	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Fathead Minnow LC50 96h (EPA) (version 1.0.8)**
	Printing Date: **20-10-2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Fathead Minnow LC50 96h (EPA) (version 1.0.8)

1.2.Other related models:

Version 1.0.1

First official release published in the VEGA platform.

Version 1.0.2

Prediction is now provided also in mg/L units.

This version is updated with the new calculation core (1.0.23) where 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 an exact isomorph compound).

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

Version 1.0.3

This version is updated with the new calculation core (1.0.26). Some minor bugs in the procedure for reading molecule structures have been fixed; some compounds, previously not loaded, could now be correctly processed. All values are now given with explicit unit of measurement. Also the experimental value (if available) is now provided in mg/L units.

There are NO changes in prediction values, but as similarity is changed some small 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 1.0.6

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.

Version 1.0.7

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.

The VEGA model has several differences that can lead to prediction that can be slightly different from the ones produced by the original US EPA T.E.S.T. model. Mainly, the values of descriptors could be different for some molecules, even if the algorithm definition is the same, due to some different compound's preprocessing (like for the detection of aromaticity).

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

T.E.S.T. software for US EPA

It is a linear regression made on 21 molecular descriptors - Todd Martin

https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

2. General information

2.1. Date of QMRF:

October 2022

2.2. QMRF author(s) and contact details:

[1] 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] Erika Colombo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy erika.colombo@marionegri.it https://www.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] Todd Martin Research Chemical Engineer in EPA's National Risk Management Research Laboratory martin.todd@epa.gov

[2] 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:

[1] Martin, T.M., and D.M. Young. (2001). "Prediction of the Acute Toxicity (96-h LC50) of Organic Compounds in the Fathead Minnow (Pimephales Promelas) Using a Group Contribution Method." Chemical Research in Toxicology, 14, 10: 1378–1385

[2] 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:

The fish Fathead Minnow (Pimephales promelas).

3.2.Endpoint:

ECOTOX 6.1.1. Short-term toxicity to fish

3.3.Comment on endpoint:

The fathead minnow LC50 endpoint represents the concentration in water, which kills half of fathead minnow (Pimephales promelas) in 4 days (96hours).

3.4.Endpoint units:

LC50 in –log(mol/L) and its conversion in mg/L

3.5.Dependent variable:

LC50 96h

3.6.Experimental protocol:

OECD TG 203 “Fish, Acute Toxicity Test” (1981, 1984 & 1992) [6]

3.7.Endpoint data quality and variability:

The data set for this endpoint was obtained by downloading the ECOTOX aquatic toxicity database 31.

The database was then filtered using the following criteria:

The ECOTOX “Media Type” field = “FW” (fresh water);

The ECOTOX “Test Location” field = “Lab” (laboratory);

The ECOTOX “Conc 1 Op (ug/L)” field cannot be <, >, or ~ (i.e. use only discrete LC50 values)

The ECOTOX “Effect” field = “Mor” (mortality)

The ECOTOX “Effect Measurement” field =“MORT” (mortality)

The ECOTOX “Exposure Duration” field = “4” (4 daysor 96 hours)

Compounds can only contain the following element symbols: C, H, O, N, F, Cl, Br, I, S, P, Si, As

Compounds must represent a single pure component (i.e. salts, undefined isomeric mixtures, polymers, or mixtures were removed)

The LC50 values were taken from the “Conc 1 (ug/L)” field in ECOTOX.

For chemicals with multiple LC50 values, the median value was used.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model provides a quantitative prediction for fathead minnow (Pimephales promelas) EC50 (96 hour), given in -log(mol/l) and its conversion in mg/L. It is implemented inside the VEGA online platform, accessible at: http://www.vegahub.eu/

The model is a re-implementation of the *single model* developed by Todd Martin inside T.E.S.T. software for US EPA. The T.E.S.T. software is freely available at: https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

4.2.Explicit algorithm:

Linear regression based on 21 descriptors

The regression coefficients have been calculated on the T.E.S.T. original dataset, that contains 816 compounds extracted from the ECOTOX aquatic toxicity database (http://cfpub.epa.gov/ecotox/),splitted in 652 compounds for the training set and 164 for the test set. More details on the model can be found in the T.E.S.T. documentation: https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test#install

4.3.Descriptors in the model:

