	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **QSARs for water solubility (LogS)**
	Printing Date: **Aug 24, 2016**

1.QSAR identifier

1.1.QSAR identifier (title):

QSARs for water solubility (LogS)

1.2.Other related models:

1.3.Software coding the model:

The R statistical computing environment for Windows (version 3.2.1)
http://cran.r-project.org/

2.General information

2.1.Date of QMRF:

August 19, 2016

2.2.QMRF author(s) and contact details:

[1]Qingda Zang, Integrated Laboratory Systems, Inc., dan.zang@nih.gov
[2]Nicole C. Kleinstreuer, National Toxicology Program, National Institute of Environmental Health Sciences, nicole.kleinstreuer@nih.gov
[3]Kamel Mansouri, National Center for Computational Toxicology, Office of Research and Development, the U.S. Environmental Protection Agency, mansouri.kamel@epa.gov
[4]Antony J. Williams, National Center for Computational Toxicology, Office of Research and Development, the U.S. Environmental Protection Agency, Williams.Antony@epa.gov
[5]Richard S. Judson, National Center for Computational Toxicology, Office of Research and Development, the U.S. Environmental Protection Agency, Judson.Richard@epa.gov
[6]David G. Allen, Integrated Laboratory Systems, Inc., dallen@ils-inc.com
[7]Warren M. Casey, National Toxicology Program, National Institute of Environmental Health Sciences, warren.casey@nih.gov

2.3.Date of QMRF update(s):

This is a new QMRF.

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1]Qingda Zang, Integrated Laboratory Systems, Inc., dan.zang@nih.gov
[2]Nicole C. Kleinstreuer, National Toxicology Program, National Institute of Environmental Health Sciences, nicole.kleinstreuer@nih.gov

2.6.Date of model development and/or publication:

August 19, 2016

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

[1]Qingda Zang, Kamel Mansouri, Antony J. Williams, Richard S. Judson, David G. Allen, Warren Casey, and Nicole C. Kleinstreuer. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning (manuscript submitted to Journal of Chemical Information and Modeling)
[2]The R statistical computing environment for Windows (version 3.2.1) http://cran.r-project.org/

2.8.Availability of information about the model:

Algorithms are available.

Training and test sets are available.

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

NA

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Not applicable

3.2.Endpoint:

QMRF 1. Physical Chemical Properties QMRF 1. 3. Water solubility

3.3.Comment on endpoint:

End point data was based on experimental measurements contained in the US EPA Estimation Program Interface (EPI) Suite database.

3.4.Endpoint units:

Water solubility: mol/L

3.5.Dependent variable:

Water solubility (LogS)

3.6.Experimental protocol:

NA

3.7.Endpoint data quality and variability:

The data set was retrieved from US EPA EPI Suite (Estimation Program Interface).

LogS - Max: 1.58; Min: -12.06; Mean: -2.60; Deviation: 2.19.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

Support Vector Regression (SVR)

SVR can model both linear and non-linear relationships between the property and molecular descriptors by utilizing an appropriate kernel function to map the input variables from a lower dimensional space to a higher dimensional feature space and transform the non-linear relationship into a linear form. SVR with a Gaussian radial basis function (RBF) kernel was employed to explore the possible nonlinear dependency between molecular fingerprints and the property.

```
LogSmodel <- svm(LogS~., data=LogSdataTraining, cost = 260,
    epsilon = 0.145, gamma = 0.000031)
```

4.3.Descriptors in the model:

Molecular fingerprints: the chemicals were represented by fingerprints derived from their molecular structures. A total of 8097 binary bits were generated with 1 and 0 denoting the presence and absence of a specific structural fragment.

4.4.Descriptor selection:

To obtain reliable and robust regression models with high predictive performance, genetic algorithm (GA) was employed to select the most information-rich subset of fingerprint bits.

GA is an efficient stochastic optimization tool and randomized search technique, and can deal with a great number of descriptors and effectively select a subset from them.

4.5.Algorithm and descriptor generation:

A wide variety of fingerprints were calculated using publicly available SMARTS systems implemented in PADEL: Estate (79bits), Extended (1024 bits), Substructure (307 bits), Klekota Roth (4860 bits), PubChem (881 bits), Atom Pairs 2D (780 bits), and MACCS (166 bits).

Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 2011, *32(7)*, 1466-1474.

4.6.Software name and version for descriptor generation:

Software: PaDEL-Descriptor; Version: 2.21.

http://www.yapcwsoft.com/dd/padeldescriptor/

4.7.Chemicals/Descriptors ratio:

LogS: 1507/350 = 4.31

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The QSAR models were developed using training sets and thus, their applicability to external chemicals depends on the structural similarity between the training chemicals and the external test chemicals. The models are presumed to provide more reliable predictions for chemicals that fall in the AD, as defined by the three distance measures below. In this study, only if the thresholds from all three distance measures are exceeded is a test chemical deemed to be outside the AD. Otherwise, if only one or two thresholds are exceeded, the chemical is considered to be potentially outside the AD.

5.2.Method used to assess the applicability domain:

Three distance-based measures (i.e., leverage, distance from centroid and k-nearest neighbors (kNN)), were applied to assess the applicability domain (AD) of each regression model. The distance of a test chemical from a defined point in the descriptor space of the training set was calculated and compared to a predefined threshold. The test chemical is located inside the AD if its distance is less than or equal to the threshold. Leverage is the diagonal element of the covariance matrix for a given dataset, and the leverage of a test chemical is proportional to Hotellings T^2 statistic and its Mahalanobis distance. The threshold was set to three times the average of the leverage (3 *m*/*n*, with *m* being the number of variables and *n* the number of training chemicals). For the measure of distance from centroid, the distance of a test chemical from the training set centroid is compared with a threshold, which is

determined as follows: (1) calculate the distances of training chemicals from their centroid; (2) sort the vector of distances in ascending order; (3) set the distance value corresponding to 95th percentile as the threshold. The kNN measure defines the model's AD based on the similarity between a test chemical and the training chemicals. The average distance of the test chemical from its five nearest neighbors in the training set is compared with a threshold, which is the 95th percentile of average distance of training chemicals from their five nearest neighbors.

5.3.Software name and version for applicability domain assessment:

The R statistical computing environment for Windows (version 3.2.1) http://cran.r-project.org/

5.4.Limits of applicability:

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: Yes
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:

NA

6.6.Pre-processing of data before modelling:

Fingerprint bits with zero variance (i.e. uniform observations across the set) were removed. To obtain reliable models, sufficient occurrences of the fingerprint bits throughout the entire data sets are necessary, and thus bits with low occurrences were eliminated. Following the removal of highly correlated and sparsely occurring bits, finally 1061 bits were retained and employed to build the regression models.

6.7.Statistics for goodness-of-fit:

Coefficient of determination (R^2)
LogS: 0.983

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

NA

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

Coefficient of determination (R^2) (10-fold cross-validation)
LogS: 0.928

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

NA

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: Yes

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:

NA

7.6.Experimental design of test set:

The data sets were randomly partitioned into training sets (75% of the chemicals) and test sets (25% of the chemicals) to build the models and validate their predictive power, respectively. The distribution of the test set is very similar to that of the training set.

7.7.Predictivity - Statistics obtained by external validation:

Coefficient of determination (R^2)

LogS: 0.932

7.8.Predictivity - Assessment of the external validation set:

As shown in Section 7.7, the predictivity is high.

7.9.Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Here it is not practical to make an interpretation linking each and every selected fingerprint bit to the modeled endpoints. However, we assume that the statistically selected fingerprint bits represent fragments that are relevant to the studied endpoints.

8.2.A priori or a posteriori mechanistic interpretation:

NA

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

https://www.epa.gov/tsca-screening-tools

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC