	QMRF identifier (JRC Inventory):**Q13-33-0041**
	QMRF Title:**Nonlinear QSAR: artificial neural network for acute aquatic toxicity**
	Printing Date:**Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear QSAR: artificial neural network for acute aquatic toxicity

1.2.Other related models:

1.3.Software coding the model:

ADMET Predictor™ 3.0

Software for estimating certain ADMET (Absorption, Distribution, Metabolism, Elimination, and Toxicity) properties of a drug-like chemical from its molecular structure; 1998-2008; Simulations Plus Inc

Simulations Plus, Inc. 42505 10th Street West Lancaster, CA 93534-7059 USA, Phone: +1.661.723.7723 (international), Toll free: 888.266.9294 (in the U.S. & Canada), Fax: +1.661.723.5524.

http://www.simulations-plus.com/Products.aspx?grpID=1&cID=11&pID=13

2.General information

2.1.Date of QMRF:

25/10/2009

2.2.QMRF author(s) and contact details:

JRC Computational Toxicology Group EC Joint Research Centre, Institute for Health and Consumer Protection Via E. Fermi 2749, 21027 Ispra (VA), Italy JRC-IHCP-COMPUTOX@ec.europa.eu

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Klaus Daginnus klaus.daginnus@bsu.hamburg.de

2.6.Date of model development and/or publication:

21/08/2009 (model development)

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

Software package: ADMET Predictor™ 3.0; Simulations Plus, Inc. 42505 10th Street West Lancaster, CA 93534-7059 USA; http://www.simulations-plus.com/Products.aspx?grpID=1&cID=11&pID=13

2.8.Availability of information about the model:

The training and test sets are available.

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

None to date.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Fathead Minnow (Pimephales promelas)

3.2.Endpoint:

3.Ecotoxic effects 3.3.Acute toxicity to fish (lethality)

3.3.Comment on endpoint:

Acute toxicity to fish.The experiments were performed on juvenile fathead minnows (28 to 36 day old) exposed to test substances via a 96h flow-through system (sect 9.2; ref 2).

3.4.Endpoint units:

Molar 96h lethal concentration (LC50) in fathead minnow was expressed in (mmol/L) and on the decimal logarithmic scale: Log (96h LC50) (mmol/L).

3.5.Dependent variable:

3.6.Experimental protocol:

The experimental protocols of biological/chemical investigations were described by Brooke et al. (sect 9.2; ref 3) and Geiger et al. (sect 9.2; ref 4). Organometallics, inorganic substances and chemicals for which the data were unavailable were excluded.

3.7.Endpoint data quality and variability:

Experimental data on 96h LC50 (mmol/L) in fathead minnow for 577 chemicals were obtained from the DSSTox (Distributed Structure Searchable Toxicity) Database, which originated in the US-EPA Fathead Minnow Acute Toxicity (EPAFHM) Database (sect 9.2; ref 1). The quality of data from DSSTox/EPAFAHM Database was assessed by Russom et al. (sect 9.2; ref 2) Experimental data on 96h LC50 (mmol/L) in fathead minnow for 577 chemicals were obtained from the DSSTox (Distributed Structure Searchable Toxicity) Database, which originated in the US-EPA Fathead Minnow Acute Toxicity (EPAFHM) Database (sect 9.2; ref 1). The quality of data from DSSTox/EPAFAHM Database was assessed by Russom et al. (sect 9.2; ref 2).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Artificial Neural Network

4.2.Explicit algorithm:

Artificial Neural Network

MLP-ANNE - Multilayer Perceptron Artificial Neural Network Ensembles Regression Model

MLP-ANNE model was calculated with ADMET Predictor^{TM}3.0 software. After the procedures of (i) selecting model descriptors (i.e. removing invariant or highly correlated ones and performing sensitivity analysis to find the most relevant combination); (ii) splitting the input data into training pool (303 training set compounds + 173 verification test compounds) and test set (101 compounds) using Kohonen self-organising map (SOM) method; and (iii) training MLP-ANNE for different network architectures, the final model was selected. The best MLP-ANNE model was characterized by the following architecture: 11-3-1 (i.e. by 11 neurons in the input layer (selected molecular descriptors), 3 neurons in the hidden layer and 1 neuron in the output layer [Log (96h LC50), mmol/L]).