[1] SsssN: Sum of >N- E-States

[2] SdssS_acnt: Count of =S<

[3] MDEC33:Molecular distance edge between all tertiary carbons

[4] MDEO11: Molecular distance edge between all primary oxygens

[5] BEHm2: Highest eigenvalue n.2 of Burden matrix / weighted by atomic masses

[6] nDB: Number of double bonds - nS: Number of Sulfur atoms

[7] nR09: Number of 9-membered rings

[8] ATS5p: Broto-Moreau autocorrelation of a topological structure

[9] lag 5 /weighted by atomic polarizabilities

[10] MATS7e: Moran autocorrelation

[11] lag 7 / weighted by atomic Sanderson electronegativities

[12] GATS3e: Gearyautocorrelation

[13] lag 3 / weighted by atomic Sanderson electronegativities

[14] SRW03: Self-returning walk count of order 3

[15] ALOGP: Ghose-Crippen octanol water coefficient

[16] NO2 [aromatic attach]: -NO2 [aromatic attach] fragment count

[17] CHO [aliphatic attach]: -CHO [aliphatic attach] fragment count

[18] OH [phosphorus attach]: -OH[phosphorus attach] fragment count

[19] >NN=O: >NN=O fragment count

[20] C(=O)- [2 nitrogen attach]: -C(=O)- [2 nitrogen attach] fragment count

[21] C#N [aliphatic attach]: -C#N [aliphatic attach] fragment count

[22] Cl [olefinic attach]: -Cl [olefinic attach] fragment count

[23] CHO [aromatic attach]: -CHO [aromatic attach] fragment count

4.4.Descriptor selection:

The original code for descriptors calculation has been kindly provided by Todd Martin

4.5.Algorithm and descriptor generation:

Linear regression based on 21 descriptors

The original code for descriptors calculation, provided by Todd Martin, was entirely re-calculated by an in-house software module in which they are implemented as described in Todeschini and Consonni, 2009 (see [1] in section 9.2)

4.6.Software name and version for descriptor generation:

The original code for descriptors calculation has been kindly provided by Todd Martin

4.7.Chemicals/Descriptors ratio:

816/21 = 39

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 ≥ 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 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 [5]. 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 > 0.85, 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 atoms 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: 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:

The final fathead minnow EC50 data set contained 823 mono-constituent organic chemicals. For use in QSAR modeling, the experimental values in mg/L were converted to –Log10 (EC50 mol/L). For single model, the data set were divided randomly into a training set (80% of the overall set) and a test set (20% of the overall set).

6.6.Pre-processing of data before modelling:

See section 6.5

6.7.Statistics for goodness-of-fit:

Training set: RMSE = 0.83, R^2 = 0.69, n = 652

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:

External validation set not available.

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:

Calibration set: RMSE =0.89, R^2 = 0.63, n = 164

Test set in AD: n = 30; R^2 = 0.47; RMSE = 0.68

Test set could be out of AD: n = 75; R^2 = 0.69; RMSE = 0.69

Test set out of AD: n = 58; R^2 = 0.69; RMSE = 0.69

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:

Mechanism it´s associated to the combination of the chosen descriptors (see [2], [3], [4] in *section* 9.2)

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] R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009, DOI:10.1002/9783527628766

[2] Martin, T. M.; Young, D. M. Prediction of the Acute Toxicity (96-h LC50)of Organic Compounds ti the Fathead Minnow (Pimephales promelas) Using aGroup Contribution Method. Chem. Res. Toxicol. 2001, 14, 1378-1385

[3] Martin, T. M.; Grulke, C. M.; Young, D. M.; Russom, C. L.; Wang, N. Y.;Jackson, C. R.; Barron, M. G. Prediction of Aquatic Toxicity Mode ofAction Using Linear Discriminant and Random Forest Models. J. Chem. Inf.Model. 2013, 53, 2229-2239.

[4] Martin, T. M.; Young, D. M.; Lilavois, C. R.; Barron, M. G. Comparisonof global and mode of action-based models for aquatic toxicity. SAR QSAREnviron. Res. 2015, 26, 245-262.

[5] 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.

[6] OECD (1981, 1984 and 1992), Test No. 203: Fish, Acute Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, https://www.oecd.org/env/ehs/testing/section2-effects-on-biotic-systems-replaced-and-cancelled-test-guidelines.htm

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