
	<b>QMRF identifier (ECB Inventory):</b>	
	<b>QMRF Title:</b> <i>QSAR model for Acute toxicity to Daphnia magna (LC50)</i>	
	<b>Printing Date:</b> <i>Jun 8, 2010</i>	

## 1. QSAR identifier

### 1.1. QSAR identifier (title):

QSAR model for Acute toxicity to Daphnia magna (LC50)

### 1.2. Other related models:

-

### 1.3. Software coding the model:

QSARModel 4.0.4 Molcode Ltd., Turu 2, Tartu, 51014, Estonia <http://www.molcode.com>

## 2. General information

### 2.1. Date of QMRF:

19.05.2010

### 2.2. QMRF author(s) and contact details:

[1] Indrek Tulp Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[2] Tarmo Tamm Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[3] Gunnar Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[4] Dimitar Dobchev Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[5] Dana Martin Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[6] Kaido Tämm Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[7] Deniss Savchenko Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[8] Jaak Jänes Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[9] Eneli Härk Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[10] Andres Kreegipuu Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

[11] Mati Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com)  
<http://www.molcode.com>

### 2.3. Date of QMRF update(s):

-

### 2.4. QMRF update(s):

-

### 2.5. Model developer(s) and contact details:

Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia [models@molcode.com](mailto:models@molcode.com) <http://www.molcode.com>

## 2.6. Date of model development and/or publication:

12.05.2010

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

[1]Karelson M, Dobchev D, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Karelson G (2008). Correlation of blood-brain penetration and human serum albumin binding with theoretical descriptors. ARKIVOC 16, 38-60.

[2]Karelson M, Karelson G, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Dobchev D (2009). QSAR study of pharmacological permeabilities. ARKIVOC 2, 218–238.

## 2.8. Availability of information about the model:

All information in full detail is available

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

None to date

## 3. Defining the endpoint - OECD Principle 1

### 3.1. Species:

Daphnia magna (water flea)

### 3.2. Endpoint:

3. Ecotoxic effects 3.1. Short-term toxicity to Daphnia (immobilisation)

### 3.3. Comment on endpoint:

Acute toxicity 48h LC50 (50% of lethal concentration). This is the concentration which immobilizes 50% of the Daphnia in a test batch within 48 h.

### 3.4. Endpoint units:

mol/L

### 3.5. Dependent variable:

log(LC50)

### 3.6. Experimental protocol:

Acute toxicity for Daphnia is expressed as the median effective concentration EC50. The concentrations of the substances are given in mol per litre (mol/L). Those animals which are not able to swim within 15 seconds after gentle agitation of the test batch are considered to be immobile.

Some studies use mortality (LC50) and immobilization (EC50) as identical endpoints in the context of daphnid toxicity, as is, for example, reported in the toxicity analysis of parathion that is also included in the presently selected AQUIRE data set [2].

From the U.S. EPA database AQUIRE, acute toxicity values (48 h LC50) for the Daphnia magna were collected for a total of 380 compounds.

When multiple test values were found for one substance, these values were checked for consistency. If values differed by more than a factor of 30 from the closest one in a group of at least three other references, the aberrant value was discarded so as to remove outliers from the data set. Of all the remaining values for a given substance, the arithmetic mean was taken as the valid experimental value.

From the initial set of 1067 LC50 data, 77 values were excluded as outliers as described above, which led to a set of 349 chemicals with at least one LC50 value per substance. Subsequently, 49 chemicals were excluded because their LC50 values exceeded the predicted water solubility or because they contained metal atoms or were inorganic, leading to the final set of 300 organic compounds that cover a log Kow (octanol/water partition coefficient) range from -2 to 8.

### 3.7. Endpoint data quality and variability:

Experimental data from different labs has been used. On previously explained reason (see 3.6), the average experimental error, which accounts as well an error caused by interlaboratory differences, might be reasonably large. Since the authors do not provide the results from interlaboratory calibrations, it is difficult or even impossible to estimate exact error.

Statistics:

max value: -0.460

min value: -10.1

standard deviation: 1.75

skewness: -0.259

## 4. Defining the algorithm - OECD Principle 2

### 4.1. Type of model:

2D and 3D regression-based QSAR

### 4.2. Explicit algorithm:

multilinear regression QSAR

multilinear regression QSAR derived with BMLR (Best Multiple Linear Regression) method

$\log(\text{LC50}) = -4.904$

$-2.272 \cdot \text{Average Bonding Information content (order 2)}$

$+0.377 \cdot \text{HOMO - LUMO energy gap (AM1)}$

$+4.653\text{E-}003 \cdot \text{HPSA Polar (AM1) part of SASA}$

$-1.240\text{E-}002 \cdot \text{Molecular weight}$

$+0.256 \cdot \text{min}(\#\text{HA}, \#\text{HD}) \text{ (AM1)}$

### 4.3. Descriptors in the model:

[1] Average Bonding Information content (order 2) [unitless] Information theoretic index showing the complexity of structure

[2] HOMO - LUMO energy gap (AM1) [eV] Energy difference between highest occupied and lowest unoccupied molecular orbitals

[3] HPSA Polar (AM1) part of SASA [ $\text{\AA}^2$ ] Polar part of solvent accessible surface area

[4] Molecular weight [g/mol] Molecular weight

[5]  $\text{min}(\#\text{HA}, \#\text{HD})$  (AM1) [unitless] minimum value of the count of hydrogen-acceptor sites and the count of hydrogen-donor sites

### 4.4. Descriptor selection:

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on a set of statistical selection rules (one-parameter equations: Fisher criterion and  $R^2$  over threshold, variance and t-test value over threshold, intercorrelation with another descriptor not over threshold),

(two-parameter equations: intercorrelation coefficient below threshold, significant correlation with endpoint, in terms of correlation coefficient and t-test)

Stepwise trial of additional descriptors not significantly correlated to any already in the model.

### 4.5. Algorithm and descriptor generation:

1D, 2D, and 3D theoretical calculations. Quantum chemical descriptors derived from AM1 calculation. Model developed by using multilinear regression.

#### 4.6. Software name and version for descriptor generation:

QSARModel 4.0.4

QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling

Molcode Ltd, Turu 2, Tartu, 51014, Estonia

<http://www.molcode.com>

#### 4.7. Chemicals/Descriptors ratio:

38.8 (194 chemicals / 5 descriptors)

### 5. Defining the applicability domain - OECD Principle 3

#### 5.1. Description of the applicability domain of the model:

Applicability domain based on training set:

a) by chemical identity: Organic Compounds (hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives; isothiocyanates; thioles; phosphorothionate and phosphate esters; and halogenated derivatives)

b) by descriptor value range: The model is suitable for compounds that have the descriptors

in the following minimal-maximal range:

Average Bonding Information content (order 2): 0.279 - 0.976

HOMO - LUMO energy gap (AM1): 4.97 - 14.7

HPSA Polar (AM1) part of SASA: 0 - 392

Molecular weight: 44.1 - 505

min(#HA, #HD) (AM1): 0 - 5

#### 5.2. Method used to assess the applicability domain:

Chemicals in the same structural domain as training set (similar functionality)

Range of descriptor values in training set with  $\pm 30\%$  confidence. Descriptor values must fall between maximal and minimal descriptor values of training set  $\pm 30\%$ .

#### 5.3. Software name and version for applicability domain assessment:

QSARModel 4.0.4

QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling

Molcode Ltd, Turu 2, Tartu, 51014, Estonia

<http://www.molcode.com>

#### 5.4. Limits of applicability:

See 5.1

### 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: No

Formula: Yes

INChI: No

MOL file:Yes

**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:**

Two compounds (2,4,6-trinitro-1,3-benzenediol, CAS:15245-44-0 and mancozeb, CAS:8018-01-7) were eliminated from original data because they are metal salts and they do not fit into applicability domain.

During the modeling procedure five compounds (paclobutrazol, CAS:76738-62-0; pirimiphos-methyl, CAS:29232-93-7; TEDP, CAS:3689-24-5; 2,4-dichlorophenoxyacetic acid, CAS:94-75-7, and dichlorvos, CAS:62-73-7) were excluded as a statistical outlier (which residuals exceeded 2 times standard deviation), final training set consist 194 data points.

194 negative values

0 positive values

**6.6.Pre-processing of data before modelling:**

n/a

**6.7.Statistics for goodness-of-fit:**

$R^2 = 0.741$  (Correlation coefficient)

$s^2 = 0.903$  (Standard error of the estimate)

$F = 108$  (Fisher function)

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

$R^2_{CV} = 0.725$

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

$R^2_{CVMO} = 0.719$

**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:**

ABC analysis (2:1 training : prediction) on sorted (in increased order of endpoint value) data divided into 3 subsets (A;B;C). Training set formed with 2/3 of the compounds (set A+B, A+C, B+C) and validation set consisted of 1/3 of the compounds (C, B, A).

average  $R^2$  (fitting) = 0.747

average  $R^2$  (prediction) = 0.712

**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:No

Formula:Yes

INChI:No

MOL file:Yes

**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:**

One compound () was excluded from test set because it does not fit into applicability domain with descriptor "min(#HA, #HD) (AM1)".

98 data points,

98 negative values,

0 positive values

**7.6.Experimental design of test set:**

From sorted data source each 3rd was subjected to the test set.

**7.7.Predictivity - Statistics obtained by external validation:**

R2 = 0.621 (Coefficient of determination)

**7.8.Predictivity - Assessment of the external validation set:**

After excluding that compound, the rest are all in range of applicability domain:

Average Bonding Information content (order 2): 0.393 - 0.980

HOMO - LUMO energy gap (AM1): 6.53 - 14.2

HPSA Polar (AM1) part of SASA: 0 - 353

Molecular weight: 41.1 - 420

min(#HA, #HD) (AM1): 0 - 6

**7.9.Comments on the external validation of the model:**

The validation coefficient of determination (R2) is relatively low but still acceptable bearing in mind the diversity of the compounds and the possible differences in experimental protocols (see 3.6 and 3.7). Also, large chemical diversity (complexity) in the test set affects R2. Investigation of descriptor value ranges of test set compounds reveals also, that quite often the values are on the edge of the applicability domain.

**8.Providing a mechanistic interpretation - OECD Principle 5**

**8.1.Mechanistic basis of the model:**

The toxicity baseline (as it is usually modeled by logP) is defined here with combination of "Molecular weight", "Average Bonding Information content (order 2)", "HPSA Polar (AM1) part of SASA" and "min(#HA, #HD) (AM1)". "Molecular weight" defines generally the mass and size of the structure; "Average Bonding Information content (order 2)" accounts the bonding complexity, i.e. aromatic, single, double, triple bonds, where also taking into account a heteroatoms; "HPSA Polar (AM1) part of SASA" shows the amount of polar surface area; and "min(#HA, #HD) (AM1)" counts the hydrogen bonding. All these descriptors affect more or less hydrophobicity - the baseline. Indirectly they are also related with other mode of action (like polar narcosis). For instance, heteroatoms, polar surface area and hydrogen bonding are important factors for different MOA. Specifically "HOMO - LUMO energy gap (AM1)" is defining the electronic hardness of molecules and is an important descriptor to define the deviation from baseline.

**8.2.A priori or a posteriori mechanistic interpretation:**

a posteriori mechanistic interpretation,

### 8.3. Other information about the mechanistic interpretation:

Interpretation consistent with scientific literature [1,3]

## 9. Miscellaneous information

### 9.1. Comments:

The data are gathered from different sources and therefore the quality of the data suffers. This means, that error term includes also a component from interlaboratory experimental differences. Thus, high quality QSAR models cannot be expected.

### 9.2. Bibliography:

[1] Structural Alerts A New Classification Model to Discriminate Excess Toxicity from Narcotic Effect Levels of Organic Compounds in the Acute Daphnid Assay  
<http://dx.doi.org/10.1021/tx0497954>

[2] U.S. Environmental Protection Agency (2002) AQUIRE (Aquatic Toxicity Information Retrieval Database), National Health and Environmental Effects Research Laboratory, Duluth, MN.

[3] Review of (Quantitative) Structure – Activity Relationships for Acute Aquatic Toxicity  
Tatiana I. Netzeva, Manuela Pavan and Andrew P. Worth QSAR Comb. Sci. 27, 2008, No. 1, 77 – 90 DOI: 10.1002/qsar.200710099

### 9.3. Supporting information:

Training set(s)

Daphnia#2_trainingset.sdf	<a href="http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf214_Daphnia#2_trainingset.sdf">http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf214_Daphnia#2_trainingset.sdf</a>
---------------------------	---

Test set(s)

Daphnia#2_testset.sdf	<a href="http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf214_Daphnia#2_testset.sdf">http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf214_Daphnia#2_testset.sdf</a>
-----------------------	---

## 10. Summary (ECB Inventory)

10.1. QMRF number:

10.2. Publication date:

10.3. Keywords:

10.4. Comments: