

	<b>QMRF identifier (JRC Inventory): Q13-42-0055</b>
	<b>QMRF Title: Nonlinear QSAR: artificial neural network for acute oral toxicity (rat cell line)</b>
	<b>Printing Date: Dec 11, 2019</b>

## 1. QSAR identifier

### 1.1. QSAR identifier (title):

Nonlinear QSAR: artificial neural network for acute oral toxicity (rat cell line)

### 1.2. Other related models:

### 1.3. Software coding the model:

QSARModel 3.3.8

Turu 2, Tartu, 51014, Estonia

<http://www.molcode.com>

Statistica 7

StatSoft Ltd.

<http://www.statsoft.com>

## 2. General information

### 2.1. Date of QMRF:

21.04.2010

### 2.2. QMRF author(s) and contact details:

Molcode model development team Molcode Ltd Turu 2, Tartu, 51014, Estonia  
models@molcode.com [www.molcode.com](http://www.molcode.com)

### 2.3. Date of QMRF update(s):

### 2.4. QMRF update(s):

### 2.5. Model developer(s) and contact details:

Molcode model development team Molcode Ltd Turu 2, Tartu, 51014, Estonia  
models@molcode.com [www.molcode.com](http://www.molcode.com)

### 2.6. Date of model development and/or publication:

12.04.2010

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

### 2.8. Availability of information about the model:

Training, selection and test sets are available. Model algorithm is available (snn file)

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

None to date.

## 3. Defining the endpoint - OECD Principle 1

### 3.1. Species:

Rat, cell line

### 3.2. Endpoint:

4. Human Health Effects 4.2. Acute Oral toxicity

### **3.3.Comment on endpoint:**

EU B.1 bis and B.1; OECD TG 420 and TG 423

### **3.4.Endpoint units:**

IC50 [mg/l]

### **3.5.Dependent variable:**

Log (IC50)

### **3.6.Experimental protocol:**

The acute oral toxicity is determined using the OECD 420 and OECD 423 (EU B.1.bis. and 1.B.tris.) test guidelines. Acute oral toxicity testing allows to obtain the information on the biologic/toxic activity of a chemical. Currently, the basis for toxicologic classification of chemicals is the median lethal dose (LD50, mg/kg b.w.), which is defined as the statistically derived dose required to kill half the members of a tested population. Animals are observed individually after dosing at least once during the first 30 minutes, periodically during the first 24 hours, with special attention given during the first 4 hours, and daily thereafter, for a total of 14 days.

Developed in vitro systems (biokinetic, cellular and molecular elements) can significantly reduce the number of animals needed for the toxicity testing of a broad range of compounds. There are 50 compounds in the in vitro cytotoxicity dataset (median of several IC50 values of different experiments) consisting of drugs, agrochemicals and industrial chemicals. The IC50 values of tested substances were translated to logarithmic scale (logIC50) to reduce the range of the data.

### **3.7.Endpoint data quality and variability:**

The experimental data, taken from different sources, is considered reliable:

Freidig et al (2007). <http://dx.doi.org/10.1016/j.toxlet.2007.03.008>;

NIEHS (2001).

[http://iccvam.niehs.nih.gov/docs/acutetox\\_docs/guidance0801/iv\\_guide.pdf](http://iccvam.niehs.nih.gov/docs/acutetox_docs/guidance0801/iv_guide.pdf)

## **4.Defining the algorithm - OECD Principle 2**

### **4.1.Type of model:**

Neural network

### **4.2.Explicit algorithm:**

Neural network

Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression

### **4.3.Descriptors in the model:**

[1]Schultz principal eigenvalue of D matrix Descriptors 1 and 2 are based on the graph theory where a compound is represented as connected graphs with vertex and edges. See Schultz et al (1993)

[2]Schultz principal eigenvalue of A+D matrix Descriptors 1 and 2 are based on the graph theory where a compound is represented as connected graphs with vertex and edges. See Schultz et al (1993)

[3]Number of rings

#### **4.4.Descriptor selection:**

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on a set of statistical selection rules as F statistic and p. The first highest F (low p) descriptors (3) were selected from the whole (~1000) descriptors. These 3 descriptors were used as inputs to the network. 18 networks with different structures were tested in order to find the best ANN with lowest RMS(root-mean-squared error). Then 197 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with Levenberg-Marquardt algorithm using logistic activation function.

#### **4.5.Algorithm and descriptor generation:**

All descriptors were generated using QSARModel on structures optimized by the AM1 semiempirical quantum mechanical model.

#### **4.6.Software name and version for descriptor generation:**

QSARModel 3.3.8  
<http://www.molcode.com>

#### **4.7.Chemicals/Descriptors ratio:**

8.75 (35 chemicals / 4 descriptors)

### **5.Defining the applicability domain - OECD Principle 3**

#### **5.1.Description of the applicability domain of the model:**

Applicability domain based on training set:

By descriptor value range (between min and max values): The model is suitable for compounds that have the descriptors in the following range:

Desc ID

See 4.3 1 2 3

Min 2.0000 3.0000 0.000000

Max 488.6895 490.6001 5.000000

#### **5.2.Method used to assess the applicability domain:**

Presence of functional groups in structures (as methyl, esters, nitro, amine, benzene)

Range of descriptor values in training set with  $\pm 30\%$  confidence

Descriptor values must fall between maximal and minimal descriptor values (see 5.1) of training set  $\pm 30\%$

#### **5.3.Software name and version for applicability domain assessment:**

QSARModel 3.3.8  
<http://www.molcode.com>

#### **5.4.Limits of applicability:**

See 5.2

### **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: No  
Formula: No  
INChI: No  
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:**

data points: 35

**6.6.Pre-processing of data before modelling:**

Standardization and normalization by taking into account the mean and standard deviation

**6.7.Statistics for goodness-of-fit:**

Training log (IC50); Selection log (IC50); Test log (IC50)

Data Mean 0.495; 0.790; 0.573

Data S.D. 1.870; 1.721; 1.851

Error Mean 0.001; -0.168; -0.198

Error S.D. 0.848; 0.571; 0.472

Abs E. Mean 0.619; 0.537; 0.398

S.D. Ratio 0.454; 0.332; 0.255

Correlation 0.891; 0.978; 0.970

**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:**

RMS (Training) = 0.120004, RMS (Selection) = 0.084220, RMS (Test) = 0.072431, See 6.7

**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: No

Formula: No

INChI: No

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:**

The method used two validation sets – selection (5) and test(10)

#### **7.6.Experimental design of test set:**

Randomly selected 10 and 5 data points

#### **7.7.Predictivity - Statistics obtained by external validation:**

See 6.7 and 6.12

#### **7.8.Predictivity - Assessment of the external validation set:**

The descriptors for the test set are in the limit of applicability, see 6.7 and 6.12

#### **7.9.Comments on the external validation of the model:**

Overall predictions for the selection set (used to stop the ANN training and not to overfit it) and the test set (used to test the external prediction of the net after training) are quite high according to the Pearson correlation coefficient and the standard deviation ratio (S.D.Ratio), see 6.7 and 6.12

### **8.Providing a mechanistic interpretation - OECD Principle 5**

#### **8.1.Mechanistic basis of the model:**

The most statistically significant descriptors (among ~1000) are those 3 selected for inputs of the ANN model. The descriptors Schultz principal eigenvalue of D and D+A are unique for a given compound - they can be used to characterize the compounds structural complexation related to the IC50. These two descriptors are unique for each compound in the set thus they can be used to characterize the compounds IC50. However, deep mechanistic interpretations for ANN models is difficult.

#### **8.2.A priori or a posteriori mechanistic interpretation:**

A posteriori mechanistic interpretation, consistent with published scientific interpretations of experiments.

#### **8.3.Other information about the mechanistic interpretation:**

### **9.Miscellaneous information**

#### **9.1.Comments:**

Supporting information for: training set; selection set; test set

ANN.snn file -includes the ANN model, in order to be used the user must have Statistica 7 or higher with ANN modules to make predictions.

The methodology and software (QSARModel) used to create the present model were applied also to obtain the results published in these papers:

Katritzky et al (2006), Karelson et al (2006) (ref 1/2, sect 9.2)

#### **9.2.Bibliography:**

- [1]Katritzky AR, Dobchev DA, Fara DC, Hur E, Tamm K, Kurunczi L, Karelson M, Varnek A & Solov'ev VP (2006). Skin Permeation Rate as a Function of Chemical Structure. Journal of Medicinal Chemistry 49, 3305-3314.
- [2]Karelson M, Dobchev DA, Kulshyn OV & Katritzky A (2008). Neural Networks Convergence Using Physicochemical Data. Journal of Chemical Information and Modeling 46, 1891-1897.
- [3]Freidig AP, Dekkers S, Verwei M, Zvinavashe E, Bessems JGM & van de Sandt JJM (2007). Development of a QSAR for worst case estimates of acute toxicity of chemically reactive compounds. Toxicology Letters 170, 214–222. <http://dx.doi.org/10.1016/j.toxlet.2007.03.008>

