



# QsarDB – first 100 DOIs for predictive models

Uko Maran

Institute of chemistry, University of Tartu, Estonia

LOD: ★★★★★

# Content

---

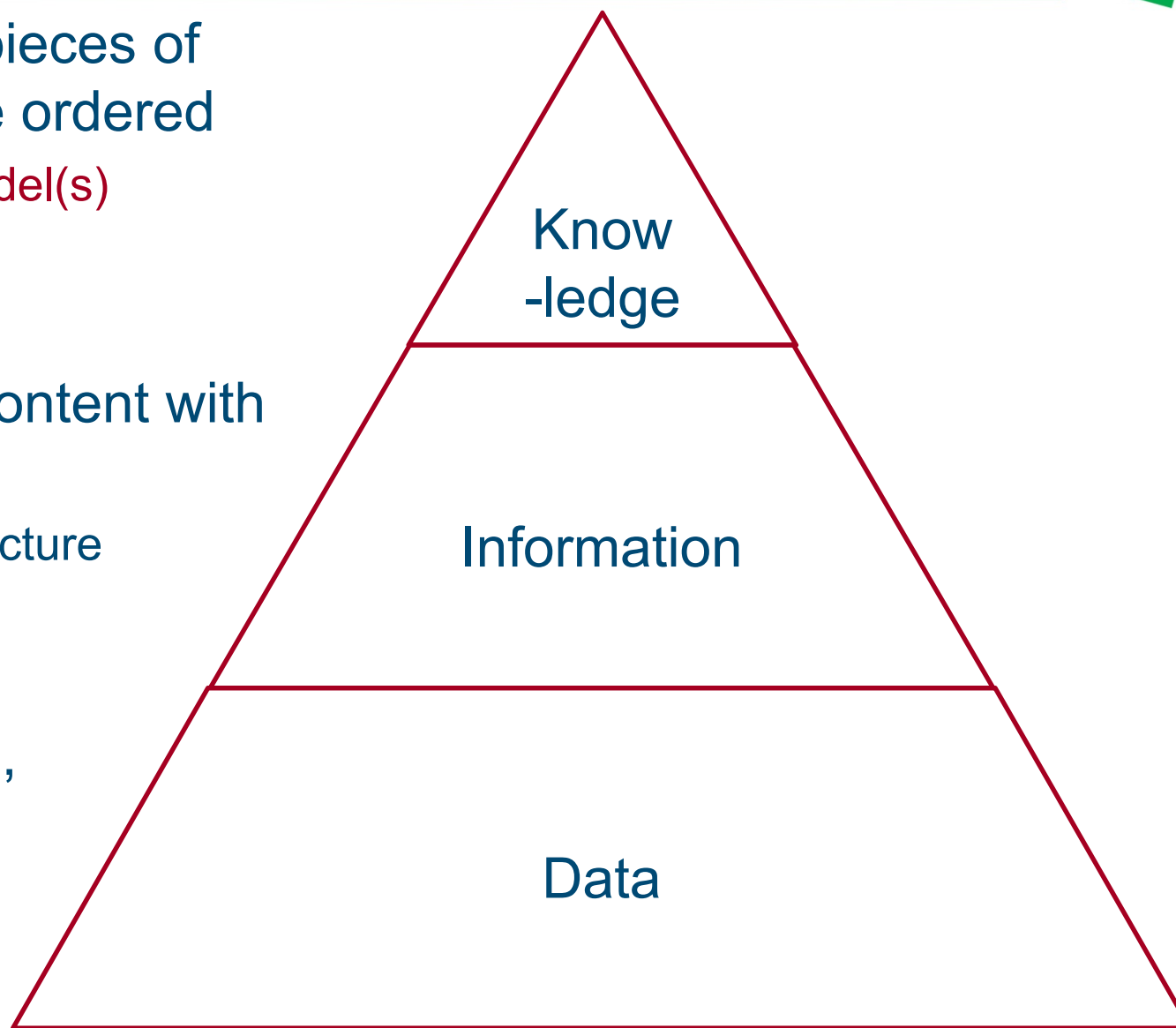


- Data
- Predictive (and descriptive) models?
- Goal
- Components
- Persistent digital identifiers
- First 100+ DOIs for predictive models

# From **data** through **information** to **knowledge**



- Abstraction – pieces of information are ordered
  - In form of model(s)
- Data put into content with other data ...
  - molecular structure
  - annotations
- Measurements, calculations

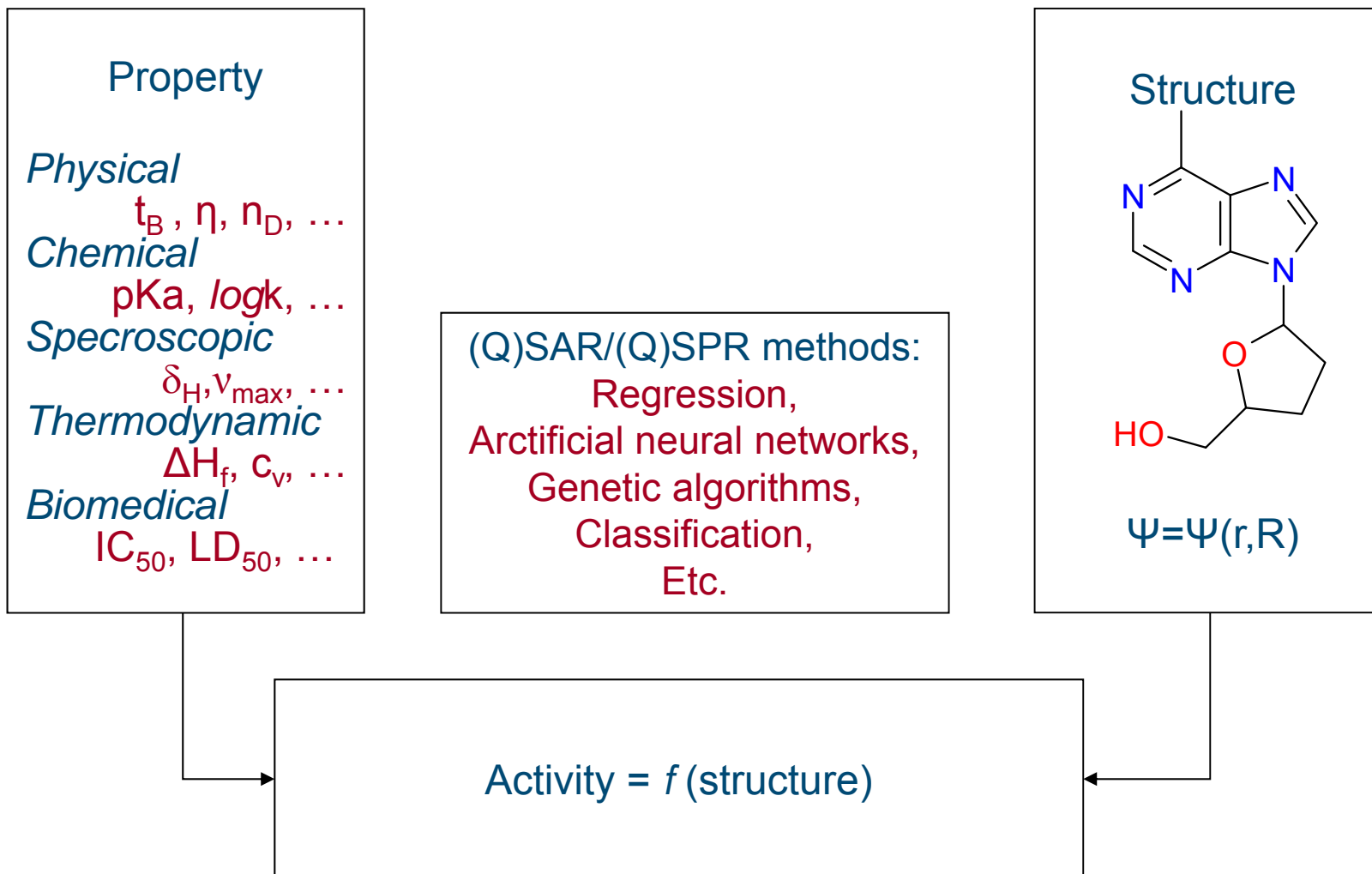


Gasteiger & Engel, Chemoinformatics, Wiley 2003

# Predictive and descriptive models?



## (Quantitative) Structure-Activity Relationships – (Q)SAR





**QsarDB** extends  
the value of (published) predictive *in silico* models  
in chemistry and  
related areas of biomedicine, biotechnology,  
predictive toxicology, etc.  
*via* improving accessibility, transparency and reproducibility

# What for QsarDB is designed?



- To preserve QSAR information
- To make QSAR information transparent
- To make QSAR information reproducible
- To make QSAR information accessible
- To make QSAR information easily transferable
- To adapt to the change in the structure of QSAR community ... **more model users than developers**

# Communication of *in silico* models



- Dominating approach ... *printed media*
- The *main advantage* is
  - peer review process for the independent evaluation of the scientific work and
  - established distribution channels to reach the intended audience.
- The *disadvantage* is ... static nature of printed media ...
  - **accessibility, traceability and reproducibility**

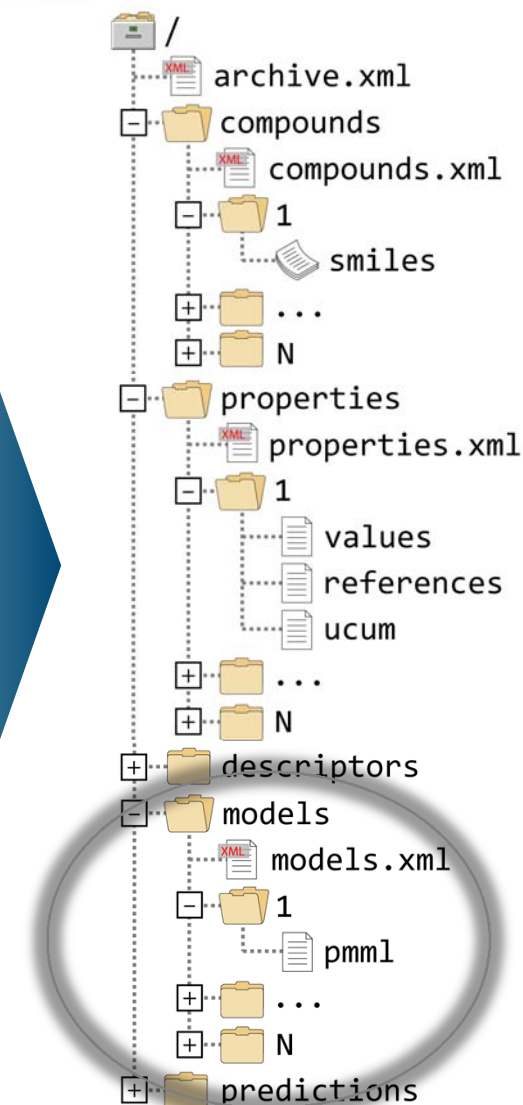
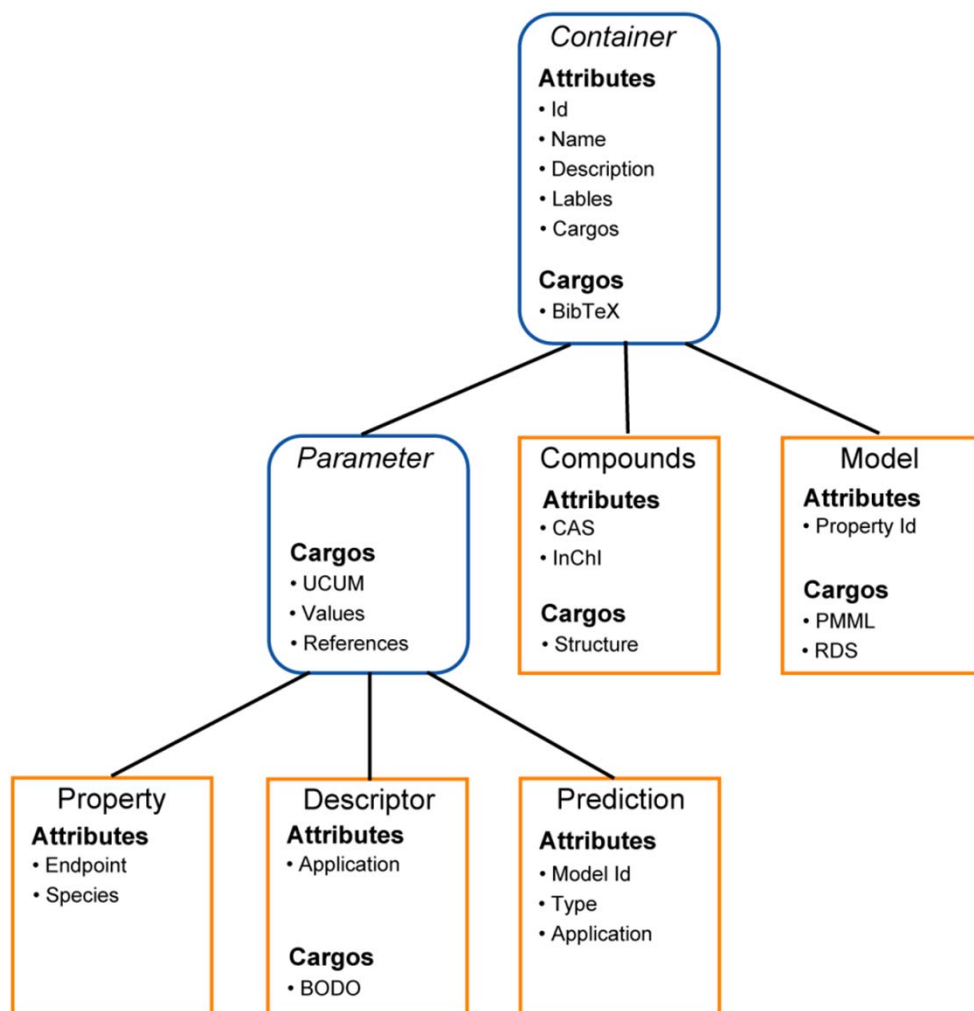
# QsarDB has three major components



- **Data format**
  - QSAR model archive (ie. small database - QDB)
- **Smart Repository**
  - collection of archives
- **Tools for QDB archive creation**
  - Command line
  - Graphical user interface



# QsarDB data schema & format



Villu Ruusmann, Sulev Sild, Uko Maran\*,  
QSAR DataBank - an approach for the digital organization and archiving of QSAR model information.  
*Journal of Cheminformatics*, 2014, 6:25. <http://dx.doi.org/10.1186/1758-2946-6-25>

# Electronic representation of predictive models



- **PMML**
  - Open standard for representing data mining models in XML format
  - PMML covers the following topics
    - Data preprocessing described through data dictionary, mining schema, transformations
    - Model representation
    - Post-processing (e.g. scaling model outputs)
- Other options are possible:
  - For example **RDS** data format (R native model representation mechanism)
- **PMML** format supports:
  - Association rules
  - Cluster models
  - Neural network
  - Regression
  - Random forest
  - Tree models
  - Support-vector machines
  - Ensemble models (all of the above)

# Smart repository: [www.qsardb.org](http://www.qsardb.org)



- DSpace platform ([www.dspace.org](http://www.dspace.org))
- Introduced QSAR specific metadata
- QsarDB archive submission process
- User interface of the web application
  - Item view to display information about QsarDB archive content
  - Explorer tool
  - Prediction tool
- Web service for predictions



# Repository: Model uploading policy



- Must have scientific publication
- (or have otherwise practical value ...)

A screenshot of the QsarDB Home website. The browser address bar shows "qsar.db.org/repository/". The page features a search bar, a "Browse" section with filters like "Communities &amp; Collections", "By Submit Date", "Authors", "Titles", "Endpoints", "Species", "Descriptor calculation software", "Modeling software", and "Model type". There is also a "My Account" section with "Login" and "Register" buttons, and "RSS Feeds" for "RSS 1.0", "RSS 2.0", and "Atom". The main content area includes sections for "Open digital repository", "Explorer tools", "Predictor tools", and "List of communities". A "Recently Added" section lists several publications with their titles, authors, and dates. A circular diagram on the right side of the page illustrates the QsarDB process: "Experimentation" (orange arrow), "Modeling" (green arrow), and "Publishing" (blue arrow), all centered around the "QsarDB" logo.

# Repository: Persistent digital identifiers



- Handle service: <http://hdl.handle.net/10967/106>
- DOI support available starting from August 21-st

<http://hdl.handle.net/10967/106> → <http://qsardb.org/repository/handle/10967/106>

Chemosphere 96 (2014) 23–32

Contents lists available at SciVerse ScienceDirect



Chemosphere

journal homepage: [www.elsevier.com/locate/chemosphere](http://www.elsevier.com/locate/chemosphere)

## Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga *Pseudokirchneriella subcapitata*

Villem Aruoja<sup>a,\*</sup>, Maikki Moosus<sup>b</sup>, Anne Kahru<sup>a</sup>, Mariliis Sihtmäe<sup>a</sup>, Uko Maran<sup>b,\*</sup>

<sup>a</sup>Laboratory of Environmental Toxicology, National Institute of Chemical Physics and Biophysics, Akadeemia tee 23, Tallinn 12618, Estonia  
<sup>b</sup>Institute of Chemistry, University of Tartu, Ravila 14a, Tartu 50411, Estonia

### HIGHLIGHTS

- REACH-relevant algal toxicity data were obtained for 50 nonpolar narcotic chemicals.
- Most of the tested compounds so far lacked published algal growth inhibition values.
- Toxicity of non-polar narcotic compounds correlated with hydrophobicity:  $R^2 = 0.95$ .
- MLR QSAR model was derived for non-polar and polar narcotic compounds:  $R^2 = 0.92$ .
- The Verhaar classification of non-polar narcotics appears to apply for algae.

### ARTICLE INFO

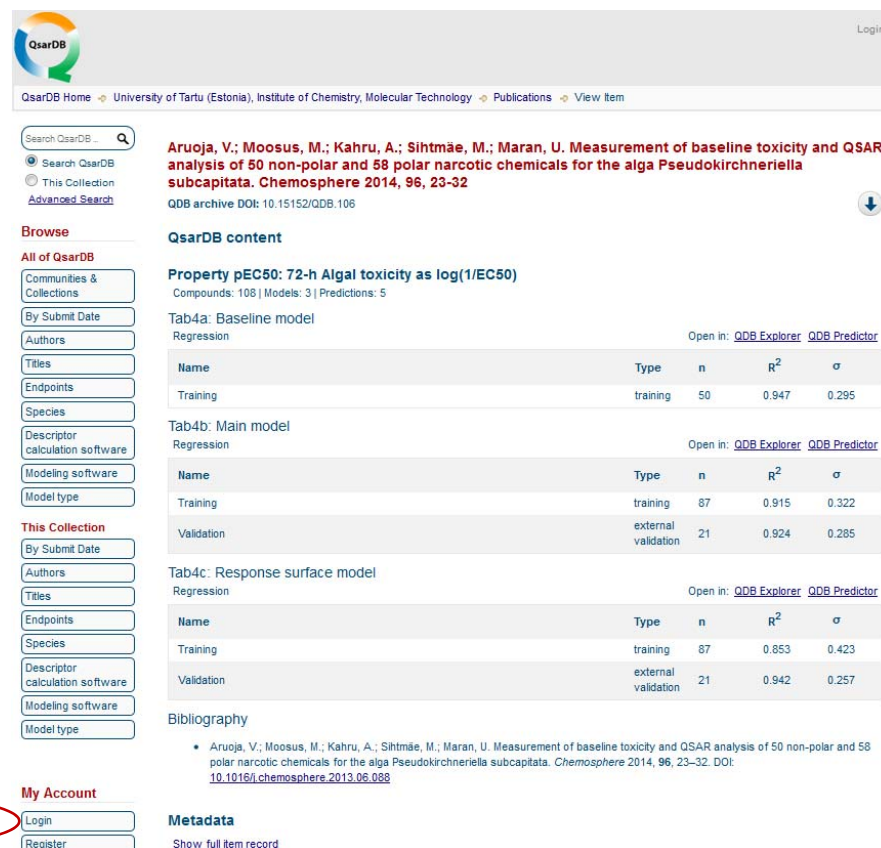
Article history:  
Received 20 March 2013  
Received in revised form 28 June 2013  
Accepted 30 June 2013  
Available online 26 July 2013

### Keywords:

REACH  
Baseline toxicity  
QSAR  
Non-polar narcosis  
Algae  
*Pseudokirchneriella subcapitata*

### ABSTRACT

In this paper a set of homogenous experimental algal toxicity data was measured for 50 non-polar narcotic chemicals using the alga *Pseudokirchneriella subcapitata* in a closed test with a growth rate endpoint. Most of the tested compounds are high volume industrial chemicals that so far lacked published REACH-compliant algal growth inhibition values. The test protocol fulfilled the criteria set forth in the OECD guideline 201 and had the same sensitivity as the open test which allowed direct comparison of toxicity values. Baseline QSAR model for non-polar narcotic compounds was established and compared with previous analogous models. Multi-linear QSAR model was derived for the non-polar and 58 previously tested polar (anilines and phenols) narcotic compounds modulating hydrophobicity, molecular size, electronic and molecular stability effects coded in the molecular descriptors. Descriptors in the model were analyzed and applicability domain was assessed providing further guidelines for the *in-silico* prediction purposes in decision support while performing risk assessment. QSAR models in the manuscript are available online through QsarDB repository for exploring and prediction services (<http://hdl.handle.net/10967/106>).  
© 2013 Elsevier Ltd. All rights reserved.



QsarDB Home University of Tartu (Estonia), Institute of Chemistry, Molecular Technology Publications View Item

Search QsarDB

Search QsarDB  
This Collection  
Advanced Search

**Browse**

All of QsarDB

Communities & Collections  
By Submit Date  
Authors  
Titles  
Endpoints  
Species  
Descriptor calculation software  
Modeling software  
Model type

**This Collection**

By Submit Date  
Authors  
Titles  
Endpoints  
Species  
Descriptor calculation software  
Modeling software  
Model type

**My Account**

Login  
Register

Aruoja, V.; Moosus, M.; Kahru, A.; Sihtmäe, M.; Maran, U. Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga *Pseudokirchneriella subcapitata*. *Chemosphere* 2014, 96, 23–32. DOI: 10.1016/j.chemosphere.2013.06.088

**QsarDB content**

**Property pEC50: 72-h Algal toxicity as log(1/EC50)**  
Compounds: 108 | Models: 3 | Predictions: 5

Tab4a: Baseline model  
Regression  
Open in: [QDB Explorer](#) [QDB Predictor](#)

Name	Type	n	R <sup>2</sup>	σ
Training	training	50	0.947	0.295

Tab4b: Main model  
Regression  
Open in: [QDB Explorer](#) [QDB Predictor](#)

Name	Type	n	R <sup>2</sup>	σ
Training	training	87	0.915	0.322
Validation	external validation	21	0.924	0.285

Tab4c: Response surface model  
Regression  
Open in: [QDB Explorer](#) [QDB Predictor](#)

Name	Type	n	R <sup>2</sup>	σ
Training	training	87	0.853	0.423
Validation	external validation	21	0.942	0.257

**Bibliography**

- Aruoja, V.; Moosus, M.; Kahru, A.; Sihtmäe, M.; Maran, U. Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga *Pseudokirchneriella subcapitata*. *Chemosphere* 2014, 96, 23–32. DOI: 10.1016/j.chemosphere.2013.06.088

**Metadata**  
[Show full item record](#)

# Repository: Explorer – visualization



- Visualizes
  - property data;
  - residuals;
  - descriptors;
  - applicability domain;

## Property analysis

▶ 40-h Tetrahymena toxicity as log(1/IC50)

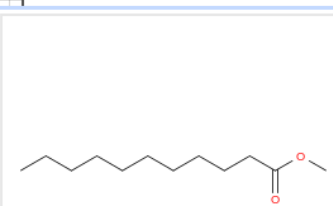
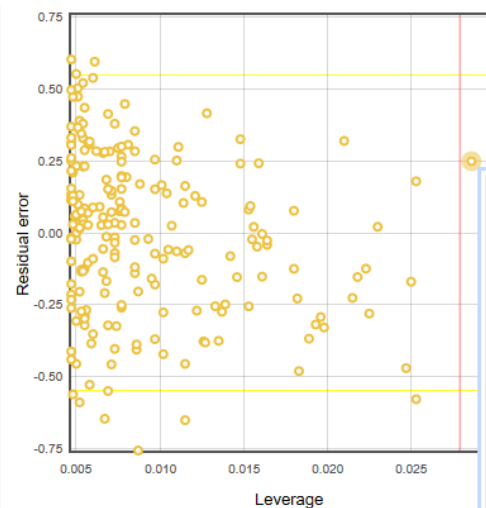
▶ Residual error

## Descriptor analysis

▶ Hydrophobicity

## Applicability domain analysis

### ▼ Williams plot



<b>Id</b>	133
<b>Name</b>	<u>methyl undecanoate</u>
<b>CAS</b>	1731-86-8

▶ Gramatica plot

# Repository: Predictor



- Predict:
  - from structure
  - from descriptors  
(safe for commercial users)

## Model input

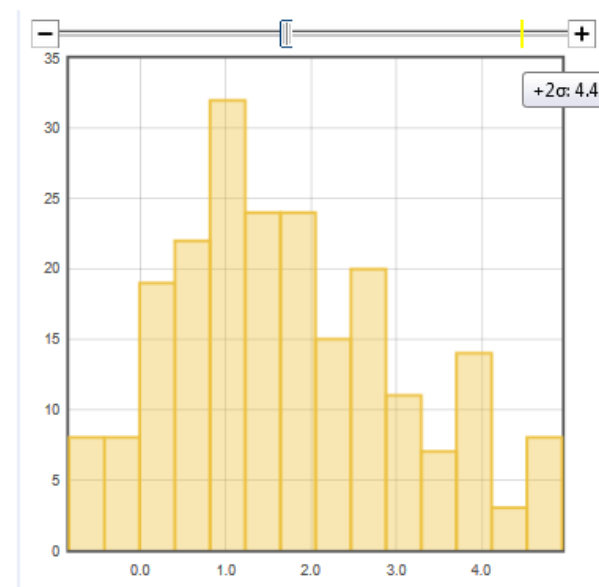
Chemical structure input (InChI or SMILES format):

Calculate

## Descriptor input:

▼ Hydrophobicity

1.73



## Model output

40-h Tetrahymena toxicity as  
 $\log(1/IC50)$

$$= 0.723 * \text{Hydrophobicity} - 1.79$$
$$= 0.723 * 1.73 - 1.79$$
$$= -0.5392$$

# First 100+ DOIs for predictive models



- <http://dx.doi.org/10.15152/QDB.106>
- 108 DOI-s (as of 23.10.2014)
- 244 descriptive & predictive models (23.10.2014)
- QsarDB qualifies in all five Linked Open Data concept criteria:
  - OL ... Open License (★)
  - +RE ... machine REadable (★★)
  - +OF ... Open Format (★★★)
  - +URI ... Uniform Resource Identifier (★★★★)
  - +LD ... Linked Data (★★★★★)

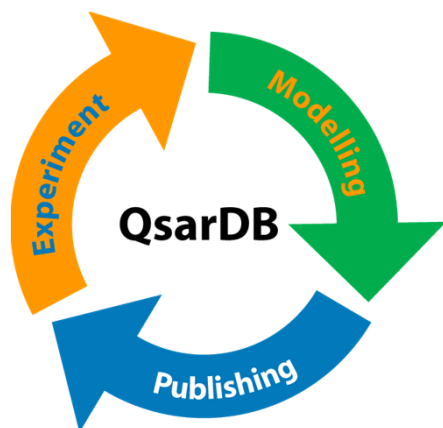


# Concluding Phrases



- One of our aims and reason for QsarDB is to
  - help modelers to keep their published (static PDF) models alive ... (for example <http://dx.doi.org/10.1016/j.chemosphere.2013.06.088>)
- Making Smart Repository even smarter:
  - structure & similarity search, etc.
  - new model types are coming in ...
  - access for the scientific communities ...
  - ...
- Preparing for interactive scientific publications!

# Thank you!



Investing in your future



European Union  
Regional Development Fund

[www.qsardb.org](http://www.qsardb.org)

(project # 3.2.1201.13-0021)