	QMRF identifier (JRC Inventory): **Q17-42-0038**
	QMRF Title: **BIOVIA toxicity prediction model – rat oral LD50**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

BIOVIA toxicity prediction model – rat oral LD50

1.2.Other related models:

None.

1.3.Software coding the model:

BIOVIA Discovery Studio v4.5

Optimize your drug discovery process with a flexible application that delivers predictive science to its required depth.

Dassault Systèmes, BIOVIA Corp., 5005 Wateridge Vista Drive, San Diego, CA92121, USA http://www.3dsbiovia.com

2.General information

2.1.Date of QMRF:

9/4/2015

2.2.QMRF author(s) and contact details:

Deqiang Zhang Dassault Systemes, BIOVIA Corp. 5005 Wateridge Vista Drive, San Diego, CA 92121, USA Deqiang.Zhang@3ds.com http://www.3dsbiovia.com

2.3.Date of QMRF update(s):

N/A

2.4.QMRF update(s):

N/A

2.5.Model developer(s) and contact details:

Deqiang Zhang Dassault Systemes, BIOVIA Corp. 5005 Wateridge Vista Drive, San Diego, CA 92121, USA Deqiang.Zhang@3ds.com http://www.3dsbiovia.com

2.6.Date of model development and/or publication:

2015

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

BIOVIA Discovery Studio v4.5 http://www.3dsbiovia.com/products/discovery-studio/

2.8.Availability of information about the model:

The model and data are proprietary (available as a commericcial product), but the algorithm is public. The training set is also proprietary, however, it is embedded with the model and can be retrieved with similarity search when a prediction is conducted. No external test is conducted except cross-validation.

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

Q32-48-43-425 ACD/Percepta model for rat acute oral toxicity

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rat (Rattus rattus or Rattus norvegicus)

3.2.Endpoint:

4.Human Health Effects 4.2.Acute Oral toxicity

3.3.Comment on endpoint:

LD stands for "Lethal Dose". LD50 is the amount of a material, given all at once, which causes the death of 50% (one half) of a group of test animals. The LD50 is one way to measure the short-term poisoning potential (acute toxicity) of a material. Toxicologists can use many kinds of animals but most often testing is done with rats and mice. The LD50 can be found for any route of entry or administration but dermal (applied to the skin) and oral (given by mouth) administration methods are the most common.

.

3.4.Endpoint units:

LD50 is usually expressed as the amount of chemical administered (e.g., milligrams) per 100 grams (for smaller animals) or per kilogram (for bigger test subjects) of the body weight of the test animal.

3.5.Dependent variable:

pLD50 = -log(LD50)

3.6.Experimental protocol:

Generally, the LD50 assay consists of a single administration of the chemical to several groups of rats at different dosages, followed by a 14-day observation period. The LD50 value is calculated from the number of deaths in the different dosage groups.

3.7.Endpoint data quality and variability:

This model was trained using 3597 experimental rat oral LD50 values from open literatures selected after critical review of experimental data.
Most of the data were extracted from various editions of the Registry of Toxic Effects of Chemical Substances (RTECS). To minimize data errors from that source, two types of procedures were performed:
A sample of data from RTECS was checked against the original literature. Deviations were found in less than 1% of the case.
Those compounds found to be outliers during the modeling process were also checked against their literature values. A few inconsistencies were corrected.
It is important to consider that RTECS lists the most toxic value when there are multiple values exit. For this reason, this model will tend to overestimate the toxicity of a query structure.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Partial least squares regression

4.2.Explicit algorithm:

Partial least squares regression

Partial least squares regression is a multivariate linear regression method that takes into account the latent structure in both the dependent variable and the explanatory variables. As in multiple linear

regression, the main purpose of PLS regression is to build a linear model: Y = X × B + E where Y is a response matrix (or vector) formed by the dependent variables, X is a matrix formed by the independent variables, B is a matrix of the regression coefficients, and E is an error term for the model. Usually, the variables in X and Y are centered by subtracting their means and scaled by dividing by their standard deviations. In PLS regression, a procedure called factor extraction is applied to produce a new matrix: T = X × W T and W are called the factor score matrix and the weight matrix, respectively. A new linear regression model is represented as: Y = T × Q + E, where Q is a matrix of regression coefficients (called loadings) for T, and E is an error (noise) term. Once the loadings Q are computed, the above regression model is equivalent to the predictive regression model Y = X * B + E, where B = W * Q. In a principal component analysis, a set of principal components can be obtained by diagonalizing the covariance matrix of the independent predictor variables. This is done similarly in PLS regression, with the exception that the covariance matrix includes both the predictor and response variables. For establishing the model, PLS regression produces a weight matrix W for X such that T = X × W, i.e., the columns of W are weight vectors for the X columns producing the corresponding factor score matrix T. These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of Y on T are then performed to produce Q, the loadings for Y (or weights for Y). One additional matrix which is necessary for a complete description of PLS regression procedures is the factor loading matrix P which gives a factor model X = T × P + F, where F is the unexplained part of the X scores. The true regression is done on a small number of latent variables in PLS regression. As a result, PLS is capable of handling a large number of independent variables without overfitting.

