	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Daphnia Magna Chronic (NOEC) toxicity model (IRFMN) version1.0.0**
	Printing Date: **Mar 1, 2020**

1.QSAR identifier

1.1.QSAR identifier (title):

Daphnia Magna Chronic (NOEC) toxicity model (IRFMN) version1.0.0

1.2.Other related models:

Algae (*Raphidocelis subcapitata*, ex *Pseudokirchneriella subcapitata*): EC50 72h (growth rate)

Algae (*Raphidocelis subcapitata*, ex *Pseudokirchneriella subcapitata*): NOEC 72h (growth rate)

Daphnids (Daphnia magna): EC50 48h, acute (immobilisation)

Daphnids (*Daphnia magna*): NOEC 21d, chronic (reproduction)

Fish (*Oryzias latipes*): LC50 96h, acute (mortality)

Fish (*Oryzias latipes*): NOEC, chronic (ELS-test).

1.3.Software coding the model:

VEGAHUB

https://www.vegahub.eu/contact/

https://www.vegahub.eu/

2.General information

2.1.Date of QMRF:

2.2.QMRF author(s) and contact details:

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

IRFMN IRFMN https://www.vegahub.eu/contacts/ https://www.vegahub.eu/contacts/

2.6.Date of model development and/or publication:

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

VEGA https://www.vegahub.eu

2.8.Availability of information about the model:

2.9.Availability of another QMRF for exactly the same model:

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnids (*Daphnia magna*)

3.2.Endpoint:

ECOTOX 6.1.4. Long-term toxicity to aquatic inverterbrates

3.3.Comment on endpoint:

3.4.Endpoint units:

3.5.Dependent variable:

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The Daphnia Chronic (NOEC) toxicity model (IRFMN) –v.1.0.0 is based on 306 experimental data retrieved from the Japanese Ministry of Environment (http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2016.pdf), and selected according to the OECD TG 211 requirements. The model is a Tree Ensemble Random Forest.

4.2.Explicit algorithm:

The model is a Tree Ensemble Random Forest.

Among the several algorithms used, we obtained the best results in terms of performance with a Random Forest called Tree ensemble. Tree ensemble builds a series of regression trees with different rows and different variables (according to certain parameters) and then it aggregates the results as an ensemble of models. It chooses the parameters for the variables of each tree and the number of compounds evaluating the performance of several models (Hyperparameter tuning Research) using as metric R^2 of a Bootstrap (100 iterations) cross-validation on training set.

4.3.Descriptors in the model:

dragon 7.0

4.4.Descriptor selection:

Dragon 7.0 extension for KNIME has been used to calculate the descriptors, resulting in 3839 2D descriptors. Then we applied a pruning process both to the compounds and to the descriptors pools. Firstly, we removed the compounds for which it was not feasible to calculate AlogP (Ghose-Crippen octanol-water partition coefficient (Ghose and Crippen, 1986; Viswanadhan et al., 1993; Ghose et al., 1998)), as it is generally well acknowledged that this descriptor is the most correlated to the response. Then, to reduce the great number of variables, we removed all the descriptors with constant values (var(X) =0), or which correlate over 0.95 (Pearson) with at least one another descriptor.

In order to select the variables we used two methods implemented in R packages for each dataset: the genetic algorithm (gaselect package) and the Variable Selection Using Random Forest (VSURF) package. We imported both the pools of variables of each dataset into a KNIME workflow to derive the models.

4.5.Algorithm and descriptor generation:

4.6.Software name and version for descriptor generation:

DRAGON

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

4.7.Chemicals/Descriptors ratio:

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

5.2.Method used to assess the applicability domain:

5.3.Software name and version for applicability domain assessment:

5.4.Limits of applicability:

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:

Splitting the training and the test sets

To derive the models, we divided the data in training and test sets with the ratio of 80:20. In order to obtain a uniform distribution of the endpoint values between the two subsets we applied an activity and descriptors sampling method. We performed a Principal Component Analysis (PCA) on all the descriptors and we selected the first two principal components. We selected five random compounds, and then we picked the most dissimilar compound from the sample pool according to the first two principal components and the response using several combinations of distance metrics and scoring functions. Then we added the compound to the pool repeating the operation until we reached the desired number for the training set.

