	QMRF identifier (JRC Inventory): **Q17-201-0035**
	QMRF Title: **Artificial Intelligence Expert Predictive System (AIEPS) model for algal (Pseudokirchneriella subcapitata) toxicity**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Artificial Intelligence Expert Predictive System (AIEPS) model for algal (Pseudokirchneriella subcapitata) toxicity

1.2.Other related models:

1.3.Software coding the model:

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

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:

23 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

2.8.Availability of information about the model:

The model is 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:

Green algae - Pseudokirchneriella subcapitata

3.2.Endpoint:

3.2.Short-term toxicity to algae (inhibition of the exponential growth rate) 201Alga Growth Inhibition Test

3.3.Comment on endpoint:

This model, implemented in AIEPS v 3.0, addresses the computation of the 72-hr EC50 for the Green Algae (Pseudokirchneriella subcapitata) for organic chemicals based on basic Probabilistic Neural Network (PNN) with Gaussian kernel (statistical corrections included) corrections

3.4.Endpoint units:

mmol/L or mg/L

3.5.Dependent variable:

The relationship between Pseudokirchneriella subcapitata 72h EC50 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.

3.6.Experimental protocol:

OECD TG 201 (majority) and other protocols.

3.7.Endpoint data quality and variability:

Mainly, the data has been secured from the Japanese Database on Aquatic Toxicity, Ministry of the Environment, Japan (http://www.nite.go.jp/en/index.html).
Using the information in the AIEPS Algae Species LC50, EC50 (24-hr and Greater) knowledge database a 574 organic compounds dataset has been assembled for modeling purposes. Measured information for Pseudokirchneriella subcapitata 72-hr EC50 was available for 301 compounds. The information for the remaining 273 compounds was estimated using mathematical relationships between Pseudokirchneriella subcapitata 24,48,96 and 120-hr EC50, and Scenedesmus subspicatus 72 and 96-hr EC50

to the Pseudokirchneriella subcapitata 72-hr EC50 (for equations, see Attachment - AIEPS 3.0 - Pseudokirchneriella subcapitata 72hr EC50 PNN Model Validation Study.doc). Algae toxicity data was accepted if it was from P. subcapitata or S. subspicatus and involved a discrete chemical, and the measurement of a discrete endpoint (e.g. > EC50 values were not accepted), other specific criteria related to data quality was not applied.

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 arsenic atoms count number of arsenic atoms

[2]number of bromine atoms count number of bromine atoms

[3]number of carbon atoms count number of carbon atoms

[4]number of chlorine atoms count number of chlorine atoms

[5]number of fluorine atoms count number of fluorine atoms

[6]number of iron atoms count number of iron atoms

[7]number of hydrogen atoms count number of hydrogen atoms

[8]number of iodine atoms count number of iodine atoms

[9]number of manganese atoms count number of manganese atoms

[10]number of nitrogen atoms count number of nitrogen atoms

[11]cumulative number of sodium, potassium and lithium atoms count cumulative number of sodium, potassium and lithium atoms

[12]number of oxygen atoms count number of oxygen atoms

[13]number of phosphorus atoms count number of phosphorus atoms

[14]number of sulphur atoms count number of sulphur atoms

[15]number of tin atoms count number of tin atoms

[16]number of zinc atoms count number of zinc atoms

[17]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) count 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)

[18]number of methyl groups count number of methyl groups

[19]number of triple bonds between carbon atoms count number of triple bonds between carbon atoms

[20]number of nitrile groups, carbonitrile excluded count number of nitrile groups, carbonitrile

excluded
[21]number of C-C#N groups count number of C-C#N groups
[22]number of N=C=S groups count number of N=C=S groups
[23]number of S=C groups, isothiocyanat excluded count number of S=C groups, isothiocyanat excluded
[24]number of S-C#N groups count number of S-C#N groups
[25]number of S-C groups, thiocyanat excluded count number of S-C groups, thiocyanat excluded
[26]number of N-N, N=N, and N#N groups count number of N-N, N=N, and N#N groups
[27]number of amide groups count number of amide groups
[28]number of amine groups attached to carbons from rings count number of amine groups attached to carbons from rings
[29]number of amine groups attached to carbons not part of rings, amides excluded count number of amine groups attached to carbons not part of rings, amides excluded
[30]number of amine groups not attached to carbons count number of amine groups not attached to carbons
[31]number of carbon-halogen bonds where the carbons are in rings count number of carbon-halogen bonds where the carbons are in rings
[32]number of CF3 groups count number of CF3 groups
[33]number of CCl3 groups count number of CCl3 groups
[34]number of carbon-halogen bonds where the carbons are not part of rings (CF3 and CCl3 excluded) count number of carbon-halogen bonds where the carbons are not part of rings (CF3 and CCl3 excluded)
[35]number of OH groups attached to carbons from rings count number of OH groups attached to carbons from rings
[36]number of C-O groups where C is part of a ring, RingC-OH excluded count number of C-O groups where C is part of a ring, RingC-OH excluded
[37]number of ester bridges number of ester bridges
[38]number of ether bridges, ester bridges excluded number of ether bridges, ester bridges excluded

[39]number of carboxyl groups attached to carbons from rings number of carboxyl groups attached to carbons from rings
[40]number of carboxyl groups, RingC-carboxyl excluded number of carboxyl groups, RingC-carboxyl excluded
[41]number of C-OH groups where C is not is ring, carboxyls excluded number of C-OH groups where C is not is ring, carboxyls excluded
[42]number of O-C(=O)([]) bridges, carboxyls and esthers excluded number of O-C(=O)([]) bridges, carboxyls and esthers excluded
[43]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
[44]number of OH groups attached to nitrogen number of OH groups attached to nitrogen
[45]number of nitrogen-halogens bonds number of nitrogen-halogens bonds
[46]number of NO2 groups attached to carbons from aromatic rings number of NO2 groups attached to carbons from aromatic rings

[47]number of nitrate groups number of nitrate groups
[48]number of N=O groups, AroRingC-NO2 and nitrate excluded number of N=O groups, AroRingC-NO2 and nitrate excluded
[49]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)
[50]number of aldehyde groups number of aldehyde groups
[51]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
[52]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
[53]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
[54]number of S=O groups not part of S(=O)=O bridges number of S=O groups not part of S(=O)=O bridges
[55]number of CC(=O)C groups number of CC(=O)C groups
[56]number of C=O groups where C is in ring, CC(=O)C groups where all carbons are in ring excluded number of C=O groups where C is in ring, CC(=O)C groups where all carbons are in ring excluded
[57]number of sulphur-hydrogen bonds number of sulphur-hydrogen bonds
[58]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
[59]number of S=P(S)(O)O bridges number of S=P(S)(O)O bridges
[60]number of S=P(O)(O)O bridges number of S=P(O)(O)O bridges
[61]number of C1CC1 rings number of C1CC1 rings
[62]number of single phosphorus-nitrogen bonds number of single phosphorus-nitrogen bonds
[63]number of P-OH groups number of P-OH groups
[64]number of P-O- groups except P-OH number of P-O- groups except P-OH
[65]number of single carbon-metal bonds number of single carbon-metal bonds
[66]number of single oxygen-metal bonds number of single oxygen-metal bonds
[67]number of single sulphur-metal bonds. number of single sulphur-metal bonds.
[68]number of carbon atoms in rings number of carbon atoms in rings
[69]number of nitrogen atoms in rings number of nitrogen atoms in rings
[70]number of sulphur atoms in rings number of sulphur atoms in rings
[71]ratio of the number of atoms in aromatic rings over the total number of atoms in the molecule ratio of the number of atoms in aromatic rings over the total number of atoms in the molecule
[72]ratio of the number of atoms in non-aromatic rings over the total number of atoms in the molecule ratio of the number of atoms in non-aromatic rings over the total number of atoms in the molecule
[73]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 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

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

[75]number of vinyl groups number of vinyl groups

[76]molecular weight g/mol molecular weight

4.4.Descriptor selection:

76 descriptors were chosen in the final model.. A random initial list was generated. Partial modeling experiments were conducted to identify superfluous descriptors. The fragments whose presence had no impact on the resulting models behavior or resulted in a model with weaker overall generalization capability were removed, others were added.

4.5.Algorithm and descriptor generation:

See attachment (AIEPS 3.0 - AIEPS 3.0 - Pseudokirchneriella subcapitata 72hr EC50 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:

Accelerys - 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 528/76 = 6.95

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 models training set is higher than a professionally determined value. For each prediction, the AIEPS provides the functionality of generating a similarity with the

models 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 is limited only to organics.
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:

301 data were available for 72 hr EC50 values for Pseudokirchneriella subcapitata following OECD 201, the remaining 273 were extrapolated from other measured endpoints e.g. Pseudokirchneriella subcapitata 24,48,96 and 120-hr EC50, and Scenedesmus subspicatus 72 and 96-hr EC50.

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

Minimum Residuals -3.23862
Maximum Residuals 2.0869
Average Residuals 4.94E-08
Standard Deviation of Residuals 0.6838
Sum of Square Residuals 246.4234
Average Square Residuals 0.4667
Coefficient of Determination Between Measured and Predicted Values 0.7367

Coefficient of Correlation Between Measured and Predicted Values 0.8583

Training/Learning Set Size 528

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:

Randomly selected external test set (not used in training) of 46 compounds:

Minimum Residuals -1.8946

Maximum Residuals 1.5065

Average Residuals 0.0346

Standard Deviation of Residuals 0.7940

Sum of Square Residuals 28.4256

Average Square Residuals 0.6179

Coefficient of Determination Between Measured and Predicted Values 0.6053

Coefficient of Correlation Between Measured and Predicted Values 0.7780

Shapiro-Wilk W Test Statistic for Residuals 0.9763

Prob<W 0.6215

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:

46 substances were used as the validation set for the acute algae toxicity PNN model.

These were selected randomly from the whole dataset of 574 substances.

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 -1.8946

Maximum Residuals 1.5065

Average Residuals 0.0346

Standard Deviation of Residuals 0.7940

Sum of Square Residuals 28.4256

Average Square Residuals 0.6179

Coefficient of Determination Between Measured and Predicted Values 0.6053

Coefficient of Correlation Between Measured and Predicted Values 0.7780

Shapiro-Wilk W Test Statistic for Residuals 0.9763

Prob<W 0.6215

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 46 compounds is normal at ?=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 the specific atoms, bonds, and molecular fragments involved in the model descriptors.

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

9.3.Supporting information:

AIEPS 3.0 - Algae Training Set_582	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf517_AIEPS 3.0 - Algae Training Set_582.sdf

AIEPS 3.0 -Algae Validation_47	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf517_AIEPS 3.0 -Algae Validation_47.sdf

Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-201-0035

10.2.Publication date:

2017-09-27

10.3.Keywords:

Artificial Intelligence Expert Predictive System;AIEPS;Algal Growth Inhibition Test;Pseudokirchneriella subcapitata;

10.4.Comments:

old# Q52-55-56-517