The equation contains 20 latent variables. Each latent variable is a linear combination of the input descriptors. The equation is too long to be included here.

4.3.Descriptors in the model:

[1]ALogP unitless The calculated partition-coefficient of a compound between 1-octanol and water

[2]Molecular_Weight gram/mole The calculated molecular weight by summing the average atomic weight of all the atoms in the molecule.

[3]Num_H_Donors unitless Number of hydrogen bond donors.

[4]Num_H_Acceptors unitless Number of hydrogen bond acceptors in the molecule.

[5]Num_RotatableBonds unitless Number of rotatable bonds in the molecule.

[6]Molecular_PolarSurfaceArea Angstrom-squared The polar surface areaof the molecule.

[7]Num_AromaticRings unitless Number of aromatic rings in the structure.

[8]ECFP_6 Unitless Extended-connectivity fingerprint with a maximum length of 6 bonds

[9]FCFP_6 Unitless Functional class fingerprint with a maximum length of 6 bonds

4.4.Descriptor selection:

A pool of most commonly used descriptors (ALogP, Molecule_Weight, Num_H_Donors, Num_H_Acceptors, Num_RotatableBonds, Num_AromaticRings, Molecular_PolarSurfaceArea, ECFP_2, ECFP_4, ECFP_6, ECFP_8, ECFP_10, ECFP_12, FCFP_2, FCFP_4, FCFP_6, FCFP_8, FCFP_10, FCFP_12, SCFP_2, SCFP_4, SCFP_6, SCFP_8, SCFP_10, SCFP_12, MDLPublicKeys) were selected randomly to build models. The model with the best 20-fold cross-validated q-squared score is selected to build the final model.

The number of components (ltent variables) is also set based on the cross-validated q-squared.

4.5.Algorithm and descriptor generation:

(1) The ALogP is the Ghose/Crippen group-contribution estimate for LogP, where P is the relative solubility of a compound in octanol versus water. See Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J., "Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using
Fragment Methods: An Analysis of AlogP and CLogP Methods." J. Phys. Chem. A, 1998, 102, 3762-3772.
(2) Molecular weight is calculated using the atomic weights of the individual atoms in the molecule.
(3) Hydrogen bond acceptors are defined as heteroatoms (O, N, S, or P) with one or more lone pairs, excluding atoms with positive formal charges, amide and pyrrole-type nitrogens, and aromatic oxygen and sulfur atoms in heterocyclic rings.
(4) Hydrogen bond donors are defined as heteroatoms (O, N, S, or P) with one or more attached hydrogen atoms.
(5) Molecular_FractionPolarSurfaceArea is calculated from the polar surface area and total surface area using a 2D approximation to each molecule.
(6) The fingerprint generation method is based on one of the original algorithms in computational organic chemistry called the Morgan algorithm. The goal of the Morgan algorithm is to assign a unique identity to each atom in a molecule so that a molecule can be described in a way that is invariant to the original numbering of atoms. The algorithm has two parts: the assignment of an initial code to each atom, and an iterative part in which each atom code is updated to reflect the codes of each atom's neighbors.
(1) The ALogP is the Ghose/Crippen group-contribution estimate for LogP, where P is the relative solubility of a compound in octanol versus water. See Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J., "Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using
Fragment Methods: An Analysis of AlogP and CLogP Methods." J. Phys. Chem. A, 1998, 102, 3762-3772.
(2) Molecular weight is calculated using the atomic weights of the individual atoms in the molecule.
(3) Hydrogen bond acceptors are defined as heteroatoms (O, N, S, or P) with one or more lone pairs, excluding atoms with positive formal charges, amide and pyrrole-type nitrogens, and aromatic oxygen and sulfur atoms in heterocyclic rings.
(4) Hydrogen bond donors are defined as heteroatoms (O, N, S, or P) with one or more attached hydrogen atoms.
(5) Molecular_PolarSurfaceArea is the polar surface area calculated using a 2D approximation to each molecule.
(6) Num_AromaticRings is the count of aromatic rings in the molecule.(7) The fingerprint generation method is based on one of the original

algorithms in computational organic chemistry called the Morgan algorithm. The goal of the Morgan algorithm is to assign a unique identity to each atom in a molecule so that a molecule can be described in a way that is invariant to the original numbering of atoms. The algorithm has two parts: the assignment of an initial code to each atom, and an iterative part in which each atom code is updated to reflect the codes of each atom's neighbors.ECFP_6 and FCFP_6 are calculated by first assigning atom types (ECFP_0 and FCFP_0) using atom type and functional class rule, and an n iterative process is used to generate features that represent each atom in progressively larger structural neighborhoods. After each iteration, the new feature codes for the atoms are added to the set of features from all previous steps. The process completes when the desired size is reached and the set of all features is returned as the fingerprint. MDLPublicKeys are bitset fingerprints calculated by searching the structure using predefined queries representing the 166 MDL public keys.

4.6.Software name and version for descriptor generation:

Dassult Systemes BIOVIA Pipeline Pilot Server

Built on the BIOVIA Foundation, Pipeline Pilot enables scientists to rapidly create, test and publish scientific services that automate the process of accessing, analyzing and reporting scientific data, either for the scientist’s personal use or for sharing across the scientific community. Using Pipeline Pilot, scientist, researchers, engineers, and analysts with little or no software development experience can create scientific protocols that can be executed through a variety of interfaces including Accelrys Web Port, other Accelrys solutions such as Accelrys Electronic Lab Notebook, Isentris, Chemical Registration and third-party applications such as Microsoft SharePoint or customer-developed applications. These protocols aggregate and provide immediate access to volumes of disparate research data locked in silos. They automate the scientific analysis of the data and enable researchers to rapidly explore, visualize and report results.

U.S. 6am -4pm (Pacific Time) Toll Free: 1-800-756- 4674 Tel: (858) 799-5509 support@accelrys.com U.K. 9:00 to 16:00 (UK time) Tel: +44 1223 228822 UK local rate: +44 845 741 3375 Central Europe 9:00 to 16:00 (Central European time) Switzerland: Tel: +41 61 588 0480 Germany: Tel: +49 221 8282 9020 support@accelrys.com Japan 10:00 to 17:00 (Tokyo time) Toll Free: 0120-712655 Tel: +81 3 4321 3906 support-japan@accelrys.com http://accelrys.com/products/pipeline-pilot/

4.7.Chemicals/Descriptors ratio:

Number of chemicals = 3597

Number of descriptors = 10

Chemicals/Descriptors = 359.7

Number of latent variables = 20

Number of chemicals/Number of latent variables = 179.85

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model is defined by the range of descriptors of training set chemicals. The applicability domain is only a qualititive measure on how reliable the prediction is. There is no

quantative measure on how reliable the prediction is.

5.2.Method used to assess the applicability domain:

If a continuous descriptor is out of range of the training set, a warning is issued for the input compound. For the fingerprint descriptors, if a new feature not seen in the training set is found, a warning message is issued for that feature.

5.3.Software name and version for applicability domain assessment:

Dassult Systemes BIOVIA Pipeline Pilot Server

Built on the BIOVIA Foundation, Pipeline Pilot enables scientists to rapidly create, test and publish scientific services that automate the process of accessing, analyzing and reporting scientific data, either for the scientist's personal use or for sharing across the scientific community. Using Pipeline Pilot, scientist, researchers, engineers, and analysts with little or no software development experience can create scientific protocols that can be executed through a variety of interfaces including Accelrys Web Port, other Accelrys solutions such as Accelrys Electronic Lab Notebook, Isentris, Chemical Registration and third-party applications such as Microsoft SharePoint or customer-developed applications. These protocols aggregate and provide immediate access to volumes of disparate research data locked in silos. They automate the scientific analysis of the data and enable researchers to rapidly explore, visualize and report results.