6.6.Pre-processing of data before modelling:

SMILES creation and neutralization

Firstly, we generated the SMILES structures from the chemical name and CAS RN for each substance using ChemCell (2019) and Marvin View (Marvin 17.28.0, 2012017, ChemAxon, 2019). We manually checked the correspondence and correctness among the obtained structures, chemical name and CAS RN among several websites and public database like

ChemIDplus Advanced (NIH, 2019), PubChem (NCBI, 2019), ChemSpider (Royal Society of Chemistry, 2019), DSSTox. Then, we added several structures, which have not automatically generated.

We normalized the SMILES with istMolBase 1.0.3. (in-house software), then we neutralized them using KNIME 3.5. Since pH is a critical issue in the experimental assays on algae, we considered ionized normalized SMILES and we calculated the major microspecies at pH 7.5 and 8.1 using JChem for Excel. We removed the compounds for which the SMILES changed depending on pH (in range 7.5-8.1).

Cleaning of the structure

We cleaned the datasets excluding the following compounds:

metal complexes

inorganicsmixtures of structural isomers

ambiguous structures

non-ionic surfactant mixturescomplex disconnected structures (e.g. polymers)chemicals whose correspondence name-CAS was not foundUVCBsalts; only the acid form was kept

Values cleaning

We selected continuous experimental values excluding those reported as a range, greater/less than a certain threshold, or approximate values. We converted each experimental value from mg/l to mmol/l, on the basis of the molecular weight calculated from the chemical structure. We also removed the compounds for which the experimental toxicity values were higher than the experimental water solubility values. For this pourposewe retrieved the experimental water solubility values mainly from a large database of more than 4,000 chemicals that we pruned in the LIFE project ANTARES and from GuideChem and Sigma-Aldrich websites in the case we did not find the water solubilities elsewhere.

Dealing with multiple values

To deal with multiple continuous data we referred to the procedures described in ECHA guidance R.10 (2008) for ecotoxicological continuous endpoints. In case the experimental conditions and the reliability of the studies were the same, we considered the ratio between the maximum and the minimum values; if it was higher than one log unit we eliminated the data. Then, we calculated the median, the arithmetic and geometric mean in mmol/l to check if there were differences among them. We found a very good correlation (R^2?1) between the values of each combination (arithmetic vs geometric mean, arithmetic mean vs median, geometric mean vs median) and finally the geometric mean was preferred (ECHA guidance R.10, 2008). To normalize the data we

performed two types of transformation,
the logarithm of the geometric mean and the Box-cox transformation (with ? value optimized for each dataset). Since the box-cox transformation gave better results in terms of normalization of the data, it was finally used to normalize the data. We excluded data falling outside the range (mean of the box-cox transformed values) ± 3*(standard deviation).

6.7.Statistics for goodness-of-fit:

Tot RMSE 0.71 R2 0.61 mean obs -2.50 n 307 Training RMSE 0.66 R2 0.64 n 215 mean obs -2.52 Test RMSE 0.81 R2 0.57 n 92 mean obs -2.4

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

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: Yes

INChI: Yes

MOL file: Yes

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:

7.6.Experimental design of test set:

7.7.Predictivity - Statistics obtained by external validation:

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

8.2.A priori or a posteriori mechanistic interpretation:

The mechanistic interpretation of the model is provided a posteriori, i.e. by interpretation of the final set of the selected descriptors.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2. Bibliography:

9.3. Supporting information:

Training set(s)

dataset_DAPHNIA_NOEC_training.csv	file:///C:\Users\Lenovo\Documents\lavoro_QMRF\daphnia_NOEC_IRFMN_1.0.0\dataset_DAPHNIA_NOEC_training.csv

Test set(s)

dataset_DAPHNIA_NOEC_test.csv	file:///C:\Users\Lenovo\Documents\lavoro_QMRF\daphnia_NOEC_IRFMN_1.0.0\dataset_DAPHNIA_NOEC_test.csv

Supporting information

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