	QMRF identifier (JRC Inventory): **Q17-105-0032**
	QMRF Title: **Artificial Intelligence Expert Predictive System (AIEPS) model for aqueous solubility**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Artificial Intelligence Expert Predictive System (AIEPS) model for aqueous solubility

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) 289-298. 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. 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

2.8.Availability of information about the model:

The model is not proprietary. The Accelrys Chemistry Control Runtime might be required to run the interface. Consult with Accelrys/Biovia.

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

N/A

3.2.Endpoint:

1.3.Water solubility 105Water Solubility

3.3.Comment on endpoint:

The maximum concentration of a chemical dissolved in water at a given temperature and pressure, when the water is both in contact and at equilibrium with the pure substance. The model implemented in the expert system targeting the prediction of aqueous solubility of non-miscible substances at room temperature (25 C), neutral pH, and 760 mm Hg atmospheric pressure is a proof-of-concept model, and is based only on a very limited data set.

3.4.Endpoint units:

mmol/L or mg/L

3.5.Dependent variable:

The relationship between water solubility and selected molecular fragment descriptors is implemented through a basic Probabilistic Neural Network (PNN) with Gaussian kernel [sect.9.2/ ref.1] (statistical corrections included). Atoms and fragments information is generated directly from molecular structure using fragment chemistry data mining. The model may handle both inorganic and organic compounds.

3.6.Experimental protocol:

Not specified

3.7.Endpoint data quality and variability:

Original source of the data was known and recorded. The conditions of the solubility study had to have been conducted at room temperature (~25 C), neutral pH, 760 mm Hg atmospheric pressure. Each time the relationships between temperature, pH, and solubility info associated with the replicates reveals presence of conflicts then the associated information on that compound is labeled as unreliable and discarded. The implemented PNN model targeting aqueous solubility uses a training/learning set consisting of 2400 compounds randomly selected from the 2558 chemicals data set.

References: [sect.9.2/ ref.6, 7, 8, 9]

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

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
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 fluorine atoms count number of fluorine atoms
[5]number of iron atoms count number of iron atoms
[6]number of hydrogen atoms count number of hydrogen atoms
[7]number of mercury atoms count number of mercury atoms
[8]number of iodine atoms count number of iodine atoms
[9]number of nitrogen atoms count number of nitrogen atoms
[10]cumulative number of sodium, potassium and lithium atoms count cumulative number of sodium, potassium and lithium atoms
[11]number of oxygen atoms count number of oxygen atoms
[12]number of phosphorus atoms count number of phosphorus atoms
[13]number of sulphur atoms count number of sulphur atoms
[14]number of tin atoms count number of tin atoms
[15]number of zinc atoms count number of zinc atoms
[16]number of triple bonds between carbon atoms count number of triple bonds between carbon atoms
[17]number of C-C#N groups count number of C-C#N groups
[18]number of S-C#N groups count number of S-C#N groups
[19]number of C#N groups, carbonitrile excluded count number of C#N groups, carbonitrile excluded

[20]number of amine groups, excluding those included inside amide groups count number of amine groups, excluding those included inside amide groups
[21]number of amide groups count number of amide groups
[22]number of hydroxyl groups connected to carbons count number of hydroxyl groups connected to carbons
[23]number of C-O- groups where the carbon is part of a ring and the oxygen is not connected to a

hydrogen count number of C-O- groups where the carbon is part of a ring and the oxygen is not connected to a hydrogen
[24]number of carboxyl groups count number of carboxyl groups
[25]number of carboxy bridges excluding carboxyls and esters count number of carboxy bridges excluding carboxyls and esters
[26]number of double nitrogen-oxygen bonds count number of double nitrogen-oxygen bonds
[27]number of nitrogen-halogens bonds count number of nitrogen-halogens bonds
[28]number of ether bridges excluding ester bridges count number of ether bridges excluding ester bridges
[29]number of ester bridges number of ester bridges
[30]number of aldehyde groups number of aldehyde groups
[31]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
[32]number of double sulphur-oxygen bonds number of double sulphur-oxygen bonds
[33]number of single sulphur-oxygen bonds number of single sulphur-oxygen bonds
[34]number of sulphur-hydrogen bonds number of sulphur-hydrogen bonds
[35]number of double phosphorus-oxygen bonds number of double phosphorus-oxygen bondsv
[36]number of single phosphorus-oxygen bonds number of single phosphorus-oxygen bonds
[37]number of double phosphorus-sulphur bonds number of double phosphorus-sulphur bonds
[38]number of single phosphorus-sulphur bonds number of single phosphorus-sulphur bonds
[39]number of single carbon-metal bonds number of single carbon-metal bonds
[40]number of single sulphur-metal bonds count number of single sulphur-metal bonds
[41]number of quinone groups number of quinone groups
[42]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
[43]number of carbon atoms in rings number of carbon atoms in rings
[44]number of nitrogen atoms in rings number of nitrogen atoms in rings
[45]number of sulphur atoms in rings number of sulphur atoms in rings
[46]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)(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 (empirical variable)(1 for inorganics)
[47]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
[48]number of fluorine atoms connected to non-ring carbons number of fluorine atoms connected to non-ring carbons
[49]number of chlorine atoms connected to non-ring carbons number of chlorine atoms connected to non-ring carbons
[50]number of bromine atoms connected to non-ring carbons number of bromine atoms connected to non-ring carbons
[51]number of iodine atoms connected to non-ring carbons number of iodine atoms connected to non-ring carbons

