

Leadscope Enterprise model for acute toxicity to *Fathead minnow* (96h mortality, LC₅₀)

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for acute toxicity to *Fathead minnow* (96h mortality, LC₅₀), Danish QSAR Group at DTU Food.

1.2 Other related models

SciMatics SciQSAR model for acute toxicity to *Fathead minnow* (96h mortality, LC₅₀), Danish QSAR Group at DTU Food.

1.3. Software coding the model

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

2. General information

2.1 Date of QMRF

January 2015.

2.2 QMRF author(s) and contact details

QSAR Group at DTU Food;

Danish National Food Institute at the Technical University of Denmark;

<http://qsar.food.dtu.dk/>;

qsar@food.dtu.dk

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Marianne Dybdahl;

National Food Institute at the Technical University of Denmark;

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

2.3 Date of QMRF update(s)

2.4 QMRF update(s)

2.5 Model developer(s) and contact details

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Jay Russel Niemelä;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Danish QSAR Group at DTU Food;

National Food Institute at the Technical University of Denmark;

<http://qsar.food.dtu.dk/>;

qsar@food.dtu.dk

2.6 Date of model development and/or publication

January 2014.

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

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., and Blower, P. E. J. (2000) LeadScope: Software for Exploring Large Sets of Screening Data. *Chem. Inf. Comput. Sci.*, 40, 1302-1314.

Cross, K.P., Myatt, G., Yang, C., Fligner, M.A., Verducci, J.S., and Blower, P.E. Jr. (2003) Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.*, 46, 4770-4775.

Valerio, L. G., Yang, C., Arvidson, K. B., and Kruhlak, N. L. (2010) A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, 6:4, 505-518.

2.8 Availability of information about the model

The training set is non-proprietary and was compiled from the US EPA MED-Duluth Fathead minnow database (MED-Duluth) in 1999. The model algorithm is proprietary from commercial software.

2.9 Availability of another QMRF for exactly the same model

3. Defining the endpoint

3.1 Species

Fish (Fathead minnow, i.e. *Pimephales promelas*).

3.2 Endpoint

QMRF 3. Ecotoxic effects

QMRF 3. 3. Acute toxicity to fish (lethality)

OECD 203 Fish, Acute Toxicity Test

3.3 Comment on endpoint

Water pollution has become a major threat to the existence of living organisms in aquatic environment. A huge quantity of pollutants in the form of domestic and industrial effluents is discharged directly or indirectly into the water bodies, which has severe impacts on its biotic and abiotic environment. A typical endpoint used in initial effect assessment of a chemical on aquatic organisms is the 96h LC50 value for the fish, fathead minnow. The fathead minnow (*Pimephales promelas*) is a species of temperate freshwater fish belonging to the *Pimephelas* genus.

The training set consists of data for acute toxicity to fathead minnow from the US EPA MED-Duluth Fathead minnow database (MED-Duluth). MED-Duluth tested a series of industrial organic compounds using the fathead minnow for the purpose of developing an expert system to predict the acute mode of toxic action from chemical structure. The entire Duluth fathead minnow database and results related to the acute mode of action are presented in Russom *et al.* (1997), see references 15. to 19. in Russom *et al.* (1997).

3.4 Endpoint units

$-\log(\text{LC}_{50})$.

3.5 Dependent variable

Acute toxicity to fathead minnow (96h lethal concentration): LC_{50} , in μM .

3.6 Experimental protocol

The experimental protocol is described in OECD guideline 203 (1992). Briefly, the fish are exposed to the different concentrations of the test substance preferably for a period of 96 hours. Mortalities are recorded at 24, 48, 72 and 96 hours and the concentration that kills 50% of the fish (lethal concentration, LC_{50} , in mg/L) after 96 hours is estimated.

3.7 Endpoint data quality and variability

The data is of good quality and as all experimental results originate from the same source (MED-Duluth) the variability in data is expected to be low.

4. Defining the algorithm

4.1 Type of model

A continuous (Q)SAR model based on structural features and numeric molecular descriptors.

4.2 Explicit algorithm

This is a continuous (Q)SAR model made by use of partial least squares (PLS) regression. The specific implementation is proprietary within the Leadscope software.

4.3 Descriptors in the model

structural features,

aLogP,

polar surface area,

number of hydrogen bond donors,

Lipinski score,

number of rotational bonds,

parent atom count,

parent molecular weight,

number of hydrogen bond acceptors

4.4 Descriptor selection

Leadscope Predictive Data Miner is a software program for systematic sub-structural analysis of a chemical using predefined structural features stored in a template library, training set-dependent generated structural features (scaffolds) and calculated molecular descriptors. The feature library contains approximately 27,000 pre-defined structural features and the structural features chosen for the library are motivated by those typically found in small molecules: aromatics, heterocycles, spacer groups, simple substituents. Leadscope allows for the generation of training set-dependent structural features (scaffold generation), and these features can be added to the pre-defined structural features from the library and be included in the descriptor selection process. It is possible in Leadscope to remove redundant structural features before the descriptor selection process and only use the remaining features in the descriptor selection process. Besides the structural features Leadscope also calculates eight molecular descriptors for

each training set structure: the octanol/water partition coefficient (alogP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), Lipinski score, atom count, parent compound molecular weight, polar surface area (PSA) and rotatable bonds. These eight molecular descriptors are also included in the descriptor selection process.

Leadscope has a default automatic descriptor selection procedure. This procedure selects the top 30% of the descriptors (structural features and molecular descriptors) according to χ^2 -test for a binary variable, or the top and bottom 15% descriptors according to *t*-test for a continuous variable. Leadscope treats numeric property data as ordinal categorical data. If the input data is continuous such as IC₅₀ or cLogP data, the user can determine how values are assigned to categories: the number of categories and the cut-off values between categories. (Roberts *et al.*2000).

When developing this model, intermediate models with application of different modelling approaches in Leadscope were tried:

1. 'Single model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors for descriptor selection.
2. 'Single model' using both the Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.
3. 'Single model' using both Leadscope pre-defined structural features and the training set dependent features (scaffolds generation), with subsequent removal of redundant structural features, and calculated molecular descriptors for descriptor selection.

Based on model performance as measured by a preliminary cross-validation the model developed using approach number 3. was chosen.

For this model scaffolds were generated by Leadscope for the training set structures and added to the Leadscope library of structural features. The number of structural features was then reduced further using the built-in filter to remove similar (redundant) features (the "less similar" features removed). Descriptors were then automatically selected among the remaining structural features and the eight molecular descriptors.

4.5 Algorithm and descriptor generation

For descriptor generation see 4.4.

After selection of descriptors the Leadscope Predictive Data Miner program performs partial least squares (PLS) regression for a continuous response variable, or partial logistic regression (PLR) for a binary response variable, to build a predictive model. By default the Predictive Data Miner performs leave-one-out or leave-groups-out (in the latter case, the user can specify any number of repetitions and percentage of structures left out) cross-validation on the training set depending on the size of the training set. In the cross-validation made by Leadscope the descriptors selected for the 'mother model' are used when building the validation submodels and they therefore have a tendency to be overfitted and give overoptimistic validation results.

In this model because of the continuous outcome in the response variable PLS regression was used to build the predictive model. For this model 147 descriptors were selected to build the model. These include 9

Leadscope calculated molecular descriptors, 80 hierarchy features, and 58 dynamic features. The 147 descriptors were distributed on 4 PLS factors.

4.6 Software name and version for descriptor generation

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

4.7 Descriptors/chemicals ratio

In this model 147 descriptors were used and distributed on 4 PLS factors. The training set consists of 364 compounds. The descriptor/chemical ratio is 1:2.5 (147:364).

5. Defining Applicability Domain

5.1 Description of the applicability domain of the model

The definition of the applicability domain consists of two components; the definition of a structural domain in Leadscope and the in-house further probability refinement algorithm on the output from Leadscope to reach the final applicability domain call.

1. Leadscope

For assessing if a test compound is within the structural applicability domain of a given model Leadscope examines whether the test compound bears enough structural resemblance to the training set compounds used for building the model (i.e. a structural domain analysis). This is done by calculating the distance between the test compound and all compounds in the training set (distance = 1 - similarity). The similarity score is based on the Tanimoto method. The number of neighbours is defined as the number of compounds in the training set that have a distance equal to or smaller than 0.7 with respect to the test compound. The higher the number of neighbours, the more reliable the prediction for the test compound. Statistics of the distances are also calculated. Effectively no predictions are made for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated in Leadscope.

2. The Danish QSAR group

In addition to the general Leadscope structural applicability domain definition the Danish QSAR group has applied two further requirements to the applicability domain of the model. First, the logP value of the query compound should fall within the logP interval of the model's training set [-4.34;6]. Secondly, only predictions that falls within the response variable LC₅₀ interval (μM) [0.0004;707945.78] of the model's training set are considered reliable and therefore accepted.

5.2 Method used to assess the applicability domain

Leadscope does not generate predictions for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated.

Only compounds with a logP value within the logP interval [-4.34;6] are within the applicability domain. The generated predictions should fall within the response variable interval [0.0004;707945.78] of the training set. Any prediction outside this interval is set to the closest response variable limit (0.0004 or 707945).

5.3 Software name and version for applicability domain assessment

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

5.4 Limits of applicability

The Danish QSAR group applies an overall definition of structures acceptable for QSAR processing which is applicable for all the in-house QSAR software, i.e. not only Leadscope. According to this definition accepted structures are organic substances with an unambiguous structure, i.e. so-called discrete organics defined

as: organic compounds with a defined two dimensional (2D) structure containing at least two carbon atoms, only certain atoms (H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I), and not mixtures with two or more 'big components' when analyzed for ionic bonds (for a number of small known organic ions assumed not to affect toxicity the 'parent molecule' is accepted). Calculation 2D structures (SMILES and/or SDF) are generated by stripping off ions (of the accepted list given above). Thus, all the training set and prediction set chemicals are used in their non-ionized form. See 5.1 for further applicability domain definition.

6. Internal validation

6.1 Availability of the training set

Yes

6.2 Available information for the training set

CAS

SMILES

6.3 Data for each descriptor variable for the training set

No

6.4 Data for the dependent variable for the training set

All

6.5 Other information about the training set

364 compounds are in the training set.

6.6 Pre-processing of data before modelling

The training set LC₅₀ (96h) results were given in mg/L and were converted to $-\log(\mu\text{M})$ before modelling.

Only structures acceptable for Leadscope were used in the final training set. That is only discrete organic chemicals as described in 5.4 were used. In case of replicate structures, one of the replicates was kept if all the compounds had the same activity and all were removed if they had different activity. No further structures accepted by the software were eliminated (i.e. outliers).

6.7 Statistics for goodness-of-fit

Leadscope's own internal performance test gave the following result for predictions within the applicability domain as defined by Leadscope (i.e. the first criterion described in 5.1):

R-squared: 0.7915

6.8 Robustness – Statistics obtained by leave-one-out cross-validation

Not performed. (It is not a preferred measurement for evaluating large models).

6.9 Robustness – Statistics obtained by leave-many-out cross-validation

Leadscope's own internal leave-many-out (LMO) cross-validation procedure was used for predictions within the applicability domain as defined by Leadscope (i.e. the first criterion described in 5.1). A 10 times 50% cross-validation was done and gave the following result:

R-Square: 0.7138

6.10 Robustness - Statistics obtained by Y-scrambling

Not performed.

6.11 Robustness - Statistics obtained by bootstrap

Not performed.

6.12 Robustness - Statistics obtained by other methods

Not performed.

7. External validation

7.1 Availability of the external validation set

7.2 Available information for the external validation set

7.3 Data for each descriptor variable for the external validation set

7.4 Data for the dependent variable for the external validation set

7.5 Other information about the training set

7.6 Experimental design of test set

7.7 Predictivity – Statistics obtained by external validation

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

External validation has not been performed for this model.

8. Mechanistic interpretation

8.1 Mechanistic basis of the model

The global model identifies structural features and molecular descriptors which in the model development was found to be statistically significant associated with effect. Many predictions may indicate modes of action that are obvious for persons with expert knowledge for the endpoint.

8.2 A priori or posteriori mechanistic interpretation

A posteriori mechanistic interpretation. The identified structural features and molecular descriptors may provide basis for mechanistic interpretation.

8.3 Other information about the mechanistic interpretation

9. Miscellaneous information

9.1 Comments

The model can be used to predict if a chemical is acute toxic (96h) to *Fathead minnow* (fish). The Danish QSAR Group applies an algorithm on top of the predictions from the model in order to convert the values from $-\log(\mu\text{M})$ to mg/L, which is the normal unit for this endpoint.

9.2 Bibliography

MED-Duluth: Geiger, D.L., Call, D.J., and Brooke, L.T. (1988) Acute Toxicities of Organic Chemicals to Fathead Minnows (*Pimephales promelas*). Vol. 1-4, Center for Lake Superior Environmental Studies, University of Wisconsin-Superior, USA. The database is available online at http://www.epa.gov/med/Prods_Pubs/fathead_minnow.htm

OECD guideline 203 (1992) OECD Guidelines for the Testing of Chemicals No. 203: Fish, Acute Toxicity Test. Organisation for Economic Cooperation and Development; Paris, France. Available online at: http://www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en

Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E., and Drummond, R.A. (1997) Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). *Environmental Toxicology and Chemistry*, 16:5, 948-967.

9.3 Supporting information