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

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

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] Kamel Mansouri mansourikamel@gmail.com

2.6.Date of model development and/or publication:

2016

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

[1] An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modeling. 2016. Kamel Mansouri, Chris M. Grulke, Ann M. Richard, Richard S. Judson and Antony J. Williams. SAR & QSAR in Environ. Res; Vol. 27 , Iss. 11,2016.

[2] OPERA: A free and open source QSAR tool for physicochemical properties and environmental fate predictions. Kamel Mansouri, Chris Grulke, Richard Judson, Antony Williams, Journal of Cheminformatics (2018)

[3] 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:

Yes - please refer to Mansouri et al. 2016

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Not applicable

3.2.Endpoint:

ENV FATE 5.4.1. Adsorption / Desorption

3.3.Comment on endpoint:

The soil adsorption coefficient is the ratio of the amount of chemical adsorbed per unit weight of organic carbon in the soil or sediment to the concentration of the chemical in solution at equilibrium (in L/Kg)

3.4.Endpoint units:

Log L/Kg

3.5.Dependent variable:

LogKoc

3.6.Experimental protocol:

Experimental protocols of the different parts of data can be traced back to the original referenced literature from the database [1]

3.7.Endpoint data quality and variability:

The experimental data were downloaded from the EPI Suite data webpage(http://esc.syrres.com/interkow/EpiSuiteData.htm). These data are from PHYSPROP (The Physical Properties Database) which is a collection of a wide variety of sources built by Syracuse Research Corporation (SRC). https://www.srcinc.com/services/engineering-operational-and-environmental-services/scientific-databases.html

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model provides a quantitative estimation of the organic carbon-water partition co-efficient (KOC). The soil adsorption coefficient is the ratio of the amount of chemical adsorbed per unit weight of organic carbon in the soil or sediment to the concentration of the chemical in solution at equilibrium (in L/Kg). It is based on k nearest neighbor (k-NN) model and it is an implementation of the same model present in OPERA

4.2.Explicit algorithm:

Distance weighted k-nearest neighbors (kNN) k=5

This is a refinement of the classical k-NN classification algorithm where the contribution of each of the k neighbors is weighted according to their distance to the query point, giving greater weight to closer neighbors. The used distance is the Euclidean distance. kNN is an unambiguous algorithm that fulfills the transparency requirements of OECD principle 2 with an optimal compromise between model complexity and performance

4.3.Descriptors in the model:

[1]CrippenLogP, Unitless, Estimate of partion coefficients. Wildman, S. A., and Crippen, G. M.(1999). Prediction of Physicochemical Parameters by Atomic Contributions. J Chem Inf Comput Sci39, 868-873.

[2]nHBAcc, Unitless, Number of hydrogen bond acceptors (using CDKHBondAcceptorCountDescriptor algorithm)

[3]SpDiam_Dzi, Unitless, Barysz matrix: Spectral diameter from Barysz matrix / weighted by first ionization potential. Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 714-726

[4]VP-1, Unitless, Chi path: Valence path, order 1. Kier, L. B., and Hall, L. H. (1976). Molecular connectivity in chemistry and drug research, (New York: Academic Press).

[5]MPC3, Unitless, Path counts: Molecular path count of order 3. Todeschini, R. and Consonni, V.(2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 574-579, pg 395402

[6]TPC, Unitless, Path counts: Total path count (up to order 10). Todeschini, R. and Consonni, V.(2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 574-579, pg 395402

[7]VABC, Unitless, Van der Waals volume: Van der Waals volume calculated using the method proposed in [Zhao, Yuan H. and Abraham, Michael H. and Zissimos, Andreas M., Fast Calculation of van der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds, The Journal of Organic Chemistry, 2003, 68:7368-7373]

[8]SsssN, Unitless, Sum of atom-type E-State: >N-

[9]topoDiameter, Unitless, Topological diameter (maximum atom eccentricity)

[10]AATS1m, Unitless, Average Broto-Moreau autocorrelation - lag 1 / weighted by mass. Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim:Wiley VCH) pg 27-37

[11]MPC6, Unitless, Path counts: Molecular path count of order 6. Todeschini, R. and Consonni, V.(2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 574-579, pg 395402

[12]ATS1v, Unitless, Broto-Moreau autocorrelation - lag 1 / weighted by van der Waals volumes. Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim:Wiley VCH) pg 27-37

4.4.Descriptor selection:

PaDEL software was used to calculate 1440 molecular descriptors. A first filter was applied in order to remove descriptors with missing values, constant and near constant (standard deviation of 0.25 as a threshold) and highly correlated descriptors (96% as a threshold). The remaining693 descriptors were used in a feature selection procedure to select a minimum number of variables encoding the most relevant structural information to the modeled endpoint. This step consisted of coupling Genetic Algorithms (GA) with the weighted kNN algorithm and was applied in 5 fold cross validation on the training set (202 chemicals). This procedure was run for 200 consecutive independent runs maximizing Q 2 in cross-validation and minimizing the number of descriptors. The number of k neighbors is optimized within the range of 3 to 7. The descriptors were then ranked based on their frequency of selection during the GA runs. The best model showed an optimal compromise between the simplicity (minimum number of descriptors) and performance (Q2 in cross-validation) to ensure transparency and facilitate the mechanistic interpretation as required by OECD principles 2 and 5. More details in paper [1]

4.5.Algorithm and descriptor generation:

PaDEL descriptors were calculated based on two-dimensional (2D) chemical structures generated by the Indigo cheminformatics suite of tools implemented in KNIME. 2D descriptors were selected over 3D to avoid complicated and usually irreproducible geometrical optimizations. The calculated descriptors fall into different groups such as constitutional indices, ring descriptors, topological indices, 2D matrix-based descriptors, functional group counts and atom counts. Details and references provided in Section 4.3

4.6.Software name and version for descriptor generation:

PaDEL-Descriptors V2.21

An open source software to calculate molecular descriptors and fingerprints. Chun Wei Yap (phayapc@nus.edu.sg) http://padel.nus.edu.sg/software/padeldescriptor

4.7.Chemicals/Descriptors ratio:

405 chemicals (training set) / 12 descriptors = 33.75

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.

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 [3]. 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 > 0.75, strongly similar compounds with known experimental value in the training set have been found

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

If 1.0 > index ≥ 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1.0, 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 1.0 > index ≥ 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1.0, 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 < 0.5, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.0 > 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 ≥ 1.0, 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 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

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: 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:

The dataset used to develop the model implemented on VEGA was not splitted in training and test sets. The training set consists in the overall dataset, n= 729

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

Original statistics i.e. resulting from Mansouri et al. (2016)

Performance in training: n = 545 R2=0.81 RMSE=0.54

After the implementation in VEGA:

Training set: n 729, RMSE 0.13, R2 0.98

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

Statistics for model implemented in VEGA

n = 729 R2=0.785 RMSE=0.569

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

Original statistics i.e. resulting from Mansouri et al. (2016)

Performance in 5-fold cross-validation:

n = 545 Q2=0.81 RMSE=0.54

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: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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

All

7.4.Data for the dependent variable for the external validation set:

All

7.5.Other information about the external validation set:

For the original model (Mansouri et al. 2016), the validation set consists of 184 chemicals

7.6.Experimental design of test set:

For the original model (Mansouri et al. 2016), the structures are randomly selected to represent 25% of the available data keeping a similar normal distribution of LogKoc values in both training and test set using Venetian blinds method

7.7.Predictivity - Statistics obtained by external validation:

Original statistics i.e. resulting from Mansouri et al. (2016):

Performance in test: n = 184 R2=0.71 RMSE=0.61

7.8.Predictivity - Assessment of the external validation set:

For the original model (Mansouri et al. 2016), the validation set consisting of 184 chemicals which is equivalent to a third (1/3) of the training set is sufficient for the evaluation of the predictivity of the model and a good representation of the chemical space

7.9.Comments on the external validation of the model:

The choice of proportions between the training set and the validation set as well as the splitting method helped in accurately evaluating the model and covering most of the training set chemical space. This goal was accomplished without the need to do a structural sampling that usually shows over-optimistic evaluation of the predictivity or a complete random selection that risks biasing the evaluation towards a certain region of the chemical space

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model descriptors were selected statistically but they can also be mechanistically interpreted. KOC is the ratio between the concentration of a chemical adsorbed by the soil normalized to soil organic carbon and the concentration dissolved in the soil water. Thus soil sorption is closely related to water solubility and logP. Therefore, the chemical features which determine the soil sorption are similar to those related to water solubility and logP. In particular, size related descriptors since larger compounds tend to have higher soil sorption because they do have lower water solubility. Also electronic profile descriptors related to charges and to charge distribution are of high importance: the presence of an active functional group next to carbon leads to better water solubility, likewise higher polarity leads to better watersolubility

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori

8.3.Other information about the mechanistic interpretation:

For more details and full reference, see references in Section 4.3   and Section 9.2

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1]James R. Baker, James R. Mihelcic, Aleksandar Sabljic, Reliable QSAR for estimating Koc forpersistent organic pollutants: correlation with molecular connectivity indices, Chemosphere, Volume45, Issue 2, October 2001, Pages 213-221 http://www.sciencedirect.com/science/article/pii/S0045653500003398

[2]Mansouri, K., Grulke, C. M., Judson, R. S., & Williams, A. J. (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. Journal of cheminformatics, 10(1), 10.https://link.springer.com/article/10.1186/s13321-018-0263-1

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

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