[52]number of fluorine atoms connected to a carbon of a ring number of fluorine atoms connected to a carbon of a ring

[53]number of chlorine atoms connected to a carbon of a ring number of chlorine atoms connected to a carbon of a ring

[54]number of bromine atoms connected to a carbon of a ring number of bromine atoms connected to a carbon of a ring

[55]number of iodine atoms connected to a carbon of a ring number of iodine atoms connected to a carbon of a ring

[56]number of oxygen atoms in rings number of oxygen atoms in rings

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

[58]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

[59]molecular weight g/mole molecular weight

4.4.Descriptor selection:

59 descriptors were chosen in the final model The list started with those descriptors (78) used to develop a PNN model for estimating fathead minnow LC50. These were based on structural information in the training set. Refinements of the descriptors were based on partial model experiments where combinations of descriptors on approximately 80% of the training substances (random selection) was examined on the remaining 20% for potential impact on the predictive quality. Through this process descriptors were eliminated or added. The aqueous solubility model uses simpler more unrestrictive descriptors than the fathead minnow acute (LC50) toxicity model.

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 models 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 models 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 400 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 atoms and fragments not accounted for by the existing combination of 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:

Computer generated random selection of available data on aqueous solubility from a variety of organic and inorganic compounds from a data set of 2400 chemicals.

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

Minimum Residuals -3.8184

Maximum Residuals 2.4595

Average Residuals 0.0000

Standard Deviation of Residuals 0.6452

Average Absolute Residuals 0.4819

Sum of Square Residuals 998.8095

Average Square Residuals 0.4162

Coefficient of determination between measured and predicted values 0.9259

Coefficient of correlation between measured and predicted values 0.9622

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:

158 substances were used as the validation set for the Aqueous solubility PNN model.

These were selected randomly from the whole dataset of 2558 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 -2.0503
Maximum Residuals 2.5155
Average Residuals 0.0161
Standard Deviation Residuals 0.8138
Average Absolute Residuals 0.6212
Sum Square Residuals 104.0254
Average Square Residuals 0.6584
Coefficient of determination between measured and predicted values 0.8688
Coefficient of correlation between measured and predicted values 0.9321
Shapiro-Wilk W Test Statistic 0.9788
Prob<W 0.3221

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 158 compounds is normal at ?=0.05 significance level. Consequently, implementation of computation of confidence intervals for the unknown measured aqueous solubility values for new compounds using the inverse linear regression of measured values versus the present model's predictions on the external test set is sound. As the 158 compounds external validation set may not be representative for the slice of the chemical universe the model is supposed to handle, caution is advised in using such confidence intervals for decision purposes.

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

[6]Environment Canada (1995). COMPUTOX Toxicity Database version 5.0. National Water Research Institute, Burlington, Ontario

[7]SRC (2004). PhysProp Database, Syracuse Research Corporation, Syracuse, NY.

[8]TerraBase Inc. (1997). TerraTox/TeraFit Software Suite 1.504. TerraBase Inc., Burlington, ON.

[9]Yalkowski SH, Dannenfelser RM (1992). AQUASOL Database version 5, College of Pharmacy, University of Arizona, Tucson.

9.3.Supporting information:

qmrf522_AIEPS 3.0 - Aqueous Sol Training Set_2400.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-105-0032/attachment/A1114
qmrf522_AIEPS 3.0 -Aqueous Sol Validation_163.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-105-0032/attachment/A1115

Test set(s)

10.Summary (JRC QSAR Model Database)

Supporting information

10.1.QMRF number:

Q17-105-0032

10.2.Publication date:

2017-09-27

10.3.Keywords:

Artificial Intelligence Expert Predictive System;AIEPS;aqueous solubility;

10.4.Comments:

old# Q53-55-56-522