	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **VEGA Henry's Law model (IRFMN) - v.1.0.1**
	Printing Date: **November 2022**

1.QSAR identifier

1.1. QSAR identifier (title):

VEGA Henry's Law 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 Via Mario Negri 2,20156 Milano, Italy edoardo.carnesecchi@marionegri.it

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

Alberto Manganaro Kode srl info@kode-solutions.net

2.6. Date of model development and/or publication:

The model was developed in 2019.

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, RichardS. Judson and Antony J. Williams. SAR & QSAR in Environ. Res; Vol. 27 , Iss. 11,2016.http://www.qsar2016.com/program

[2]OPERA: A free and open source QSAR tool for physicochemical properties and environmentalfate predictions. Kamel Mansouri, Chris Grulke, Richard Judson, Antony Williams, Journal ofCheminformatics (2017) https://jcheminf.biomedcentral.com/articles/

[3] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, andScreening 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:

Not applicable

3.2. Endpoint:

ENV FATE 5.4.2. Henry's Law constant

3.3. Comment on endpoint:

Henry’s Law is defined that at a constant temperature, the amount of a given gas that dissolves in a given type and volume of liquid is directly proportional to the partial pressure of that gas in equilibrium with that liquid

3.4. Endpoint units:

Log atm-m3/mole

3.5. Dependent variable:

LogHL

3.6. Experimental protocol:

The data were downloaded from OPERA software tool

3.7. Endpoint data quality and variability:

Data as from OPERA software tool (https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NCCT&dirEntryId=340233)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

The Henry’s law model is a QSAR/QSPR model derived from weighted nearest neighbors algorithm (kNN)based on PaDEL descriptors.

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] nHBDon, Unitless, Hbond donor count: Number of hydrogen bond donors (using CDKHBondDonorCountDescriptor algorithm).

[2] MLFER_S, Unitless, Molecular linear free energy relation: Combined dipolarity/polarizability. Platts JA, Butina D, Abraham MH, Hersey A. Estimation of molecular free energy relation descriptors using a group contribution approach. J Chem Inf Comput Sci. 1999;39(5):835-45.

[3] GATS1e, Unitless, Geary autocorrelation - lag 1 / weighted by Sanderson electronegativities.

Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim:Wiley VCH) pg 27-37

[4] ndssC, Unitless, Atom type electrotopological state: Count of atom-type E-State: =C<. Hall, L. H.,and Kier, L. B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 35, 1039-1045; Liu,R., Sun, H., and So, S. S. (2001). Development of quantitative structure-property relationship models for early ADME evaluation

in drug discovery. 2. Blood-brain barrier penetration. J Chem Inf Comput Sci 41, 1623-1632.; Gramatica, P., Corradi, M., and Consonni, V. (2000). Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. Chemosphere 41, 763-777.

[5] ATS3m, Unitless, Broto-Moreau autocorrelation - lag 3 / weighted by mass.

Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 27-37

[6] nHBint6, Unitless, Atom type electrotopological state: Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 6. Hall, L. H., and Kier, L. B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 35, 1039-1045; Liu, R., Sun, H., and So, S. S. (2001).

[7] nHBAcc2, Unitless, Number of hydrogen bond acceptors (any oxygen; any nitrogen where the formal charge of the nitrogen is non-positive (i.e. formal charge <= 0) except a non-aromatic nitrogen that is adjacent to an oxygen and aromatic ring, or an aromatic nitrogen with a hydrogen atom in a ring, or an aromatic nitrogen with 3 neighboring atoms in a ring, or a nitrogen with total bond order>=4; any fluorine).

[8] AATSC0i, Unitless, Average centered Broto-Moreau autocorrelation - lag 0 / weighted by first ionization potential. Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 27-37

[9] SpAD_Dzm, Unitless, Barysz matrix: Spectral absolute deviation from Barysz matrix / weighted by mass. Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 714-726

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 remaining765 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 (441 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.

4.5. Algorithm and descriptor generation:

VEGA platform, where the descriptors have been implemented following the definition available in the Dragon software

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)

4.7. Chemicals/Descriptors ratio:

NA

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 [4]. 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

The VEGA software provides QSAR models to predict tox, ecotox, environ, and phys-chemproperties of chemical substances. emilio.benfenati@marionegri.it https://www.vegahub.eu/

5.4. Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counter ion and 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= 591

The dataset downloadable in VEGA contain prediction of original model.

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. (2018) Performance in training: R2= 0.84 RMSE= 1.91

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

Statistics for model implemented in VEGA n= 591 R2= 0.86 RMSE= 0.78

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

Original statistics i.e. resulting from Mansouri et al. (2018) Performance in 5-fold cross-validation: Q2=0.84 RMSE=1.96

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: No

INChI: No

MOL file: No

NanoMaterial: 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:

In the original model (Mansouri et al. 2018), the validation set consists of 150 chemicals. The values are ranging from ~-10 to ~0.5

7.6. Experimental design of test set:

NA

7.7. Predictivity - Statistics obtained by external validation:

Original statistics i.e. resulting from Mansouri et al. (2018) Performance in test: R2=0.85 RMSE=1.82

7.8. Predictivity - Assessment of the external validation set:

NA

7.9. Comments on the external validation of the model:

Two external validations were performed based on a test set and an exernal validation set (see 7.6).

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. Henry's Law definition: the mass of gas dissolved by a given volume of solvent is proportional to the pressure of the gas with which it is in equilibrium. So Henry's law constant is a measure of the relative affinity of a compound for the vapor phase and water. H depends mainly on interactions in the aqueous phase because in the gas phase, behavior is close to ideal. Interactions with water molecules is a constitutive property of the molecule and can involve hydrogen bonding and dipole-dipole, dipole-induced dipole, ion dipole, and ion-induced dipole interactions, which are all exoergic

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] Dunnivant FM, Elzerman AW (1988). Aqueous solubility and Henry's law constant data for PCBcongeners for evaluation of quantitative structure-property relationships (QSPRs). Chemosphere 17,525–541

[2] Dunnivant FM, Elzerman AW, Jurs PC, Hasan MN (1992). Quantitative structure-property relationships for aqueous solubilities and Henry's law constants of polychlorinated biphenyls. Environ Sci Technol 26, 1567–1573.

[3]Dearden JC, Schüürmann G (2003) Quantitative structure-property relationships for predicting henry's law constant from molecular structure 22 (8) 1755–1770

[4] 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