
	QMRF identifier (ECB Inventory): Q52-55-56-520	
	QMRF Title: Artificial Intelligence Expert Predictive System (AIEPS) model for acute toxicity to <i>Daphnia magna</i>	
	Printing Date: Aug 17, 2017	

1. QSAR identifier

1.1. QSAR identifier (title):

Artificial Intelligence Expert Predictive System (AIEPS) model for acute toxicity to *Daphnia magna*

1.2. Other related models:

1.3. Software coding the model:

[1] Accelrys Accord Chemistry SDK v 6.1 Accord Software Development Kit BIOVIA 5005 Wateridge Vista Drive, San Diego, CA 92121 USA Tel: +1 858 799 5000 <http://accelrys.com/>; <http://accelrys.com/products/datasheets/accord-chemistry-cartridge.pdf>

[2] Accelrys Accord Chemistry Control 6 Runtime Active X Chemistry control - database files used by windows installer BIOVIA 5005 Wateridge Vista Drive, San Diego, CA 92121 USA Tel: +1 858 799 5000 <http://accelrys.com/>; <http://accelrys.com/products/datasheets/accord-chemistry-control.pdf>

2. General information

2.1. Date of QMRF:

22 December, 2015

2.2. QMRF author(s) and contact details:

Mark Lewis Health Canada 99 Metcalfe St., Ottawa, Ontario, Canada, K1A 0K9
mark.lewis@canada.ca <http://www.hc-sc.gc.ca/ewh-semt/index-eng.php>

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

Stefan P. Niculescu Scientific Consultant spniculescu@gmail.com

2.6. Date of model development and/or publication:

9 November 2012

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

[1] Kaiser KLE and Niculescu SP (2001). Modeling acute toxicity of chemicals to *Daphnia magna*: A probabilistic neural network approach. *Environmental toxicology and chemistry* 20 (2) 420-431. <http://onlinelibrary.wiley.com/doi/10.1002/etc.5620200225/full>

[2] Niculescu SP, Kaiser KLE and Schultz TW (2000). Modeling the toxicity of chemicals to *Tetrahymena pyriformis* using molecular fragment descriptors and probabilistic neural networks. *Archives of environmental contamination and toxicology* 39 (3) 289-329 <http://link.springer.com/article/10.1007/s002440010107>

[3] Niculescu SP, Atkinson A, Hammond G & Lewis M (2004). Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow. *SAR and QSAR in Environmental Research* 15 (4) 293-309. <http://www.tandfonline.com/doi/abs/10.1080/10629360410001724941>

[4] Niculescu SP, Lewis MA and Tigner J (2008). Probabilistic neural networks modeling of the 48-h LC50 acute toxicity endpoint to *Daphnia magna*. *SAR and QSAR in Environmental Research* 19 (7-8) 735-750. <http://www.tandfonline.com/doi/abs/10.1080/10629360802550556>

[5]Masters T (1993) Practical Neural Network Recipes in C++. Academic Press, San Diego”
[https://books.google.ca/books?id=7Ez_Pq0sp2EC&lpg=PR17&ots=e05FixTiqW&dq=Masters%20T%20\(1993\)%20Practical%20Neural%20Network%20Recipes%20in%20C%2B%2B.%20Academic%20Press%2C%20San%20Diego%E2%80%9D&lr&pg=PR17#v=onepage&q&f=false](https://books.google.ca/books?id=7Ez_Pq0sp2EC&lpg=PR17&ots=e05FixTiqW&dq=Masters%20T%20(1993)%20Practical%20Neural%20Network%20Recipes%20in%20C%2B%2B.%20Academic%20Press%2C%20San%20Diego%E2%80%9D&lr&pg=PR17#v=onepage&q&f=false)

2.8.Availability of information about the model:

The model and training set are not proprietary.

The setup involves installation of Accelrys Chemistry Control 6.0.1 Runtime and Accord SDK 6.1 Runtime. Consult with Accelrys/Biovia on any legal obligations or limitations.

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnia magna

3.2.Endpoint:

QMRF 3. 1. Short-term toxicity to Daphnia (immobilisation) OECD 202 Daphnia sp Acute Immobilisation Test

3.3.Comment on endpoint:

Daphnia magna 48h LC50 - concentration of test chemical that kills 50% of the test subjects in a 48-h exposure test.

3.4.Endpoint units:

mmol/L or mg/L

3.5.Dependent variable:

The relationship between Daphnia magna 48h LC50 and selected molecular fragment descriptors is implemented through a basic Probabilistic Neural Network (PNN) with Gaussian kernel (statistical corrections included). Atoms and fragment information is generated directly from molecular structure using fragment chemistry data mining. The model may handle both inorganic and organic compounds. All data modeling is performed at the level of Log (mmol/L) units.

3.6.Experimental protocol:

Not specified

3.7.Endpoint data quality and variability:

The toxicity information stored in this database is the result of a critical evaluation of the data in the US Environmental Protection Agency ECOTOX database, scientific publications and protected data sources. Wherever applicable, the compounds are identified using Chemical Abstracts Service Registry Numbers (CAS RN). Retrieval and validation of CAS RN and molecular structure data has been performed using the PubChem on-line search engine. The database contains measured D. magna LC50 information for 1052 chemical structures (both organics and inorganics).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Probabilistic Neural Network with Gaussian kernel (statistical corrections) included

4.2.Explicit algorithm:

PNN Algorithm

Probabilistic Neural Network with Gaussian kernel (statistical corrections) included

see Attachment

Details on PNN methodology may be found here:

Masters T (1993) Practical Neural Network Recipes in C++. Academic Press, San Diego

4.3.Descriptors in the model:

- [1]number of silver atoms count number of silver atoms
- [2]number of arsenic atoms count number of arsenic atoms
- [3]number of boron atoms count number of boron atoms
- [4]number of barium atoms count number of barium atoms
- [5]number of berilium atoms count number of berilium atoms
- [6]number of bromine atoms count number of bromine atoms
- [7]number of carbon atoms count number of carbon atoms
- [8]number of calcium atoms count number of calcium atoms
- [9]number of cadmium atoms count number of cadmium atoms
- [10]number of chlorine atoms count number of chlorine atoms
- [11]number of cobalt atoms count number of cobalt atoms
- [12]number of chromium atoms count number of chromium atoms
- [13]number of copper atoms count number of copper atoms
- [14]number of fluorine atoms count number of fluorine atoms
- [15]number of iron atoms count number of iron atoms
- [16]number of hydrogen atoms count number of hydrogen atoms
- [17]number of mercury atoms count number of mercury atoms
- [18]number of iodine atoms count number of iodine atoms
- [19]number of magnesium atoms count number of magnesium atoms
- [20]number of manganese atoms count number of manganese atoms
- [21]number of nitrogen atoms count number of nitrogen atoms
- [22]cumulative number of sodium, potassium and lithium atoms count cumulative number of sodium, potassium and lithium atoms
- [23]number of oxygen atoms count number of oxygen atoms
- [24]number of phosphorus atoms count number of phosphorus atoms
- [25]number of lead atoms count number of lead atoms
- [26]number of sulfur atoms count number of sulfur atoms
- [27]number of selenium atoms count number of selenium atoms
- [28]number of silicon atoms count number of silicon atoms
- [29]number of tin atoms count number of tin atoms
- [30]number of uranium atoms count number of uranium atoms
- [31]number of vanadium atoms count number of vanadium atoms
- [32]number of zinc atoms number of zinc atoms
- [33]ratio between the cumulative number of nitrogen and oxygen atoms in the molecule over the cumulative number of nitrogen, oxygen and carbon atoms (1 for inorganics) ratio between the cumulative number of nitrogen and oxygen atoms in the molecule over the cumulative number of nitrogen, oxygen and carbon atoms (1 for inorganics)
- [34]number of methyl groups number of methyl groups
- [35]number of triple bonds between carbon atoms number of triple bonds between carbon atoms

[36]number of nitrile groups, carbonitrile excluded number of nitrile groups, carbonitrile excluded

[37]number of C-C#N groups number of C-C#N groups

[38]number of N=C=S groups number of N=C=S groups

[39]number of S=C groups, isothiocyanat excluded number of S=C groups, isothiocyanat excluded

[40]number of S-C#N groups number of S-C#N groups

[41]number of S-C groups, thiocyanat excluded number of S-C groups, thiocyanat excluded

[42]number of N-N, N=N, and N#N groups number of N-N, N=N, and N#N groups

[43]number of amide groups attached to carbons from rings number of amide groups attached to carbons from rings

[44]number of amide groups attached to carbons not part of rings number of amide groups attached to carbons not part of rings

[45]number of amide groups not connected to carbons number of amide groups not connected to carbons

[46]number of amine groups attached to carbons from rings number of amine groups attached to carbons from rings

[47]number of amine groups attached to carbons not part of rings, amides excluded number of amine groups attached to carbons not part of rings, amides excluded

[48]number of amine groups not attached to carbons number of amine groups not attached to carbons

[49]number of carbon-halogen bonds where the carbons are in rings number of carbon-halogen bonds where the carbons are in rings