4.3.Descriptors in the model:

[1]S+logP octanol-water partition coefficient

[2]SdCH2 atom-type electropological-state index for =CH2 groups

[3]Pi_Q4 derived from electronic properties, 4th component of the autocorrelation vector of Hückel pi atomic charges

[4]F_TpleB constitutional descriptor, triple bonds as fraction of total bonds

[5]PolarizG [3] derived from electronic properties, polarizability calculated by Glen's method

[6]EEM_XFpl derived from electronic properties, maximum sigma Fukui index on polar atoms

[7]N_Bonds constitutional descriptor, number of bonds

[8]SsO- atom-type electropological-state index for coordinated O- groups

[9]SHdsCH atom-type electropological-state index for aCHa groups (aromatic carbons)

[10]StsC atom-type electropological-state index for #C- groups

[11]Sscl atom-type electropological-state index for -Cl groups

4.4.Descriptor selection:

ADMET PredictorTM 3.0 software calculated hundreds of various descriptors for each studied compound. Thus, the pre-selection of "candidate" inputs had to be performed. This procedure aimed to exclude (based on the statistical selection rules) from the initial set of available inputs those which were: (i) identical or had low variance (i. e. coefficient of variation, CV, lower than 1%); (ii) underrepresented (i. e. had non-zero values for less than 4 compounds); (iii) highly correlated (i. e. the correlation between raw descriptors was greater than 0.99999). Removing the latter resulted in the selection of 149 "candidate" inputs. In the next step, in order to find the optimal model complexity (i.e. the best amount of relevant descriptors to ensemble training), the input gradient sensitivity analysis (SA) over all "candidates" was performed. Finally, the set of 11 descriptors was selected.

4.5.Algorithm and descriptor generation:

All the descriptors were calculated with ADMET Predictor^{TM}3.0 software.

4.6.Software name and version for descriptor generation:

ADMET Predictor™ 3.0

Software for estimating certain ADMET (Absorption, Distribution, Metabolism, Elimination, and Toxicity) properties of a drug-like chemical from its molecular structure; 1998-2008; Simulations Plus Inc

Simulations Plus, Inc. 42505 10th Street West Lancaster, CA 93534-7059 USA, Phone: +1.661.723.7723 (international), Toll free: 888.266.9294 (in the U.S. & Canada), Fax: +1.661.723.5524.

http://www.simulations-plus.com/Products.aspx?grpID=1&cID=11&pID=13

4.7.Chemicals/Descriptors ratio:

11/476 = 0.023 (43.3 chemicals per descriptor)

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on the training pool, including 476 defined organic chemicals: 471 single compounds and 5 mixtures of formulation (for details please refer to the supporting files):

(i) AD by chemical classes (FHM): the training pool compounds covered all standard chemical classes from EPAFHM Database (e.g. aliphatic and aromatic hydrocarbons, ethers, alcohols, aldehydes, ketones, amides, aliphatic and aromatic amines, sulfidespridines, barbitals, etc.); these compounds covered different modes of toxic action - the majority of them (200) was associated with baseline narcosis or electrophile/proelectrophile reactivity (82).
(ii) AD by descriptor values range: the model predictions were suitable for compounds characterized by the following descriptor values:
[1] S+logP: min. -4.31; max. 6.77;
[2] SdCH2: min. 0.00; max. 5.42;
[3] Pi_Q4: min. -0.17; max. 0.45;
[4] F_TpleB: min. 0.00; max. 0.50;
[5] PolarizG: min. 3.47; max. 48.81;
[6] EEM_XFpl: min. -0.08; max. 0.45;
[7] N_Bonds: min. 1; max. 35;
[8] SsO-: min. 0.00; max. 30.90;
[9] SHdsCH: min. 0.00; max. 5.42;
[10] StsC: min. 0.00; max. 7.42;
[11] SsCl: min. 0.00; max. 35.69.Experimental (observed) Log (96-h LC50) values for the training pool
compounds varied from min. -6.38 to max. 2.96 mmol/L; for test set compounds from min. -3.25 to max. 2.85 mmol/L.

