	QMRF identifier (JRC Inventory): **Q17-C1-0033**
	QMRF Title: **Artificial Intelligence Expert Predictive System (AIEPS) model for acute fish (fathead minnow) toxicity**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Artificial Intelligence Expert Predictive System (AIEPS) model for
acute fish (fathead minnow) 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:

18 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:

Neither the model nor training set is 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:

Fathead minnow (Pimephales promelas)

3.2.Endpoint:

3.3.Acute toxicity to fish (lethality) C.1.Acute Toxicity for Fish

3.3.Comment on endpoint:

Fathead minnow 96h LC50 - concentration of test chemical that kills 50% of the test subjects following a 96-h exposure

3.4.Endpoint units:

mmol/L or mg/L

3.5.Dependent variable:

The relationship between fathead minnow 96h 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:

Mainly, the data has been secured from the AQUIRE database, which is part of ECOTOX knowledge database (US EPA http://cfpub.epa.gov/ecotox/). Data was accepted as presented in the database for 835 chemical compounds randomly selected through computational generation from 921 compounds.

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 bromine atoms count number of bromine atoms

[2]number of carbon atoms count number of carbon atoms

[3]number of chlorine atoms count number of chlorine atoms

[4]number of copper atoms count number of copper 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 mercury atoms count number of mercury atoms

[9]number of iodine atoms count number of iodine 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 sulfur atoms count number of sulfur atoms

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

[16]number of zinc atoms number of zinc atoms

[17]number of methyl groups number of methyl groups

[18]number of triple bonds between carbon atoms number of triple bonds between carbon atoms

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

[20]number of N=C=O groups number of N=C=O groups

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

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

[23]number of nitrile groups C#N which are not carbonitriles number of nitrile groups C#N which are not carbonitriles

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

[25]number of amine groups connected to a ring carbon number of amine groups connected to a ring carbon

[26]number of amine groups connected to a non-ring carbon, excluding those included inside amide groups number of amine groups connected to a non-ring carbon, excluding those included inside amide groups

[27]number of amine groups excluding those connected to carbons which are not part of amides number of amine groups excluding those connected to carbons which are not part of amides

[28]number of amide groups connected to ring carbons number of amide groups connected to ring carbons

[29]number of amide groups connected to non-ring carbons number of amide groups connected to non-ring carbons

[30]number of amide groups excluding those connected to carbons number of amide groups

excluding those connected to carbons
[31]number of halogen atoms connected to ring carbons number of halogen atoms connected to ring carbons
[32]number of halogen atoms connected to non-ring carbons number of halogen atoms connected to non-ring carbons
[33]number of hydroxyl groups connected to ring carbons number of hydroxyl groups connected to ring carbons
[34]number of hydroxyl groups connected to non-ring carbons number of hydroxyl groups connected to non-ring carbons
[35]number of C-O- groups where the carbon is part of a ring and the oxygen is not connected to a hydrogen number of C-O- groups where the carbon is part of a ring and the oxygen is not connected to a hydrogen
[36]number of carboxyl groups connected to ring carbons number of carboxyl groups connected to ring carbons
[37]number of carboxyl groups not connected to ring carbons number of carboxyl groups not connected to ring carbons
[38]number of carboxy bridges excluding carboxyls and esters number of carboxy bridges excluding carboxyls and esters
[39]number of C-N(=O)=O groups where the carbon is part of a ring number of C-N(=O)=O groups where the carbon is part of a ring
[40]number of -O-N(=O)=O groups g/mole number of -O-N(=O)=O groups
[41]number of N(=O)=O groups excluding the ones connected to ring carbons and nitrates number of N(=O)=O groups excluding the ones connected to ring carbons and nitrates
[42]number of N=O groups excluding the ones part of N(=O)=O groups number of N=O groups excluding the ones part of N(=O)=O groups
[43]number of hydroxyl groups connected to nitrogens number of hydroxyl groups connected to nitrogens
[44]number of halogen atoms connected to nitrogens number of halogen atoms connected to nitrogens
[45]number of ether bridges excluding ester bridges number of ether bridges excluding ester bridges

[46]number of ester bridges number of ester bridges
[47]number of C=O groups where the carbon is part of a ring number of C=O groups where the carbon is part of a ring
[48]number of aldehyde groups number of aldehyde groups
[49]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.
[50]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
[51]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

[52]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
[53]number of S=O groups not part of S(=O)=O bridges number of S=O groups not part of S(=O)=O bridges
[54]number of hydrogen atoms connected to sulphur number of hydrogen atoms connected to sulphur
[55]number of S-C groups not included in thiocyanat number of S-C groups not included in thiocyanat
[56]number of S=C groups not included in isothiocyanat number of S=C groups not included in isothiocyanat
[57]number of P=O groups number of P=O groups
[58]number of P=S groups number of P=S groups
[59]number of P-S groups number of P-S groups
[60]number of P-N groups number of P-N groups
[61]number of P-O- groups except P-OH number of P-O- groups except P-OH
[62]number of hydroxyl groups connected to phosphorus number of hydroxyl groups connected to phosphorus
[63]number of single carbon-metal bonds number of single carbon-metal bonds
[64]number of single oxygen-metal bonds number of single oxygen-metal bonds
[65]number of single sulphur-metal bonds number of single sulphur-metal bonds
[66]number of quinone groups number of quinone groups
[67]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
[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 (empirical variable) 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 (empirical variable)
[69]number of carbon atoms in rings number of carbon atoms in rings
[70]number of nitrogen atoms in rings number of nitrogen atoms in rings
[71]number of sulphur atoms in rings number of sulphur atoms in rings
[72]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
[73]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
[74]ratio between the cumulative number of nitrogen and oxygen atoms in the molecule over the cumulative number of nitrogen, oxygen and carbon atoms (empirical variable) ratio between the cumulative number of nitrogen and oxygen atoms in the molecule over the cumulative number of nitrogen, oxygen and carbon atoms (empirical variable)
[75]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
[76]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
[77]number of vinyl groups number of vinyl groups

[78]molecular weight g/mol molecular weight

4.4.Descriptor selection:

78 descriptors were chosen in the final model. A random initial descriptor list was molecular weight as well as atoms, fragments and functional groups. The initial data set for fathead minnow consisted of 921 compounds, from which 86 were randomly selected through computational generation to form the external test set. To select the descriptors, the atoms or groups poorly represented or absent in the structures of the 835 study fathead minnow training dataset were eliminated from the list. Partial neural network learning experiments were conducted to identify the influence of these descriptors. Through the neural network training and statistical analysis, it was evident which descriptors had no impact on the resulting models behavior or resulted in a model with weaker overall generalization capability and these descriptors were removed.

4.5.Algorithm and descriptor generation:

See attachment (AIEPS 3.0 - Fathead minnow 96hr LC50 PNN Model Validation Study.doc), section 3, 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 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 835/78 = 10.71

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 is limited to organics.

As the available data on organometallics was very limited, caution is recommended when using the model for predicting the endpoint value for organometallics.

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 of 835 organic substances was randomly selected, through computer generation, from 921 substances having measured data for acute 96 hr LC50s with fathead minnow, Pimephales promelas.

6.6.Pre-processing of data before modelling:

There has been no preprocessing of data before modelling

6.7.Statistics for goodness-of-fit:

Minimum Residuals -2.3179

Maximum Residuals 1.5963

Average Residuals 0.0000

Standard Deviation of Residuals 0.5018

Sum of Square Residuals 210.0455

Average Square Residuals 0.2516

Coefficient of Determination Between Measured and Predicted Values 0.8844

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:

86 substances were used as the external validation set for the acute fish toxicity PNN model.

These were randomly determined thorugh a standard computational algorithm from the whole dataset of 921 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 -3.1563

Maximum Residuals 1.3460

Average Residuals -0.0614

Standard Deviation of Residuals 0.7511

Sum of Square Residuals 48.2808

Average Square Residuals 0.5614

Determination Coefficient Between Measured and Predicted Values 0.7766

Correlation Coefficient Between Measured and Predicted Values 0.8812

Shapiro-Wilk W Test Statistic 0.9222

Prob<W <0.0001

7.8.Predictivity - Assessment of the external validation set:

The Shapiro-Wilk W Test rejects the null hypothesis that the distribution of the residuals on the external test set of 86 compounds is normal at ?=0.05 significance level. Computation of confidence intervals cannot be implemented. The compound 899 (CAS RN 1484135 i.e. N-Vinylcarbazole) is not properly represented inside the training set and is a structural outlier.

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 involves the identification of presence or absence of the chemical descriptors (76) used to train the model with those in the substance of interest. The algorithm from the trained model is applied with the appropriate weights assigned to each factor reflecting the influence of those factors on the endpoint of interest. The result provides a large scope prediction of the 96hr LC50 for fathead minnow.

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 descriptors.

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

[6]ECOTOX. U.S. Environmental Protection Agency. 2002. ECOTOX User Guide: ECOTOXicology Database System. Version 3.0. http:/www.epa.gov/ecotox/

9.3.Supporting information:

qmrf521_AIEPS 3.0 - Fish Training Set_835.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-C1-0033/attachment/A1116
qmrf521_AIEPS 3.0 -Fish Validation_88.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-C1-0033/attachment/A1117

Test set(s)

10.Summary (JRC QSAR Model Database)

Supporting information

10.1.QMRF number:

Q17-C1-0033

10.2.Publication date:

2017-09-27

10.3.Keywords:

Artificial Intelligence Expert Predictive System;AIEPS;daphnia magna;fish;fathead minnow;acute toxicity;

10.4.Comments:

old# Q52-55-56-521