[50]number of CF₃ groups number of CF₃ groups

[51]number of CCl₃ groups number of CCl₃ groups

[52]number of carbon-halogen bonds where the carbons are not part of rings number of carbon-halogen bonds where the carbons are not part of rings

[53]number of OH groups attached to carbons from rings number of OH groups attached to carbons from rings

[54]number of C-O groups where C is part of a ring, RingC-OH excluded number of C-O groups where C is part of a ring, RingC-OH excluded

[55]number of ester bridges number of ester bridges

[56]number of ether bridges, ester bridges excluded number of ether bridges, ester bridges excluded

[57]number of carboxyl groups attached to carbons from rings number of carboxyl groups attached to carbons from rings

[58]number of carboxyl groups, RingC-carboxyl excluded number of carboxyl groups, RingC-carboxyl excluded

[59]number of C-OH groups where C is not is ring, carboxyls excluded g/mole number of C-OH groups where C is not is ring, carboxyls excluded

[60]number of O-C(=O)([]) bridges, carboxyls and esters excluded number of O-C(=O)([]) bridges, carboxyls and esters excluded

[61]number of C=O groups where the carbon is not part of a ring, and excluding those included in amides, carboxyls, ester bridges, isocyanat and aldehydes, but including those part of OC(=O)O groups number of C=O groups where the carbon is not part of a ring, and excluding

those included in amides, carboxyls, ester bridges, isocyanat and aldehydes, but including those part of OC(=O)O groups

[62]number of OH groups attached to nitrogen number of OH groups attached to nitrogen

[63]number of nitrogen-halogens bonds number of nitrogen-halogens bonds

[64]number of NO₂ groups attached to carbons from aromatic rings number of NO₂ groups attached to carbons from aromatic rings

[65]number of nitrate groups number of nitrate groups

[66]number of NO₂ groups not attached to carbons from rings, nitrate excluded number of NO₂ groups not attached to carbons from rings, nitrate excluded

[67]Number of N=O groups, NO₂ excluded Number of N=O groups, NO₂ excluded

[68]ratio between the cumulative number of nitrogen and oxygen atoms in the molecule which are not part of N(=O)=O groups over the number of carbons (0 for inorganics) ratio between the cumulative number of nitrogen and oxygen atoms in the molecule which are not part of N(=O)=O groups over the number of carbons (0 for inorganics)

[69]number of aldehyde groups number of aldehyde groups

[70]number of bridges consisting of a sulphur atom connected with only three oxygens and made of two S=O and one S-O subgroups number of bridges consisting of a sulphur atom connected with only three oxygens and made of two S=O and one S-O subgroups

[71]number of bridges consisting of a sulphur atom connected with four oxygens and made of two S=O and two S-O subgroups number of bridges consisting of a sulphur atom connected with four oxygens and made of two S=O and two S-O subgroups

[72]number of bridges consisting of a sulphur atom connected with two oxygens through double bonds, excluding sulfonic and sulfate bridges number of bridges consisting of a sulphur atom connected with two oxygens through double bonds, excluding sulfonic and sulfate bridges

[73]number of S=O groups not part of S(=O)=O bridges number of S=O groups not part of S(=O)=O bridges

[74]number of vinyl groups number of vinyl groups

[75]number of quinone groups number of quinone groups

[76]number of CC(=O)C groups, quinones excluded number of CC(=O)C groups, quinones excluded

[77]number of sulphur-hydrogen bonds number of sulphur-hydrogen bonds

[78]number of bridges consisting of a nitrogen atom connected through single bonds to four carbons number of bridges consisting of a nitrogen atom connected through single bonds to four carbons

[79]number of S=P(S)(O)O bridges number of S=P(S)(O)O bridges

[80]number of S=P(O)(O)O bridges number of S=P(O)(O)O bridges

[81]number of C1CC1 rings number of C1CC1 rings

[82]number of single phosphorus-nitrogen bonds number of single phosphorus-nitrogen bonds

[83]number of P-OH groups number of P-OH groups

[84]number of P-O- groups except P-OH number of P-O- groups except P-OH

[85]number of single carbon-metal bonds number of single carbon-metal bonds

[86]number of single oxygen-metal bonds number of single oxygen-metal bonds

[87]number of single sulphur-metal bonds number of single sulphur-metal bonds

[88]number of carbon atoms in rings number of carbon atoms in rings

[89]number of nitrogen atoms in rings number of nitrogen atoms in rings

[90]number of sulphur atoms in rings number of sulphur atoms in rings

[91]ratio of the number of atoms in aromatic rings over the total number of atoms in the molecule

[92]ratio of the number of atoms in non-aromatic rings over the total number of atoms in the molecule

[93]number of carbons in the longest carbon atoms chain whose bonds are not part of any ring and at least one extremity is not part of a ring

[94]number of bonds in non-isolated rings minus the corresponding number of atoms

[95]number of vinyl groups

[96]molecular weight

4.4.Descriptor selection:

96 descriptors were chosen in the final model. The descriptors used for Daphnia magna toxicity predictions were largely based the 78 descriptors used in the fathead minnow acute toxicity model. The descriptors poorly represented or absent in the structures of the 800 compounds in the fathead minnow training dataset were eliminated from the list. Examination of the chemical structures and partial modeling experiments were conducted to identify additional descriptors.