5.2.Method used to assess the applicability domain:

Applicability Domain (AD) assessment based on the training pool compounds: (i) their chemical identity (i.e. the presence of certain functional groups and their membership in particular chemical classes, e.g. organometalics and inorganic substances were excluded); (ii) the ranges of descriptor values describing the intrinsic properties of studied chemicals - the descriptor values of "predicted" compounds should fall between maximal and minimal descriptor values of the training pool compounds.

5.3.Software name and version for applicability domain assessment:

ADMET Predictor™ 3.0
Software for estimating certain ADMET (Absorption, Distribution, Metabolism, Elimination, and Toxicity) properties of a drug-like chemical from its molecular structure; 1998-2008; Simulations Plus Inc
Simulations Plus, Inc. 42505 10th Street West Lancaster, CA 93534-7059 USA, Phone: +1.661.723.7723 (international), Toll free: 888.266.9294 (in the U.S. & Canada), Fax: +1.661.723.5524.
http://www.simulations-plus.com/Products.aspx?grpID=1&cID=11&pID=13

5.4.Limits of applicability:

The model is suitable for specified chemical classes of compounds that have particular molecular descriptors in specified ranges (p. 5.1). The most sensitive descriptor was octanol-water partition coefficient (S+logP). The values of S+logP for training pool compounds varied from

-4.31 to 6.77 as the applicability domain of the model covers chemicals characterized by different modes of toxic action. Compounds characterized by S+logP values lower than 0 as well as those with S+logP higher than 6 should not be modelled as narcotics – S+logP<0 indicates unrealistically high toxic effects, while S+logP>6 indicates that the uptake of compound from water is too slow to be connected with acute toxicity. It means that the predictions performed by narcosis-type model can be associated with high uncertainty for such compounds.

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

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 presented MLP-ANNE model was developed and internally validated based on the training pool including 476 compounds (303 training set compounds for neural networks training + 173 verification set compounds for internal validation). The algorithm used for training pool selection based on Kohonen self-organizing map (SOM) method.

6.6.Pre-processing of data before modelling:

Transformation of data from 96-h LC50 to logarithmic scale: Log (96-h LC50).

6.7.Statistics for goodness-of-fit:

The MLP-ANNE model's goodness-of-fit was tested against 303 training set compounds:

Coefficient of Multiple Determination: R^2 = 0.755;

Root Mean Squared Error of Calibration: RMSE = 0.699;

Mean Absolute Error: MAE = 0.508.

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

The MLP-ANNE model was internally validated according to 173 verification set compounds. In order to find the best complexity of the model (i.e. determine the moment of stopping the training procedure and avoid overtraining) the verification set errors were monitored (early stopping technique). The finally chosen model was characterized by the following, verification-set based, statistics:

Explained variance in prediction: Q^2= 0.809

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:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: Yes

INChI: Yes

MOL file: Yes

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:

External validation (test) set with 101 compounds appended.

7.6.Experimental design of test set:

The external validation set (i.e. test set) consisted of 101 compounds from the entire data set, selected according to Kohonen Self-Organizing Map (SOM) mathematical method. The composition of test set was determined before the beginning of neural networks training procedure (external validation compounds haven't participated in training). The mapping process based on 11 previously selected descriptors, gathering the structural information on the studied compounds. The size of Kohonen map was 24x24 and all chemicals were clustered into 576 two-dimensional cells of similar structure, indicated by the values of the descriptors.

7.7.Predictivity - Statistics obtained by external validation:

External validation coefficient (based on test set compounds): QEXT2 = 0.715;

Root Mean Squared Error of Prediction (based on test set compounds): RMSE = 0.705;

Mean Average Error (based on test set compounds): MAE = 0.515.

7.8.Predictivity - Assessment of the external validation set:

The application of Kohonen SOM method was used to determine the external validation (test) set, consisting of compounds representing the structural features and toxicological classes of the entire data 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:

As the MLP-ANNE model was developed statistically, no a priori assumptions have been made.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation.

The sensitivity analysis allowed the selection of molecular descriptors giving as much relevant information on the endpoint as possible. The most sensitive one was the octanol-water partition coefficient (S+logP), which is the main mechanistically interpretable descriptor as far as acute aquatic toxicity is concerned. S+logP describes the kinetics of the process of uptaking chemicals from water via lipid membranes and thus indicates a baseline toxicity. Other descriptors represent the structural features of chemicals as well as their electronic properties (e.g. polarizability, presence of polar/certain functional groups or bonds) – they reflect to the polarity and the surface areas of compounds that can be available for solvent (water) molecules as well as for lipid membranes of aquatic biota.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

The MLP-ANNE model is an example of the result of non-linear modelling based on the application of sophisticated mathematical and statistical approaches. Since no equation describing the correlations between descriptors and the endpoint can be specified, the only way to transparently present the modelling procedure and its results is to describe it step-by-step in words.

9.2.Bibliography:

[1]US EPA DSSTox (Distributed Structure Searchable Toxicity) Database http://www.epa.gov/NCCT/dsstox/

[2]Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE & Drummond RA (2007). Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (Pimephales Promelas). Environmental Toxicology and Chemistry 16 (5), 948-967.

[3]Brooke LT, Call DJ, Geiger DL & Northcott CE, eds. (1984). Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas), Vol. 1. Center for Lake Superior Environmental Studies, University of Wisconsin, Superior, WI, USA.

[4]Geiger DL, Northcott CE, Call DJ & Brooke LT, eds. (1985). Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas), Vol. 2. Center for Lake Superior Environmental Studies, University of Wisconsin, Superior, WI, USA.

9.3.Supporting information:

qmrf143_Training_303.sdf	http://qsardb.jrc.ec.europa.eu/qmrffile:///C:\Documents and Settings\casadja\My Documents\a_QSAR\a_QMRFdb\model submissions\7 review completed\Q10-23-19-143_2009-11-13_QSAR for acute aquatic toxicity to Pimephales Promelas (Fathead Minnow)_K Dagginus\qmrf143_Training_303.sdf

qmrf143_TrainingPoolTotal_476.sdf	http://qsardb.jrc.ec.europa.eu/qmrffile:///C:\Documents and Settings\casadja\My Documents\a_QSAR\a_QMRFdb\model submissions\7 review completed\Q10-23-19-143_2009-11-13_QSAR for acute aquatic toxicity to Pimephales Promelas (Fathead Minnow)_K Dagginus\qmrf143_TrainingPoolTotal_476.sdf
qmrf143_Test_101.sdf	http://qsardb.jrc.ec.europa.eu/qmrffile:///C:\Documents and Settings\casadja\My Documents\a_QSAR\a_QMRFdb\model submissions\7 review completed\Q10-23-19-143_2009-11-13_QSAR for acute aquatic toxicity to Pimephales Promelas (Fathead Minnow)_K Dagginus\qmrf143_Test_101.sdf
qmrf143_Verification_173.sdf	http://qsardb.jrc.ec.europa.eu/qmrffile:///C:\Documents and Settings\casadja\My Documents\a_QSAR\a_QMRFdb\model submissions\7 review completed\Q10-23-19-143_2009-11-13_QSAR for acute aquatic toxicity to Pimephales Promelas (Fathead Minnow)_K Dagginus\qmrf143_Verification_173.sdf

Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q13-33-0041

10.2.Publication date:

2013-06-27

10.3.Keywords:

neural network;acute fish toxicity;Pimephales Promelas;fathead minnow;

10.4.Comments:

former Q10-23-19-143