QMRF	*QMRF identifier (JRC Inventory):To be entered by JRC* *QMRF Title: Nonlinear ANN QSAR Model for Acute toxicity (Inhalation - Rat (male; female))* *Printing Date:6.12.2011*	QMRF

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear ANN QSAR Model for Acute toxicity (Inhalation - Rat (male; female))

1.2.Other related models:

1.3.Software coding the model:

QSARModel 3.3.8; Statistica 7, StatSoft Ltd. Turu 2, Tartu, 51014, Estonia, http://www.molcode.com

2.General information

2.1.Date of QMRF:

10.10.2010

2.2.QMRF author(s) and contact details:

Dimitar Dobchev, Tarmo Tamm, Gunnar Karelson, Indrek Tulp, Dana Martin, Kaido Tämm, Deniss Savchenko, Jaak Jänes, Eneli Härk, Andres Kreegipuu, Mati Karelson, Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@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 Molcode Ltd Turu 2, Tartu, 51014, Estonia models@molcode.com www.molcode.com

2.6.Date of model development and/or publication:

12.04.2010The methodology and software (QSARModel) used to create the present model were applied also to obtain the results published in these papers.

1)Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hur, E.; Tämm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. (2006). Skin Permeation Rate as a Function of Chemical Structure . Journal of Medicinal Chemistry, 49(11), 3305 - 3314.

2)Karelson, M.; Dobchev, D. A.; Kulshyn, O. V.; Katritzky, A. (2006). Neural Networks Convergence Using Physicochemical Data. Journal of Chemical Information and Modeling, 46, 1891 - 1897.

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

[1]1)Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hur, E.; Tämm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. (2006). Skin

Permeation Rate as a Function of Chemical Structure . Journal of Medicinal Chemistry, 49(11), 3305 - 3314.
[2]2)Karelson, M.; Dobchev, D. A.; Kulshyn, O. V.; Katritzky, A. (2006). Neural Networks Convergence Using Physicochemical Data. Journal of Chemical Information and Modeling, 46, 1891 - 1897.
[3]Statistica 7 www.statsoft.com
[4]

2.8.Availability of information about the model:

All information in full detail is available

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

No other QMRF available for the same model

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rat (male; female)

3.2.Endpoint:

Inhalation toxicity LC50

3.3.Comment on endpoint:

Acute inhalation toxicity is the adverse effect caused by a substance following a single uninterrupted exposure by inhalation over a short period of time (4 h). LC50 (median lethal concentration) is a statistically derived estimate of a concentration of a substance that can be expected to cause death during exposure in 50 percent of animals. The LC50 value is expressed as amount of the test substance per unit volume of air, e.g. millimols per m3.

3.4.Endpoint units:

[mmol/m3]

3.5.Dependent variable:

LogLC50

3.6.Experimental protocol:

Male rats of the Sprague-Dawley strain aged 5 weeks were housed in polyethylene or alminium cages maintained at 23 ± 20 C and 50 ± 10 % relative humidity and wood flake bedding, which was changed twice a week. Except for the exposure period, the animals had free access to feed and filter-purified tap water. They were housed for one week and submitted to the experiment at the age of 6 weeks. The inhalation apparatus was used with test substance in a 500 g can as a generator. To check the concentration, the standard gas mixtures were poured into a gas chromatograph to draw standard calibration curves of concentration vs GC peak area. For monitoring the exposure condition, the samples were injected into the gas chromatograph, and concentrations of test substance were determined from the standard curves mentioned above. Gas was distributed evenly throughout the chamber. During exposure (4 hours), concentration of test substance inside the gas chamber was analyzed every 10 minutes and the flow rate was adjusted from the mixing chamber by observing a flow meter so that an appropriate concentration

could be maintained at all times [1 - 4].

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression

4.2.Explicit algorithm:

The algorithm is based on regression neural network predictor with structure 4-4-3-1

This is standard back-forward artificial neural network model with connected architecture 4 neurons in the first layer (descriptors), 4 neurons in the first hidden layer, 3 neurons in the second hidden layer and one neuron in the last layer responsible for the property.

4.3.Descriptors in the model:

[1]Highest n-n repulsion (AM1)
[2]Highest e-e repulsion (AM1)
[3]Max electrophilic reactivity index (AM1)
[4]Relative number of O atoms
[5]

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 (4) were selected from the whole (997) descriptors. These 4 descriptors were used as inputs to the network. 12 networks with different structures were tested in order to find the best ANN with lowest RMS (root-mean-squared error) and highest correct predictions (for training, selection and test sets). Then 555 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with Levenberg-Marquardt algorithm encoded in the backpropagation scheme using linear and hyperbolic activation functions.

4.5.Algorithm and descriptor generation:

All descriptors were generated using QSARModel on structure optimized by AM1 semiempirical quantum mechanical model. All structures were generated as random conformers and optimizaed first with molecular mechanic (MM, stopping criterion grad=0.05) followed with semiempirical quantum mechanical optimizations AM1 method (stopping criterion grad 0.05).

4.6.Software name and version for descriptor generation:

QSARModel
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

The ratio = 13.75. This is a significant ratio since roughly there are 14 data points per descriptor in the training set. However, this ratio

is not very applicable for the case of ANN model but rather for multilinear models, because in this case the ANN weights could also be considered as parameters.

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain indicates in what limits according to the chemicals structures and the numerical descriptors the model can be applied. The two types of limits are defined bellow:

a)The model is suitable for diverse compounds especially for small organic compounds with the following functional groups: functional grups as phenols, aldehydes, nitro, amino, alcohols,halides, aromatic, aliphatic functional groups and other

b)The model is suitable for compounds that have descriptors values in the followin range;

Desc 1 2 3 4
min 125.091 126.2694 0.005661 0
max 333.0623 330.2876 0.056181 0.615385

5.2.Method used to assess the applicability domain:

Presence of functional groups in structures.

Range of descriptor values in training set with ±30% confidence.

Descriptor values must fall between maximal and minimal descriptor values (see5.1) of training set ±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 5.1, 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: 55

6.6.Pre-processing of data before modelling:

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

6.7.Statistics for goodness-of-fit:

	Training Log(LC50)	Selection Log(LC50)	Test Log(LC50)
Data Mean	2.056	2.542	2.538
Data S.D.	1.028	0.625	1.039
Error Mean	0.000	-0.155	-0.738
Error S.D.	0.349	0.928	1.256
Abs E. Mean	0.258	0.733	0.984
S.D. Ratio	0.339	1.485	1.209
Correlation	0.941	0.459	0.731

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

See 6.7

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)= 0.076, RMS(Selection)= 0.204, RMS(Test) = 0.315,

In this ANN were used 2 sets randomly chosen (10) to test the network – selection set and test set, see also 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 (10) and test (10)

7.6.Experimental design of test set:

Randomly selected 10 and 10 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 significant according to the RMS error and the standard deviation ratio (S.D.Ration), see 6.7 and 6.12

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Since the ANN is a more complex predictor than a linear model, it is difficult to analyze the relation between the property and the descriptors. Most of the descriptors are related to the reactivity of the compounds. However, one could make rough estimation based on their values. Regarding the first three descriptors, it can be noted that they have slight negative correlation with the property. It might suggest that with the increase of these descriptors, the property would decrease. It seems that the oxygen atoms in the molecules contribute to lower Log LC50 values according to the 4th descriptor.

8.2.A priori or a posteriori mechanistic interpretation:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

Supporting information for :Training set(s)
Selection set(s)
Test set(s)
4-4-3-1.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.

9.2.Bibliography:

[1]ECOTOX database http://cfpub.epa.gov/ecotox/index.html
[2]Veith, G. D. , Petkova, E. P. and Wallace, K. B. A baseline inhalation toxicity model for narcosis in mammals, SAR and QSAR in Environmental Research, 20 (5), 2009, 567 – 578.
[3]1. Rusch G. M. , Bast C. B., Cavender F. L. Establishing a point of departure for risk assessment using acute inhalation toxicology data, Regulatory Toxicology and Pharmacology 54, 2009, 247–255.
[4]1. Wolf W. de , Lieder P. H. and D. Walker J. Application of QSARs: Correlation of Acute Toxicity in the Rat Following Oral or Inhalation

Exposure, QSAR Comb. Sci. 23, 2004, 521-525.
[5]

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC Inventory)

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