4.5.Algorithm and descriptor generation:

See attachment (AIEPS 3.0 - Daphnia magna 48hr LC50 PNN Model Validation Study.doc), section 4, for the discussion of the derivation and refinement of the PNN algorithm. As a starting point the multivariate Bayesian density estimator is used in combination with a mapping tool similar to the Maximum Likelihood Estimation method. The best probability density associated with the accumulative distribution of the cases in the training set is determined using Meisels' algorithm. Details can be found in Masters T (1993) Practical Neural Network Recipes in C++. Academic Press, San Diego.

4.6.Software name and version for descriptor generation:

Accelrys - Accord Chemistry Control 6.0.1 and Accord SDK 6.01

Runtime versions of these are included with the distributed program. The descriptors are automatically generated from the SMILES string during the data minning stage prior to prediction generation.

Accelrys.com

4.7.Chemicals/Descriptors ratio:

The number of chemicals in training set to descriptors ratio is $971/96 = 10.12$

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Based on the continuity of the mathematical functions involved in the model's computation algorithm, predictions are expected to be reliable when the values of the model input values are in the range between the minimum and maximum values of the corresponding descriptors encountered in the model's training data set, or outside close to them.

5.2.Method used to assess the applicability domain:

The substance of interest should have chemical descriptors which fall within the minimum or maximum values of those used in the training set. In addition, the model provides means to compare the substance of interest to those in the training set through Tanimoto indices. In other words, a prediction may be deemed acceptable when the Tanimoto maximum similarity indicator with the compounds in the model's training set is higher than a professionally determined value. For each prediction, the AIEPS provides the functionality of generating a similarity with the model's training dataset report, where the 10 most similar compounds are identified and the corresponding measured information reported in table format. Another table allows comparison between the values used as model input with the ranges of the corresponding training set descriptors. So, all necessary elements to judge the reliability of the predictions are made available to the user. Based on this information, is up to the user to decide if the predicted value is reliable or not.

5.3. Software name and version for applicability domain assessment:

5.4. Limits of applicability:

The model targets only small molecules consisting of less than 200 atoms. It is not recommended to use it for larger structures.

The model can handle both organics and inorganics.

With few exceptions the model cannot account for the differences between structural isomers. The exceptions occur when the combination of the model fragment descriptors is able to recognize them.

Predictions may not be accurate when the target structure involves active fragments not accounted for by the existing model descriptors.

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: No

INChI: No

MOL file: 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 training set consisting of 971 structures selected from the 1052 compound data set available. Data was mainly secured from the US EPA AQUIRE database (ECOTOX). The associated structures were validated mainly using PubChem structure search engine. Predictions were also based on empirical data for Daphnia magna tests of durations of 24, 96 and 504 hours which were extrapolated or interpolated to 48 hr LC50's based on generated formulae. The Shapiro-Wilk W Test Statistics were 0.9206, 0.9472, and 0.9440, respectively. Further details can be found in the attachment (AIEPS 3.0 - Daphnia magna 48hr LC50 PNN Model Validation

6.6.Pre-processing of data before modelling:

For those Daphnia magna LC50 test endpoints of 24, 96 or 504 hrs, the values (29.6%) were interpolated or extrapolated to 48hr LC50 values using the following fomulae:
DM48=1.03631*DM24-0.295 (n = 275) DM48=0.96486*DM96+0.36996 (n=25)
DM48=(2*DM24+DM96)/3 (n=4) DM48=0.80633*DM504+0.19261 (n=8) (See also Attached Model Validation study)

6.7.Statistics for goodness-of-fit:

Minimum Residuals -3.1300
Maximum Residuals 2.3455
Average Residuals -6.39E-09
Standard Deviation of Residuals 0.6579
Sum of Square Residuals 419.9016
Average Square Residuals 0.4324
Coefficient of Determination Between Measured and Predicted 0.8551
Coefficient of Correlation Between Measured and Predicted 0.9247
Training/Learning Set Size 971

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:

Yes

7.2.Available information for the external validation set:

CAS RN:Yes

Chemical Name:Yes

Smiles:Yes

Formula:No

INChI:No

MOL file: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:

81 substances, randomly generated through computer algorithm, were identified from the total of 1052 substances available. These were used as the external validation set.