U.S. 6am -4pm (Pacific Time) Toll Free: 1-800-756- 4674 Tel: (858) 799-5509 support@accelrys.com U.K. 9:00 to 16:00 (UK time) Tel: +44 1223 228822 UK local rate: +44 845 741 3375 support@accelrys.com Central Europe 9:00 to 16:00 (Central European time) Switzerland: Tel: +41 61 588 0480 Germany: Tel: +49 221 8282 9020 support@accelrys.com Japan 10:00 to 17:00 (Tokyo time) Toll Free: 0120-712655 Tel: +81 3 4321 3906 support-japan@accelrys.com http://accelrys.com/products/pipeline-pilot/

5.4.Limits of applicability:

Property	Min	Max	Mean	Std. Dev.
ALogP	-7.685	13.41	2.1842	1.7647
Molecular_Weight	27.025	835.89	220.91	97.773
Num_H_Donors	0	14	0.72811	1.0062
Num_H_Acceptors	0	15	3.0373	1.9578
Num_RotatableBonds	0	27	4.0389	3.4324
Num_AromaticRings	0	7	0.75202	0.81322
Molecular_PolarSurfaceArea	0	331.42	53.759	37.497

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 data used to train the model consisted of 3597 samples. The training set is proprietary, however, it is emmbedded with the model and can be retrieved with similarity search when a prediction is conducted.

6.6.Pre-processing of data before modelling:

The 3597 experimental rat oral LD50 values were selected from open literatures after critical review of experimental data. Most of the data were extracted from various editions of the Registry of Toxic Effects of Chemical Substances (RTECS). To minimize data errors from that source, two types of procedures were performed:

A sample of data from RTECS was checked against the original literature. Deviations were found in less than 1% of the case.

Those compounds found to be outliers during the modeling process were also checked against their literature values. A few inconsistencies were corrected.

Generally, the LD50 assay consists of a single administration of the chemical to several groups of rats at different dosages, followed by a 14-day observation period. The LD50 value is calculated from the number of deaths in the different dosage groups.

It is important to consider that RTECS lists the most toxic value when there are multiple values exit. For this reason, this model will tend to overestimate the toxicity of a query structure.

6.7.Statistics for goodness-of-fit:

r = 0.777 r-squared = 0.604 r-squared (adjusted) = 0.602 RMS error = 0.582

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

N/A

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

20-fold cross-validation:

q-squared = 0.504

RMS error = 0.655

6.10.Robustness - Statistics obtained by Y-scrambling:

N/A

6.11.Robustness - Statistics obtained by bootstrap:

N/A

6.12.Robustness - Statistics obtained by other methods:

N/A

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3.Data for each descriptor variable for the external validation set:

No

7.4.Data for the dependent variable for the external validation set:

No

7.5.Other information about the external validation set:

N/A

7.6.Experimental design of test set:

N/A

7.7.Predictivity - Statistics obtained by external validation:

N/A

7.8.Predictivity - Assessment of the external validation set:

N/A

7.9.Comments on the external validation of the model:

N/A

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

These features contribute to the LD50 the most (Each of these features coresponding to a substructure in a molecule):

Feature Coefficient

Count<MDLPublicKeys:33> -0.443772

Count<ECFP_6:2106656448> -0.351738

Count<ECFP_6:226796801> -0.320405

Count<ECFP_6:497523368> -0.3011

Count<MDLPublicKeys:36> -0.283873

Count<ECFP_6:1887306650> -0.270714

Count<ECFP_6:683445015> -0.266394

Count<ECFP_6:-817402818> -0.262524

Count<ECFP_6:-175507738> -0.261889

Count<ECFP_6:-176455838> -0.256639

Count<FCFP_6:1676877079> -0.253628

Count<ECFP_6:2077607946> -0.251821Count<MDLPublicKeys:110> -0.240325

Count<ECFP_6:655739385> -0.238826

Count<ECFP_6:412256466> -0.228989Count<ECFP_6:989418220> -0.227505

Count<FCFP_6:-1096219292> -0.224905

Count<ECFP_6:2014710090> -0.224532Count<FCFP_6:907036844> -0.222143

Count<ECFP_6:2101483135> -0.221499

8.2.A priori or a posteriori mechanistic interpretation:

posteriori: these features are selected purely based on their coefficient appearing in the final equation

8.3.Other information about the mechanistic interpretation:

N/A

9.Miscellaneous information

9.1.Comments:

The model is extensible, i.e., it can be extended by feeding new training data to create an improved model.

9.2.Bibliography:

Wold S, Ruhe A, Wold H, Dunn WJ(1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing. 5(3) 735-743 http://dx.doi.org/10.1137%2F0905052

9.3.Supporting information:

qmrf502_qmrf440_RatOralLD50-equation.txt	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-42-0038/attachment/A1107

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-42-0038

10.2.Publication date:

2017-09-27

10.3.Keywords:

rat;acute oral toxicity;LD50;BIOVIA Discovery Studio;

10.4.Comments:

old# Q51-54-55-502