For those Daphnia magna LC50 test endpoints of 24, 96 or 504 hrs, the values (29.6%) were interpolated or extrapolated to 48hr LC50 values using the following fomulae:
DM48=1.03631*DM24-0.295 (n = 275) DM48=0.96486*DM96+0.36996 (n=25)
DM48=(2*DM24+DM96)/3 (n=4) DM48=0.80633*DM504+0.19261 (n=8) (See also Attached Model Validation study)

7.6. Experimental design of test set:

Experimental data was randomly set aside before modeling

7.7. Predictivity - Statistics obtained by external validation:

Minimum Residuals -1.4084

Maximum Residuals 1.6745

Average Residuals 0.1498

Standard Deviation of Residuals 0.6754

Sum of Square Residuals 38.3092

Average Square Residuals 0.4730

Coefficient of Determination Between Measured and Predicted 0.7603

Coefficient of Correlation Between Measured and Predicted 0.8720

Shapiro-Wilk W Test Statistic for Residuals 0.9769

Prob<W 0.4465

External Test Set Size 81

7.8. Predictivity - Assessment of the external validation set:

The Shapiro-Wilk W Test accepts the null hypothesis that the distribution of the residuals on the external test set of 81 compounds is normal at $\alpha=0.05$ significance level.

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:

The mechanistic approach of the present model is supported by the use of molecular weight, and presence of specific atoms, bonds, and molecular fragments and the training of the neural network has assigned weights to each of these descriptors to correct for their influence on the 48hr LC50 to *Daphnia magna*.

8.2. A priori or a posteriori mechanistic interpretation:

The mechanistic interpretation was determined a posteriori by interpreting and modifying the final set of descriptors which contributed to the best fit.

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] Kaiser KLE and Niculescu SP (2001). Modeling acute toxicity of chemicals to *Daphnia magna*: A probabilistic neural network approach. *Environmental toxicology and chemistry* 20 (2) 420-431. <http://onlinelibrary.wiley.com/doi/10.1002/etc.5620200225/full>

[2] Niculescu SP, Kaiser KLE and Schultz TW (2000). Modeling the toxicity of chemicals to *Tetrahymena pyriformis* using molecular fragment descriptors and probabilistic neural networks. *Archives of environmental contamination and toxicology* 39 (3) 289-329 <http://link.springer.com/article/10.1007/s002440010107>

[3] Niculescu SP, Atkinson A, Hammond G & Lewis M (2004). Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow. *SAR and QSAR in Environmental Research* 15 (4) 293-309. <http://www.tandfonline.com/doi/abs/10.1080/10629360410001724941>

[4] Niculescu SP, Lewis MA and Tigner J (2008). Probabilistic neural networks modeling of the

48-h LC50 acute toxicity endpoint to Daphnia magna. SAR and QSAR in Environmental Research 19 (7-8) 735-750. <http://www.tandfonline.com/doi/abs/10.1080/10629360802550556>
[5]Masters T (1993) Practical Neural Network Recipes in C++. Academic Press, San Diego”
[https://books.google.ca/books?id=7Ez_Pq0sp2EC&lpg=PR17&ots=e05FixTiqW&dq=Masters%20T%20\(1993\)%20Practical%20Neural%20Network%20Recipes%20in%20C%2B%2B.%20Academic%20Press%2C%20San%20Diego%E2%80%9D&lr&pg=PR17#v=onepage&q&f=false](https://books.google.ca/books?id=7Ez_Pq0sp2EC&lpg=PR17&ots=e05FixTiqW&dq=Masters%20T%20(1993)%20Practical%20Neural%20Network%20Recipes%20in%20C%2B%2B.%20Academic%20Press%2C%20San%20Diego%E2%80%9D&lr&pg=PR17#v=onepage&q&f=false)

9.3.Supporting information:

Training set(s)

AIEPs 3.0 - Daphnia Training Set_971	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf520_AIEPs 3.0 - Daphnia Training Set_971.sdf
--------------------------------------	---

Test set(s)

AIEPS 3.0 -Daphnia Validation_82	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf520_AIEPS 3.0 - Daphnia Validation_82.sdf
----------------------------------	---

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q52-55-56-520

10.2.Publication date:

2016/11/11

10.3.Keywords:

Artificial Intelligence Expert Predictive System, AIEPS, daphnia magna, acute toxicity

10.4.Comments: