CADASTER

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Grant agreement no.: 212668

Collaborative Project

Sub-Priority ENV2007 3.3.1.1: In-silico techniques for hazard-, safety-, and environmental risk-assessment

Establishment of a database on properties and fate/effect parameters of chemicals within the four classes of chemicals selected (Deliverable 2.4)

Start date of project: 1 January 2009 Duration: 4 years

Due date of deliverable: 31 December, 2011
Actual submission date: 22 December, 2011

Lead Contractor: National Institute of Public Health and the Environment (RIVM), Laboratory for Ecological Risk Assessment
Corresponding authors of the document: Igor V. Tetko[1], Mojca Kos Durjava[2]

1. Helmholtz Zentrum Muenchen - German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany
2. Public Health Institute Maribor, Center for Risk Assessment of Chemicals with Laboratory, Maribor, Prvomajska 1, Slovenija

Reviewed by: Paola Gramatica -- QSAR Research Unit in Environmental Chemistry and Ecotoxicology, University of Insubria, Via J.H. Dunant 3 - 21100 Varese, Italy

Deliverable no: 2.4 (Database of properties)

Nature: Report publicly available + prototype

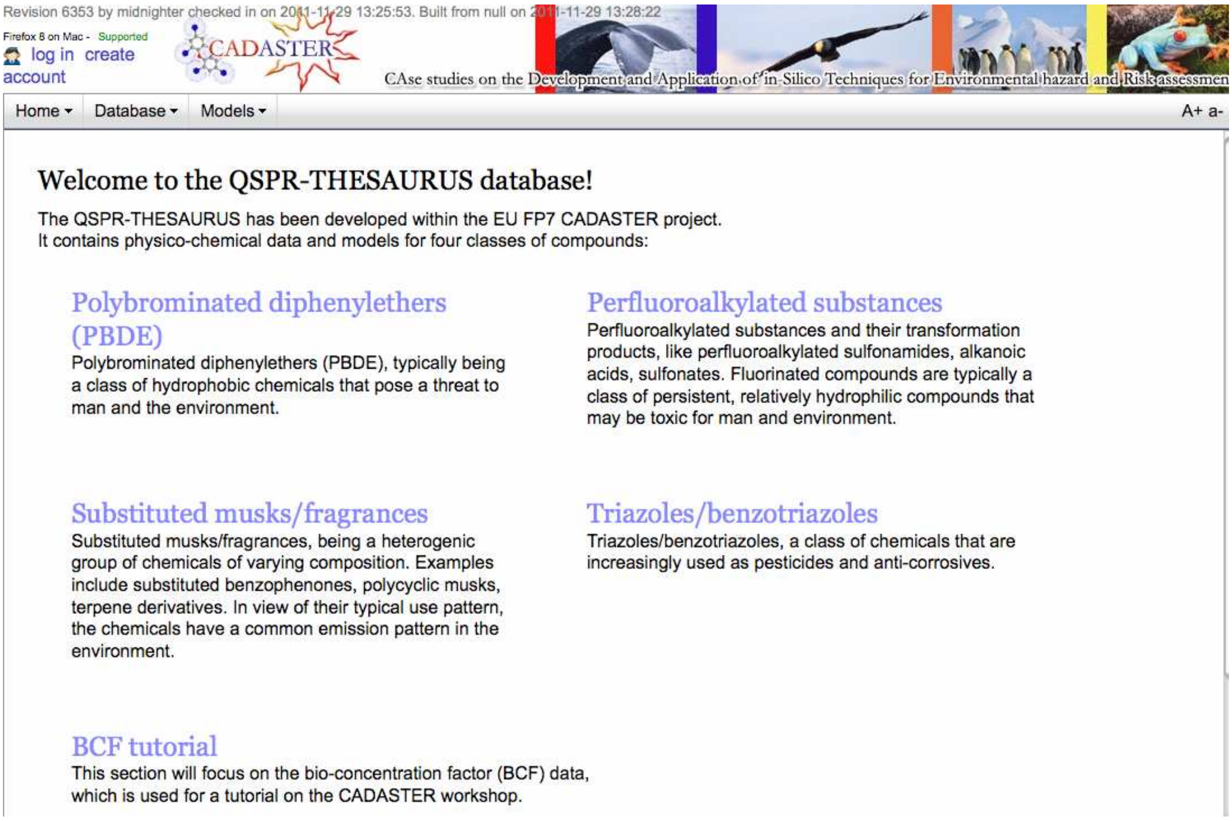
Project co-funded by the EU Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	X
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of content

WP2: ***Database on experimental parameters and (Q)SARs for chemical and biological endpoints***

Work Package Leader: Mojca Kos Durjava (Partner 2: Public Health Institute Maribor)

The CADASTER QSPR-THESAURUS database has been developed within the framework of the CADASTER project. The database is based on the On-line Chemical Modeling Environment (QSPR THESAURUS) http://www.qspr-thesaurus.eu, which has been developed by Dr Tetko's group at HMGU[1] and is currently being offered as a commercial software from eADMET GmbH http://www.eadmet.com. The database was further developed according to the request of the CADASTER project partners and database users. The database provides the main repository to store and handle endpoint data collected and measured during the CADASTER project. Below, we describe the main features of the database.


Revision 6353 by midnighter checked in on 2011-11-29 13:25:53. Built from null on 2011-11-29 13:28:22
Firefox 8 on Mac - Supported
log in create account
CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessmen
Home
Database
Models
A+ a-

Welcome to the QSPR-THESAURUS database!

The QSPR-THESAURUS has been developed within the EU FP7 CADASTER project.
It contains physico-chemical data and models for four classes of compounds:

Polybrominated diphenylethers (PBDE)

Polybrominated diphenylethers (PBDE), typically being a class of hydrophobic chemicals that pose a threat to man and the environment.

Perfluoroalkylated substances

Perfluoroalkylated substances and their transformation products, like perfluoroalkylated sulfonamides, alkanoic acids, sulfonates. Fluorinated compounds are typically a class of persistent, relatively hydrophilic compounds that may be toxic for man and environment.

Substituted musks/fragrances

Substituted musks/fragrances, being a heterogenic group of chemicals of varying composition. Examples include substituted benzophenones, polycyclic musks, terpene derivatives. In view of their typical use pattern, the chemicals have a common emission pattern in the environment.

Triazoles/benzotriazoles

Triazoles/benzotriazoles, a class of chemicals that are increasingly used as pesticides and anti-corrosives.

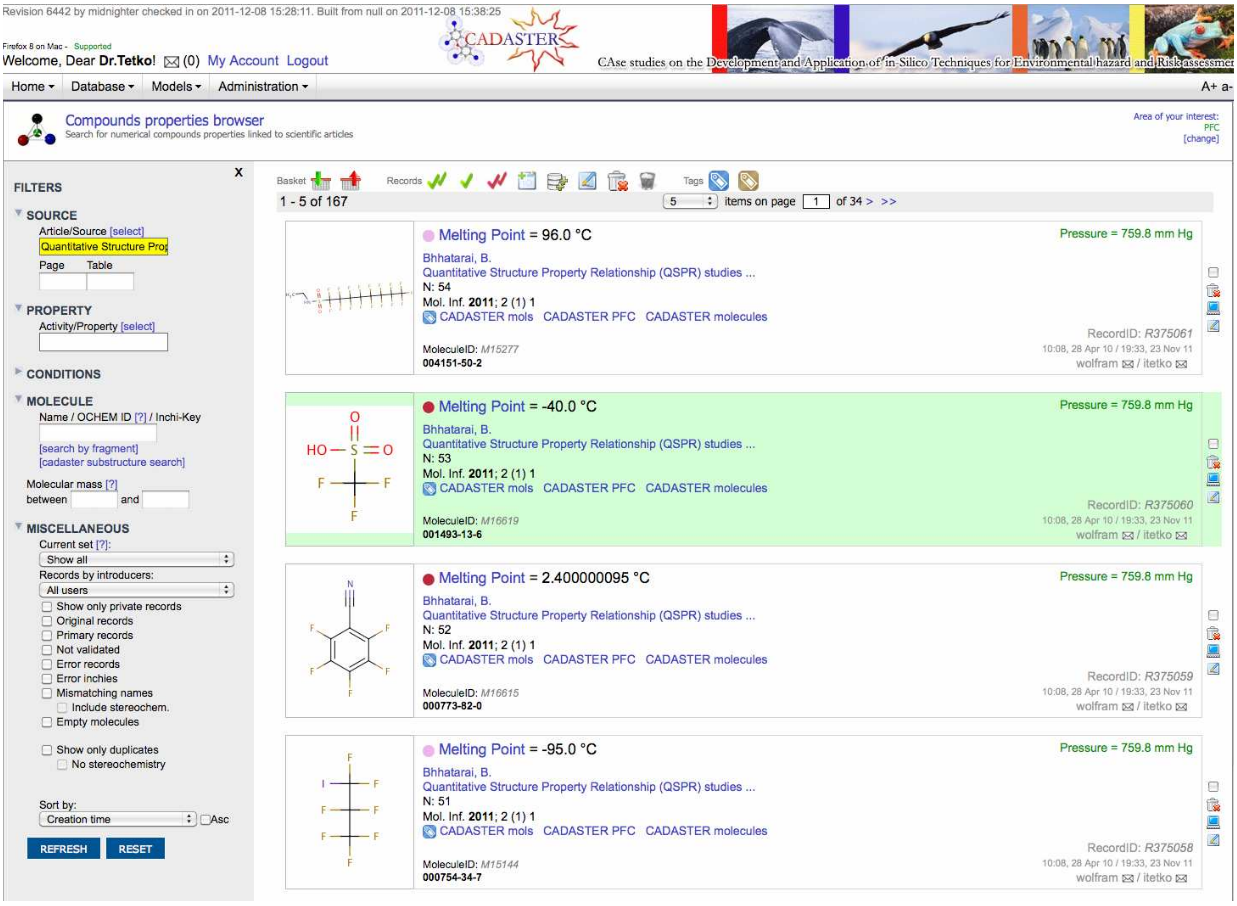
BCF tutorial

This section will focus on the bio-concentration factor (BCF) data, which is used for a tutorial on the CADASTER workshop.

Figure 1. Front page of the QSPR THESAURUS database.

Overview of the database structure

The front page of the QSPR THESAURUS database provides an access to the 4 classes of chemicals which are the focus of the CADASTER project (Figure 1). After selection of one of the classes, the user accesses the database of experimental and calculated properties. The database (Figure 2) contains experimentally measured biological and physicochemical properties of molecules belonging to the four classes, together with the conditions under which the experiments have been conducted and references to the sources where the data were published. These data were collected or measured by CADASTER partners during the project.


CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessmen
Home ▾
Database ▾
Models ▾
Administration ▾
A+ a-
Basket
Records
Tags
REFRESH
RESET
X

Revision 6442 by midnighter checked in on 2011-12-08 15:28:11. Built from null on 2011-12-08 15:38:25

Firefox 8 on Mac - Supported

Welcome, Dear **Dr.Tetko**! (0) My Account Logout

Compounds properties browser

Search for numerical compounds properties linked to scientific articles

Area of your interest:
PFC
[change]

FILTERS

SOURCE

Article/Source [select]: Quantitative Structure Proj

Page Table

PROPERTY

Activity/Property [select]

CONDITIONS

MOLECULE

Name / OCHEM ID [?] / Inchi-Key

[search by fragment]
[cadaster substructure search]

Molecular mass [?] between and

MISCELLANEOUS

Current set [?]: Show all

Records by introducers: All users

- □ Show only private records
- □ Original records
- □ Primary records
- □ Not validated
- □ Error records
- □ Error inchies
- □ Mismatching names
  - □ Include stereochem.
- □ Empty molecules
- □ Show only duplicates
  - □ No stereochemistry

Sort by: Creation time □Asc

1 - 5 of 167 — 5 items on page 1 of 34 > >>

● Melting Point = 96.0 °C — Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 54
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M15277*
004151-50-2
RecordID: *R375061*
10:08, 28 Apr 10 / 19:33, 23 Nov 11
wolfram / itetko

● Melting Point = -40.0 °C — Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 53
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M16619*
001493-13-6
RecordID: *R375060*
10:08, 28 Apr 10 / 19:33, 23 Nov 11
wolfram / itetko

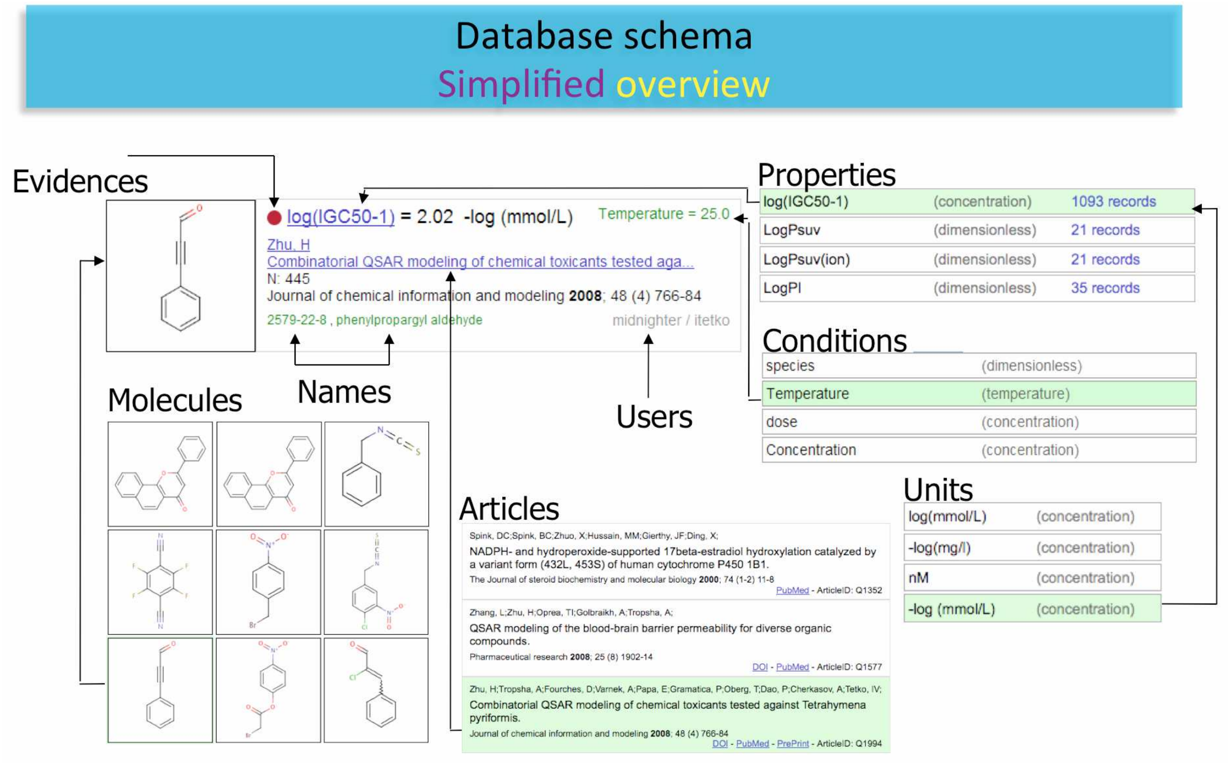
● Melting Point = 2.400000095 °C — Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 52
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M16615*
000773-82-0
RecordID: *R375059*
10:08, 28 Apr 10 / 19:33, 23 Nov 11
wolfram / itetko

● Melting Point = -95.0 °C — Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 51
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M15144*
000754-34-7
RecordID: *R375058*
10:08, 28 Apr 10 / 19:33, 23 Nov 11
wolfram / itetko

Figure 2. Database of experimental properties and fate effects available at http://www.qspr-thesaurus.eu.

The *experimental measurements* are the central entities of the database. They combine all the information related to the experiment, in particular the result of the measurement, which can be either numeric or qualitative depending on the measured property. The central system component, where the experimental measurements can be introduced, searched and manipulated, is the *compound property browser*.

The *experimental measurement* includes information about the *property* that was measured and the associated *chemical compound*. The compounds and the properties can be marked with particular keywords, also known as *tags*, that allow convenient filtering and grouping of the data. The CADASTER database uses five tags to differentiate compounds from each of four analyzed classes as well as all compounds.



● log(IGC50-1) = 2.02 -log (mmol/L) Temperature = 25.0
Zhu, H
Combinatorial QSAR modeling of chemical toxicants tested aga...
N: 445
Journal of chemical information and modeling **2008**; 48 (4) 766-84
2579-22-8 , phenylpropargyl aldehyde
midnighter / itetko

log(IGC50-1)	(concentration)	1093 records
LogPsuv	(dimensionless)	21 records
LogPsuv(ion)	(dimensionless)	21 records
LogPI	(dimensionless)	35 records

species	(dimensionless)
Temperature	(temperature)
dose	(concentration)
Concentration	(concentration)

log(mmol/L)	(concentration)
-log(mg/l)	(concentration)
nM	(concentration)
-log (mmol/L)	(concentration)

Spink, DC;Spink, BC;Zhuo, X;Hussain, MM;Gierthy, JF;Ding, X;
NADPH- and hydroperoxide-supported 17beta-estradiol hydroxylation catalyzed by a variant form (432L, 453S) of human cytochrome P450 1B1.
The Journal of steroid biochemistry and molecular biology **2000**; 74 (1-2) 11-8
PubMed - ArticleID: Q1352

Zhang, L;Zhu, H;Oprea, TI;Golbraikh, A;Tropsha, A;
QSAR modeling of the blood-brain barrier permeability for diverse organic compounds.
Pharmaceutical research **2008**; 25 (8) 1902-14
DOI - PubMed - ArticleID: Q1577

Zhu, H;Tropsha, A;Fourches, D;Varnek, A;Papa, E;Gramatica, P;Oberg, T;Dao, P;Cherkasov, A;Tetko, IV;
Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.
Journal of chemical information and modeling **2008**; 48 (4) 766-84
DOI - PubMed - PrePrint - ArticleID: Q1994

Figure 3. Rich annotation of endpoint data in the database.

For each measurement stored in QSPR THESAURUS, it is obligatory to specify the *source of the data* (Figure 3). The source is usually a publication in a scientific journal or a book. The strict policy of QSPR THESAURUS is to accept only those experimental records that have their source of information specified. This improves the quality of the data and allows it to be verified by checking the original publication. Although a user can also introduce an unpublished article and link the data to it, records from such sources should be treated with caution. The ways to browse, introduce and automatically fetch the publications from the PubMed database are described below in the “Sources of information” section.

Every numeric property has a corresponding category of *units*, for example, the category of units for *Inhibition Concentration 50%* (IC50) is “Concentration”. By default the QSPR THESAURUS database keeps experimental endpoints in the original format (i.e., in *units* as reported in the publication). For this purpose all units are grouped into strictly defined unit categories, for example Kelvin, Celsius and Fahrenheit degrees belong to the “Temperature” category. For the purpose of compatibility and for modeling of the combined sets from different publications, the system provides on the fly conversion between different units.

An important feature of our database, which is also unique among other chemical databases, is the possibility to store the *conditions of experiments*. This information is crucial for modeling: in many cases, the result of an experimental measurement is senseless without knowing the conditions under which the experiment has been conducted. For example, it does not make sense to specify the boiling point for a compound without specifying the air pressure. Such conditions should be introduced as *obligatory* conditions, i.e., a new record will be rejected by the system if there is no information about these conditions provided. Conditional values stored in the database can be numerical (with units of measurement), qualitative or descriptive (textual). Moreover, in the “conditions” section it is possible to note additional information related to the experiment, even if it is not a “condition” in the classical sense. Examples of such additional information are assay descriptions, a target of the ligand (the receptor) or species on which the biological activity has been tested. For simplicity, we further universally refer to all this information as “conditions”.

Features overview

In brief, the distinguishing features of the QSPR THESAURUS database are as follows.

- The wiki principle: most of the data can be accessed, introduced and modified by users
- Different access levels: guests, registered users, verified users, administrators
- Tracking of all the changes
- Obligatory indications of the source of the data
- Possibility to indicate conditions of the experiment, which can be later used for QSAR modeling
- Search by substructure, molecule names, by publication where the measurements were referenced, by conditions of experiments, etc.
- Control of duplicated records
- Batch upload and batch modification of large amounts of data
- Different units of measurements and utilities to interconvert between units
- Organizing the records in re-usable sets (“baskets”)

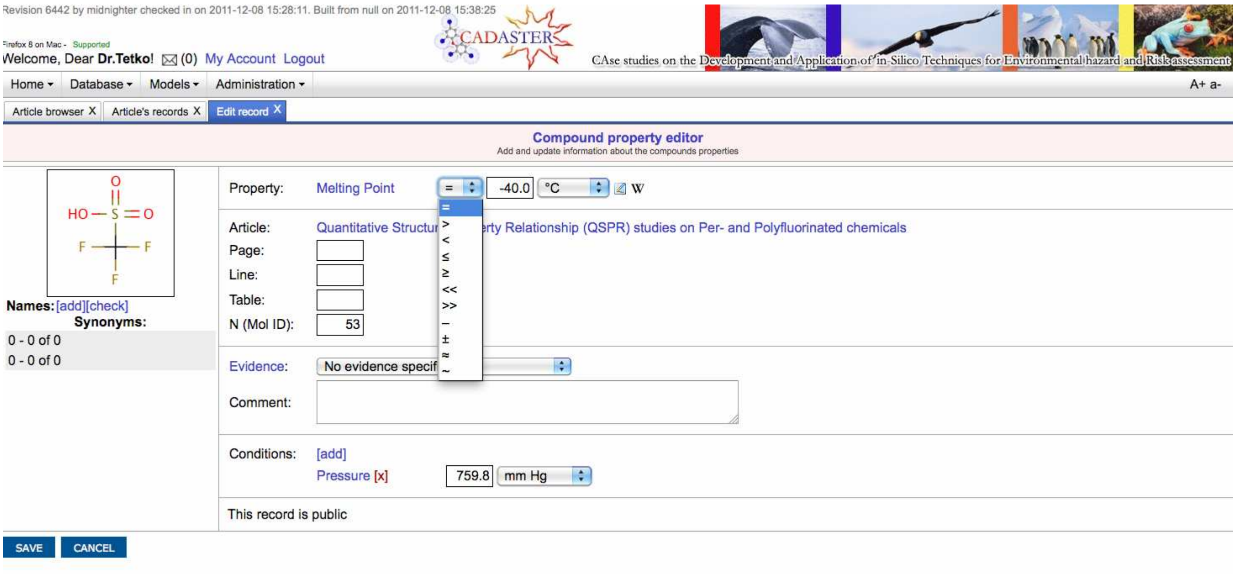
- Hidden and public records to allow collaborative development on the web

Introduction of data

Basics

There are two ways to introduce experimental data to the QSPR THESAURUS database: the first way is the manual record-by-record input, where each experimental measurement is entered separately; the second way is the batch upload facility that allows upload of large amounts of data from Excel or SDF (Structure Data File). This functionality is described in more detail below in the “Batch upload” section. Other types of database entities (new physicochemical properties, units of measurements, publications, etc.) can be introduced via special interface windows called *browsers*. Basically, every entity in the QSPR THESAURUS database has a corresponding browser: the c*ompound properties browser* for experimental measurements, the *publications browser* for articles and books, the *properties browser*, the *units browser* etc. Additionally, for every entity there is a specific dialog window where a user can create a new entry or edit an existing one.

This also applies to experimental measurements: each measurement is created and modified in the *record editor* window (Figure 4), where a user specifies all relevant information: compound structure, corresponding publication, conditions of the experiment, units of measurement etc.


Revision 6442 by midnighter checked in on 2011-12-08 15:28:11. Built from null on 2011-12-08 15:38:25
Firefox 8 on Mac - Supported
Welcome, Dear Dr.Tetko! (0) My Account Logout
CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment
Home
Database
Models
Administration
A+ a-
Article browser X
Article's records X
Edit record X
HO
S
O
O
F
F
F
SAVE
CANCEL
=
>
<
≤
≥
<<
>>
–
±
≈
~

Compound property editor

Add and update information about the compounds properties

Names: [add][check]

Synonyms:

0 - 0 of 0

0 - 0 of 0

Property: Melting Point = -40.0 °C W

Article: Quantitative Structur… …erty Relationship (QSPR) studies on Per- and Polyfluorinated chemicals

Page:

Line:

Table:

N (Mol ID): 53

Evidence: No evidence specif…

Comment:

Conditions: [add]

Pressure [x] 759.8 mm Hg

This record is public

Figure 4. The single record editor is used to introduce a new data point. In addition to exact values, the user can also specify ranges, accuracy, and predictions.

Similarly, there is the *molecule editor* that allows introduction of compound structures in several ways: a user can either explicitly draw the structure in the JME molecule editor, upload an SDF or MOL2 file, specify a SMILES string, or paste a file in one of these formats. The JME editor used at QSPR THESAURUS is probably the most used structure input tool on the internet. The program allows users to draw, edit, and display molecules and reactions as described at http://www.molinspiration.com/jme/. Additionally, the structure of a molecule can be automatically retrieved from the PubChem database by the name of molecule, though this leaves the potential for incorrect retrieval based on incorrect name-structure association and the user should validate the correctness of the structure that is downloaded.

Batch upload

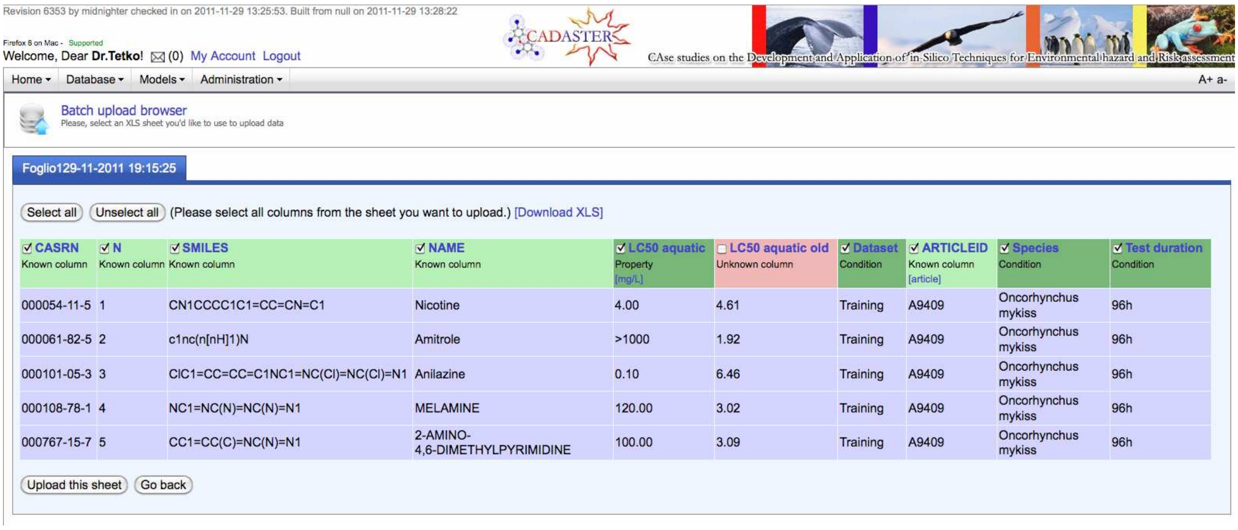
Since QSPR THESAURUS relies on user contributions, it is essential to provide the user with simple and efficient tools to introduce data into the system. The record editor mentioned above is useful for data correction or single record introduction, but is not suitable for the introduction of hundreds or thousands of records. For efficient and fast introduction of large amounts of data, QSPR THESAURUS offers the “batch upload” tool. This section provides an insight into the main features of this tool.

Input data. The input data for the batch upload tool is a specially prepared Excel workbook, CSV or SDF file. The file is processed by the tool to create records according to the provided data. To unify the upload process, SDF and CSV files are internally converted to the Excel file format with tags represented as columns and molecular structures and names put into additional columns. An example Excel file with all possible columns and explanations can be downloaded directly at the first page of the batch upload tool.

As described in the “Structure overview” section, the essential information contained in the record is the value of a biological or chemical property for a specific molecule published in a specific article. Although the Excel file format allows a user to provide all the detailed information about a record (number of a page in an article, where a particular value was published, accuracy of measurements, textual comments, record evidence, measurement units, etc.), a minimal valid file must contain only the information on property value, molecular structure (or name) and article details for every uploaded record. In case some information is not provided (i.e., unit of measurement), the default values specified for the uploaded property (or condition) are used.

Information about a molecular structure can be provided in the form of SMILES, SDF or MOL2. If the structure of the molecule is not available, it is possible to provide a molecule name or identifier, e.g., Chemical Abstract Registration Number (CAS-RN) – if possible, the structure will be retrieved automatically from the PubChem database. The publication can be specified either in the form of an internal QSPR THESAURUS article identifier or a PubMed identifier. The sheet can also contain information about the measurement conditions. The information about the property itself and all the required conditions and units should already be present in the database. For numerical properties, users can provide predicates, such as >, <, ≥, ≤, ~, >>, <<, ≈ as well as the accuracy of the measurements.

After the file has been created, the user can use the batch upload tool to introduce data to QSPR THESAURUS. The tool is created in the form of a “wizard” with a step-by-step approach for the upload process. In the first step, the user must choose the file to upload. The file should be a valid Excel, CSV or SDF file not larger than 25MB. The database also allows storing predicted and hidden data and on the first page a user should indicate whether his/her data are of one of these types.


CADASTER
CAse studies on the Development and Application of in Silico Techniques for Environmental hazard and Risk assessment
Home ▾
Database ▾
Models ▾
Administration ▾
A+ a-
Select all
Unselect all
Upload this sheet
Go back

Revision 6353 by midnighter checked in on 2011-11-29 13:25:53. Built from null on 2011-11-29 13:28:22

Firefox 8 on Mac - Supported

Welcome, Dear Dr.Tetko! ✉ (0) My Account Logout

Batch upload browser

Please, select an XLS sheet you'd like to use to upload data

Foglio129-11-2011 19:15:25

(Please select all columns from the sheet you want to upload.) [Download XLS]

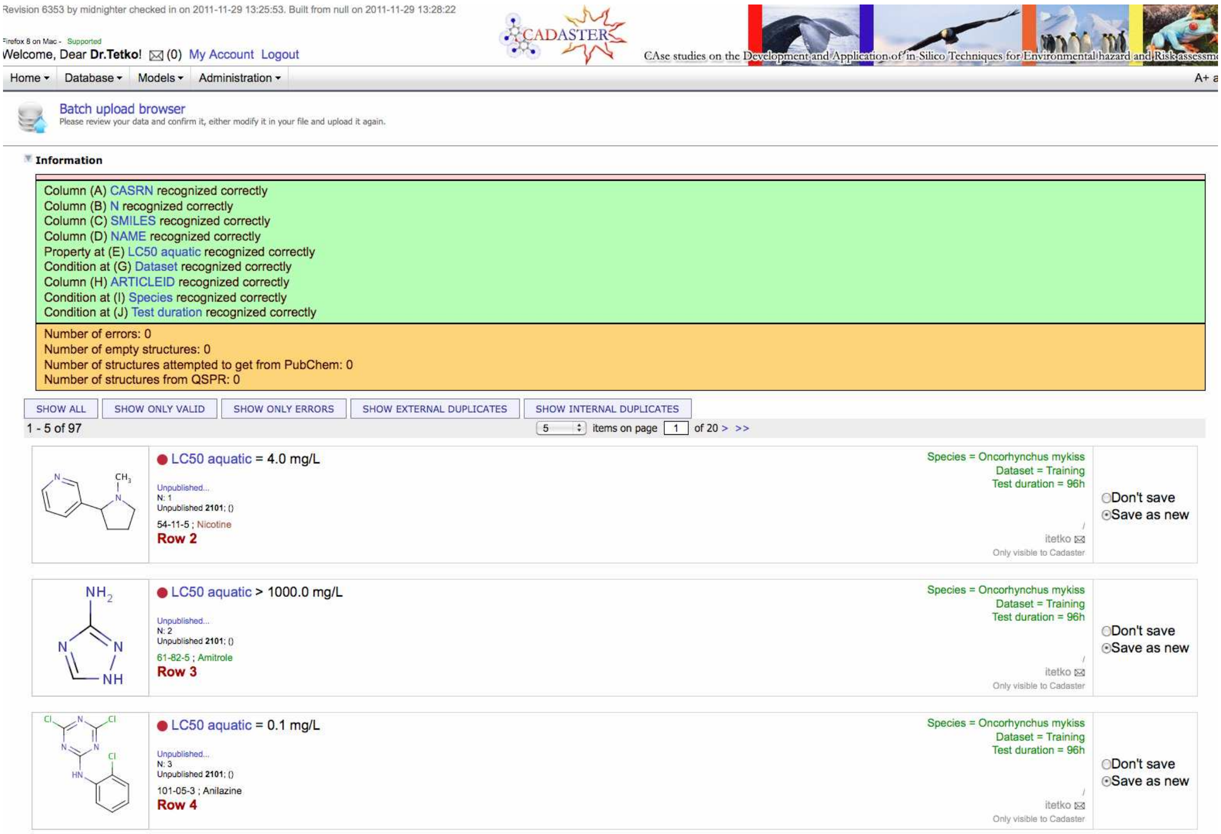
☑ CASRN Known column	☑ N Known column	☑ SMILES Known column	☑ NAME Known column	☑ LC50 aquatic Property [mg/L]	□ LC50 aquatic old Unknown column	☑ Dataset Condition	☑ ARTICLEID Known column [article]	☑ Species Condition	☑ Test duration Condition
000054-11-5	1	CN1CCCC1C1=CC=CN=C1	Nicotine	4.00	4.61	Training	A9409	Oncorhynchus mykiss	96h
000061-82-5	2	c1nc(n[nH]1)N	Amitrole	>1000	1.92	Training	A9409	Oncorhynchus mykiss	96h
000101-05-3	3	ClC1=CC=CC=C1NC1=NC(Cl)=NC(Cl)=N1	Anilazine	0.10	6.46	Training	A9409	Oncorhynchus mykiss	96h
000108-78-1	4	NC1=NC(N)=NC(N)=N1	MELAMINE	120.00	3.02	Training	A9409	Oncorhynchus mykiss	96h
000767-15-7	5	CC1=CC(C)=NC(N)=N1	2-AMINO-4,6-DIMETHYLPYRIMIDINE	100.00	3.09	Training	A9409	Oncorhynchus mykiss	96h

Figure 5. Sheet preview during the data upload. The green headers indicate automatically recognized entities.

Sheet preview. The second step of the batch upload is the data sheet preview (see Figure 5). At this stage the user can choose a particular sheet from the uploaded workbook (only one sheet at a time can be uploaded), deselect unnecessary columns (especially useful in the case of multiple property columns in a sheet since only one property can be uploaded at a time) and override default units for properties and conditions.

At this stage, it is also possible to provide details of the associated article for the records that have no article explicitly specified in the uploaded file. This is often the case for SDF files. In case of mistakes in the property and column names, the user can rename (or “remap”) columns.

Once the required columns are selected, all property units are set to the desired values and column names are remapped, the user can proceed with the upload.



Revision 6353 by midnighter checked in on 2011-11-29 13:25:53. Built from null on 2011-11-29 13:28:22
Firefox 8 on Mac - Supported
Welcome, Dear Dr.Tetko! (0) My Account Logout
CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessm
Home
Database
Models
Administration
A+ a
SHOW ALL
SHOW ONLY VALID
SHOW ONLY ERRORS
SHOW EXTERNAL DUPLICATES
SHOW INTERNAL DUPLICATES
5 items on page 1 of 20 > >>
CH3
N
NH2
NH
Cl
HN

Batch upload browser
Please review your data and confirm it, either modify it in your file and upload it again.

Information

Column (A) CASRN recognized correctly
Column (B) N recognized correctly
Column (C) SMILES recognized correctly
Column (D) NAME recognized correctly
Property at (E) LC50 aquatic recognized correctly
Condition at (G) Dataset recognized correctly
Column (H) ARTICLEID recognized correctly
Condition at (I) Species recognized correctly
Condition at (J) Test duration recognized correctly

Number of errors: 0
Number of empty structures: 0
Number of structures attempted to get from PubChem: 0
Number of structures from QSPR: 0

1 - 5 of 97

● LC50 aquatic = 4.0 mg/L
Unpublished...
N: 1
Unpublished **2101**; ()
54-11-5 ; Nicotine
Row 2
Species = Oncorhynchus mykiss
Dataset = Training
Test duration = 96h
/
itetko
Only visible to Cadaster
○Don't save
●Save as new

● LC50 aquatic > 1000.0 mg/L
Unpublished...
N: 2
Unpublished **2101**; ()
61-82-5 ; Amitrole
Row 3
Species = Oncorhynchus mykiss
Dataset = Training
Test duration = 96h
/
itetko
Only visible to Cadaster
○Don't save
●Save as new

● LC50 aquatic = 0.1 mg/L
Unpublished...
N: 3
Unpublished **2101**; ()
101-05-3 ; Anilazine
Row 4
Species = Oncorhynchus mykiss
Dataset = Training
Test duration = 96h
/
itetko
Only visible to Cadaster
○Don't save
●Save as new

Figure 6. Data preview during the batch upload.

Data preview. After an intermediate page that shows the upload progress, the user proceeds to next step, as displayed in Figure 6. At this stage, the user is presented a summary of the batch upload. The summary indicates the recognized or non-recognized columns, missing required columns, as well as some basic statistics regarding the molecular structures – the number of structures retrieved from QSPR THESAURUS or from PubChem. It also contains the preview browser, which allows the user to preview the data in the almost exactly same way as it would be stored in the QSPR THESAURUS database.

At the data preview step, the records are generally divided into four groups - v*alid records, internal duplicates, external duplicates* and *unrecoverable errors.*

A *valid record* is a record that fulfills all the requirements and is ready to be uploaded and saved. This record has a white or yellow background. It can be marked for upload or skipped.

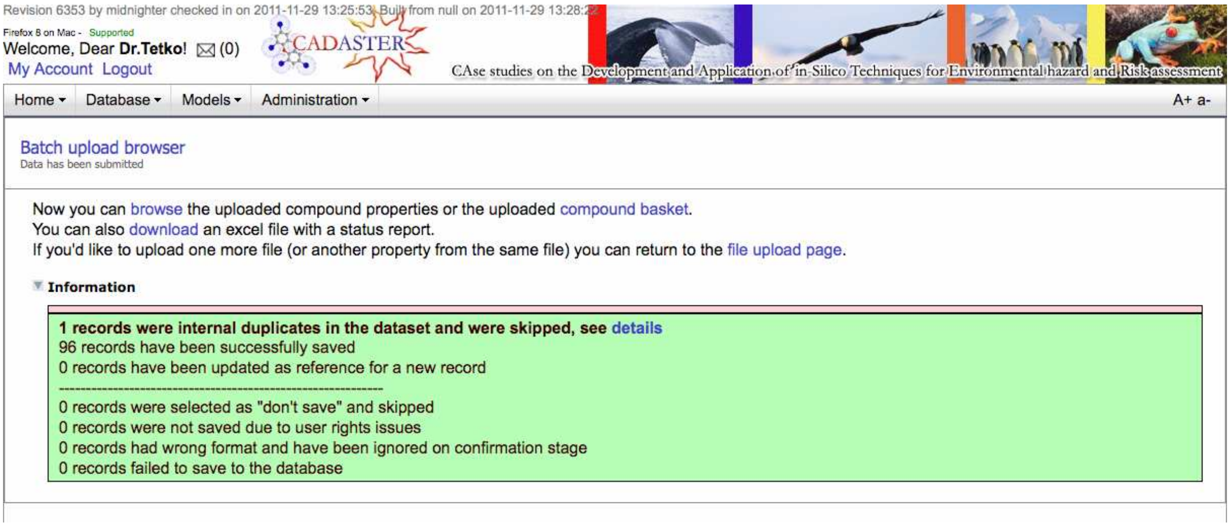
An *internal duplicate* is a record that from the QSPR THESAURUS point of view is an exact duplicate of some other record in the uploaded sheet. These records can not be uploaded.

An *external duplicate* is a record that from the QSPR THESAURUS point of view is a duplicate of some other record in the database, e.g., a duplicate of a record uploaded before. The rules for duplicate detection are described in more detail below in the “Data quality and consistency” section. The external duplicate can be skipped, saved as a duplicate, or overwritten. Saving as a duplicate saves the record, but marks it with red color. External duplicates should be reviewed and corrected after uploading. When an overwrite option is selected, the system attempts to replace the existing record with a new one only if the user has sufficient access privileges. Otherwise, any attempt to overwrite the existing record will be rejected.

An *unrecoverable error* is any record that according to the system rules is invalid, e.g., wrong format or invalid value in the property column, a missing article or obligatory condition. A user can use a drop-down button to review the cause of the error. Such records cannot be saved to the database in their current form.

The preview browser has special filters and navigation buttons to help a user to review specific subsets of the uploaded dataset. At this stage it is also possible to download the Excel file generated with all the changes made in the previous step.

Data upload. After revision and verification of the data in the preview mode, the uploading process can be finalized by saving the records into the QSPR THESAURUS database.


Revision 6353 by midnighter checked in on 2011-11-29 13:25:53 Built from null on 2011-11-29 13:28:22
Firefox 8 on Mac - Supported
Welcome, Dear Dr.Tetko! (0)
My Account Logout
CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment
Home
Database
Models
Administration
A+ a-

Batch upload browser
Data has been submitted

Now you can browse the uploaded compound properties or the uploaded compound basket.
You can also download an excel file with a status report.
If you'd like to upload one more file (or another property from the same file) you can return to the file upload page.

Information

1 records were internal duplicates in the dataset and were skipped, see details
96 records have been successfully saved
0 records have been updated as reference for a new record
----
0 records were selected as "don't save" and skipped
0 records were not saved due to user rights issues
0 records had wrong format and have been ignored on confirmation stage
0 records failed to save to the database

Figure 7. Final report of batch data upload. The statistics of the uploaded data are indicated.

A final upload report (Figure 7) is generated and displayed. It contains error messages for each skipped record with a brief explanation of the reason why the record has not been uploaded. It is possible to export the report as an Excel file for further revision.
After the upload process is complete, the user can proceed to the *compound property browser* to inspect the freshly uploaded records and to work with them.

Sources of information

One of the basic principles of the QSPR THESAURUS database policy is a strict requirement to provide the source of information for each experimental measurement introduced to the database. Most chemical databases do not store this information, which makes it difficult to verify the data and to correct them if needed.

QSPR THESAURUS supports two types of sources: articles (publications in scientific journals) and books (or chapters of books). There are a number of supplementary fields for each type of source: the title, the abstract, the journal, PubMed identifier, DOI identifier, ISBN number, web link etc. For each source it is possible to store a PDF file, which makes it easier to verify the data later on. For legal reasons a PDF file uploaded by a particular user is accessible by this user only.

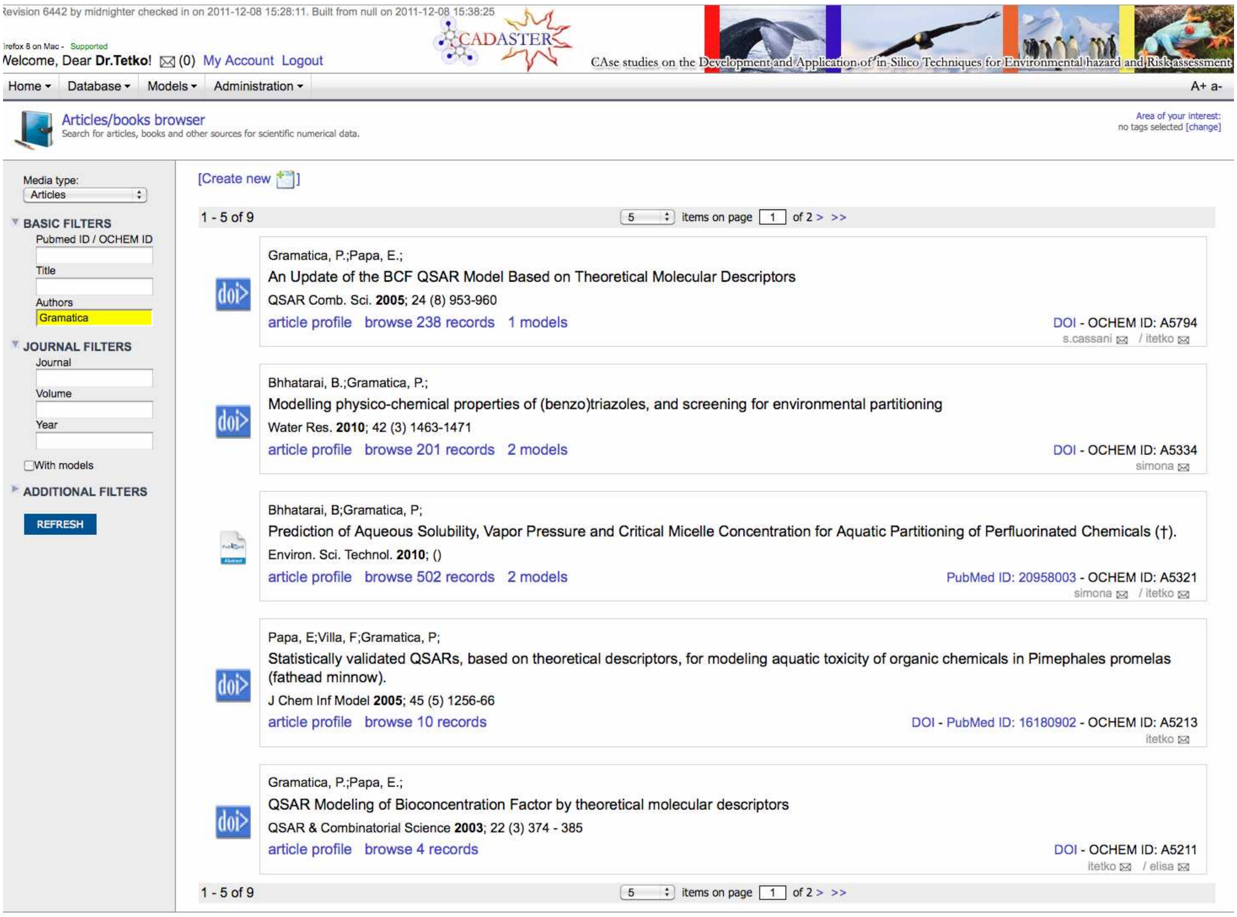
There are several ways to add a new article to the database:

- automatically retrieve the article from the PubMed database. This requires a PubMed identifier.
- upload from an external citation file. Currently the database supports the RIS, EndNote, BibTex and ISI formats. Such files are frequently provided by publishers.
- input all the fields manually. This is the most tedious and error prone way and should be used only if the PubMed identifier is not known and no article citation file is available

Similarly to the addition of a new article, there are several ways to introduce a new book:

- automatic retrieval by the ISBN number
- manual input of all fields

Manually introduced articles and books can be edited later. If a publication has been retrieved automatically (via PubMed or ISBN identifier), further modifications are forbidden to ensure the consistency of information.



Revision 6442 by midnighter checked in on 2011-12-08 15:28:11. Built from null on 2011-12-08 15:38:25
Firefox 8 on Mac - Supported
Welcome, Dear Dr.Tetko! (0) My Account Logout
CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment
Home ▾ Database ▾ Models ▾ Administration ▾
A+ a-
Articles/books browser
Search for articles, books and other sources for scientific numerical data.
Area of your interest:
no tags selected [change]
Media type:
Articles
BASIC FILTERS
Pubmed ID / OCHEM ID
Title
Authors
Gramatica
JOURNAL FILTERS
Journal
Volume
Year
□With models
ADDITIONAL FILTERS
REFRESH
[Create new]
1 - 5 of 9
5 items on page 1 of 2 > >>

Gramatica, P.;Papa, E.;
An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors
QSAR Comb. Sci. **2005**; 24 (8) 953-960
article profile browse 238 records 1 models
DOI - OCHEM ID: A5794
s.cassani / itetko

Bhhatarai, B.;Gramatica, P.;
Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning
Water Res. **2010**; 42 (3) 1463-1471
article profile browse 201 records 2 models
DOI - OCHEM ID: A5334
simona

Bhhatarai, B;Gramatica, P;
Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals (†).
Environ. Sci. Technol. **2010**; ()
article profile browse 502 records 2 models
PubMed ID: 20958003 - OCHEM ID: A5321
simona / itetko

Papa, E;Villa, F;Gramatica, P;
Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow).
J Chem Inf Model **2005**; 45 (5) 1256-66
article profile browse 10 records
DOI - PubMed ID: 16180902 - OCHEM ID: A5213
itetko

Gramatica, P.;Papa, E.;
QSAR Modeling of Bioconcentration Factor by theoretical molecular descriptors
QSAR & Combinatorial Science **2003**; 22 (3) 374 - 385
article profile browse 4 records
DOI - OCHEM ID: A5211
itetko / elisa

Figure 8. The article browser also provides information about the number of records in the article, the introducer, and the modifier of the article.

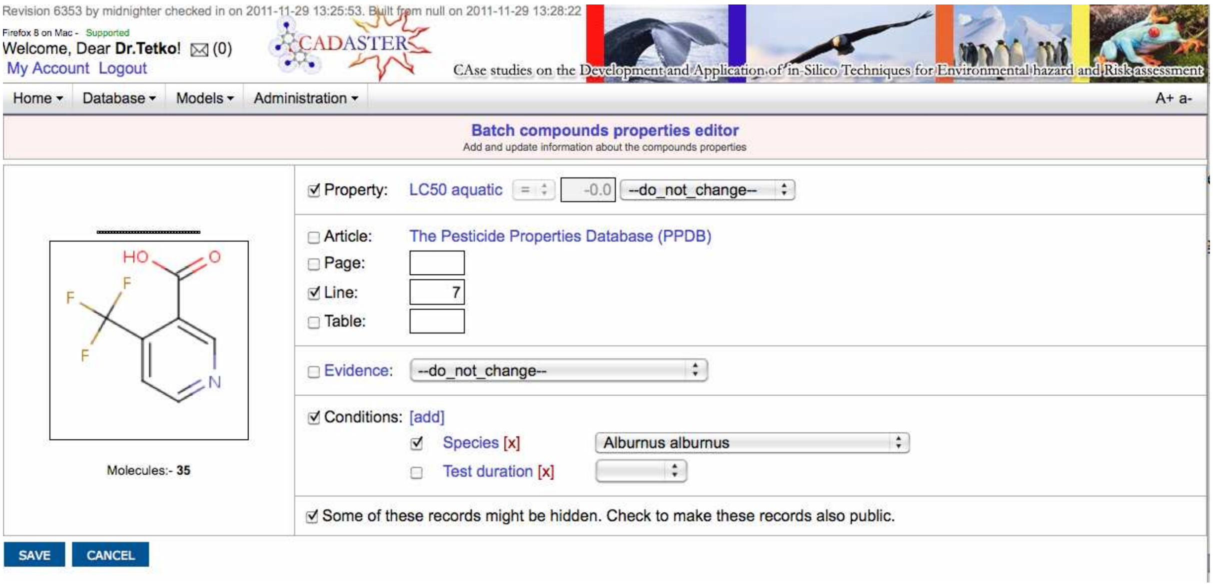
All publications stored in the QSPR THESAURUS database can be searched for and accessed from the *article browser* (Figure 8). Similar to the other browsers in our system, the *article browser* has a set of filters, which allow search by author, PubMed ID, ISBN, title, journal etc. From this browser, the user can easily navigate to the experimental measurements associated with a particular publication.

Data access and manipulation

All data records stored in the QSPR THESAURUS database can be easily modified, filtered, searched and eventually grouped and organized for further convenient work.

Batch editing

Data modification can be done in two ways: by either editing a single record separately or by using the batch editing tool (Figure 9) to work with multiple records simultaneously. In the editor window, the user can change all fields of a record: measured property, article information, structure and conditions. Batch editing is useful to correct systematic errors that might have occurred during batch upload, e.g., wrong units or missing conditions of experiment.



Revision 6353 by midnighter checked in on 2011-11-29 13:25:53. Built from null on 2011-11-29 13:28:22

Firefox 8 on Mac - Supported

Welcome, Dear Dr.Tetko! ✉ (0)

My Account Logout

CADASTER

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Home ▾ Database ▾ Models ▾ Administration ▾ A+ a-

Batch compounds properties editor

Add and update information about the compounds properties

☑ Property: LC50 aquatic = -0.0 --do_not_change--

☐ Article: The Pesticide Properties Database (PPDB)

☐ Page:

☑ Line: 7

☐ Table:

☐ Evidence: --do_not_change--

☑ Conditions: [add]

☑ Species [x] Alburnus alburnus

☐ Test duration [x]

☑ Some of these records might be hidden. Check to make these records also public.

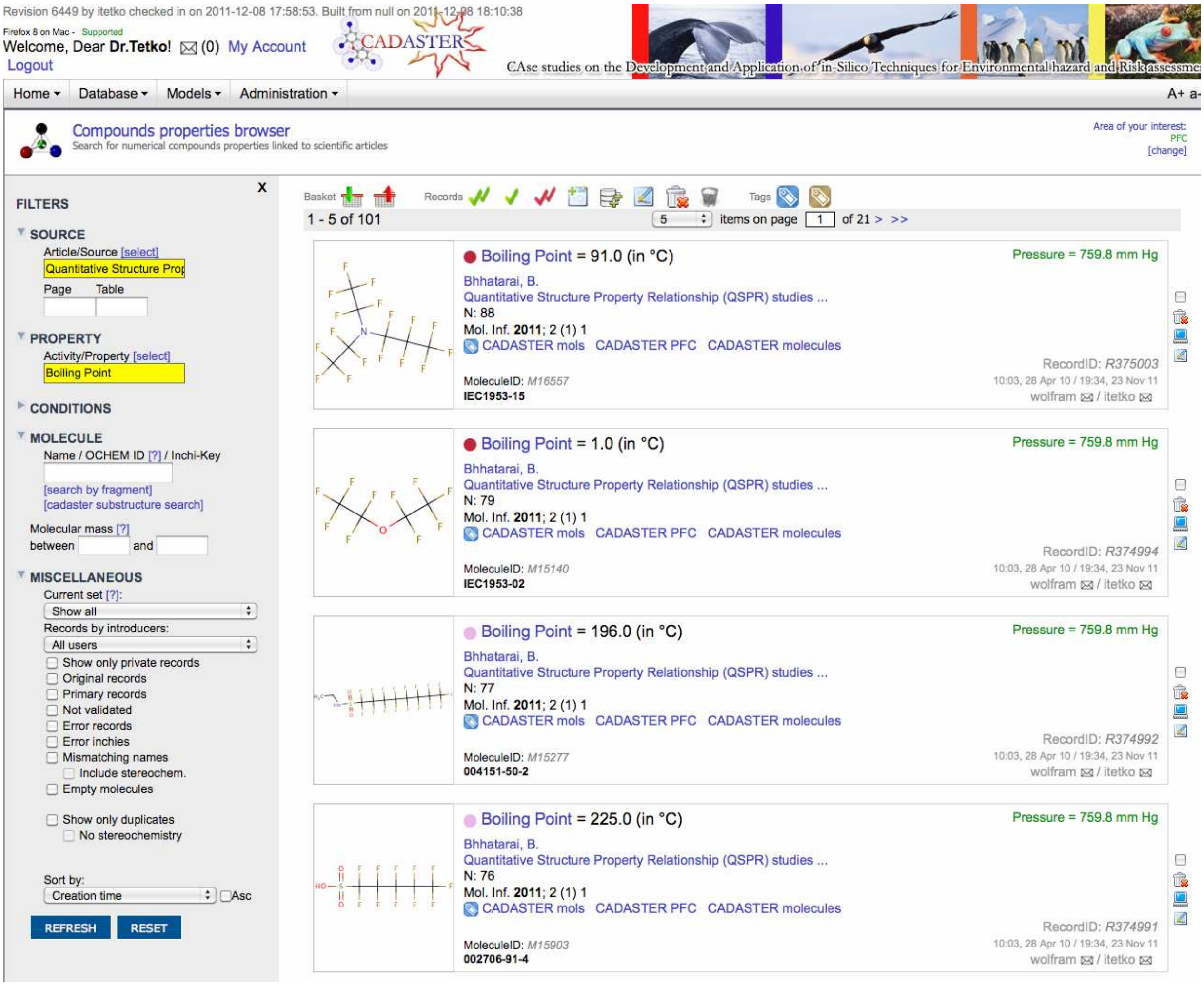
Molecules:- 35

SAVE CANCEL

Figure 9. The batch editor allows changing several features of records simultaneously. The checkboxes indicate features that will be changed.

Filtering and searching

Every browser of QSPR THESAURUS, e.g., molecules, articles, properties, etc., has a panel for data filtering. Filters are used to focus on a certain subset of the data, e.g., a set of certain properties, a set of specific organisms or conditions, etc. Records can be filtered by literature source (article or book where the data has been published), physicochemical property or experimental condition and structural information, e.g., molecule name or InChI key as well as by molecular sub-fragments. Comprehensive filter options to find duplicates, errors or non-validated entries are available.



Revision 6449 by itetko checked in on 2011-12-08 17:58:53. Built from null on 2011-12-08 18:10:38

Firefox 8 on Mac - Supported

Welcome, Dear **Dr.Tetko**! ✉ (0) My Account

Logout

CAse studies on the Development and Application of in Silico Techniques for Environmental hazard and Risk assessme

Home ▾ Database ▾ Models ▾ Administration ▾ A+ a-

Compounds properties browser

Search for numerical compounds properties linked to scientific articles

Area of your interest: PFC [change]

FILTERS

▾ SOURCE

Article/Source [select]

Quantitative Structure Proţ

Page Table

▾ PROPERTY

Activity/Property [select]

Boiling Point

▸ CONDITIONS

▾ MOLECULE

Name / OCHEM ID [?] / Inchi-Key

[search by fragment]

[cadaster substructure search]

Molecular mass [?]

between and

▾ MISCELLANEOUS

Current set [?]: Show all

Records by introducers: All users

- □ Show only private records
- □ Original records
- □ Primary records
- □ Not validated
- □ Error records
- □ Error inchies
- □ Mismatching names
  - □ Include stereochem.
- □ Empty molecules

- □ Show only duplicates
  - □ No stereochemistry

Sort by: Creation time □Asc

REFRESH RESET

Basket Records Tags

1 - 5 of 101 5 items on page 1 of 21 > >>

● Boiling Point = 91.0 (in °C) Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 88
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M16557*
IEC1953-15
RecordID: *R375003*
10:03, 28 Apr 10 / 19:34, 23 Nov 11
wolfram ✉ / itetko ✉

● Boiling Point = 1.0 (in °C) Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 79
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M15140*
IEC1953-02
RecordID: *R374994*
10:03, 28 Apr 10 / 19:34, 23 Nov 11
wolfram ✉ / itetko ✉

● Boiling Point = 196.0 (in °C) Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 77
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M15277*
004151-50-2
RecordID: *R374992*
10:03, 28 Apr 10 / 19:34, 23 Nov 11
wolfram ✉ / itetko ✉

● Boiling Point = 225.0 (in °C) Pressure = 759.8 mm Hg
Bhhatarai, B.
Quantitative Structure Property Relationship (QSPR) studies ...
N: 76
Mol. Inf. **2011**; 2 (1) 1
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: *M15903*
002706-91-4
RecordID: *R374991*
10:03, 28 Apr 10 / 19:34, 23 Nov 11
wolfram ✉ / itetko ✉

Figure 10. Filtering options in the database: article and property filters (highlighted) are used. The CADASTER substructure search can be also done using Ambit software through web services.

QSPR THESAURUS supports relational filtering (Figure 10): for a given record, the user can find other records with the same structure (preserving or ignoring stereochemistry) or the records that have been modified at approximately the same time as this record.

Data can also be filtered by tags, i.e., labels assigned to molecules or properties. A desired set of tags, referred to as an "*area of interest",* restricts all the displayed information (experimental measurements, publications, properties and compounds) to the selected tags. The tags are used to organize data according to four analyzed sets of molecules.

Data organization

By applying various filters, a user can specify and select records of interest that can be stored separately in sets called *baskets*. A typical use of a basket is to assign both training and validation sets for modeling. The content of a basket can be browsed and modified from the compound properties browser or from the basket browser.

Public and private records. By default, all records introduced to the database are publicly available unless the user explicitly makes the records private. The private records are only visible to CADASTER users. This allows data curation by the CADASTER project participants before making them publicly available. The login and registration is only required for CADASTER participants. All other users can access data after login in as a guest user.

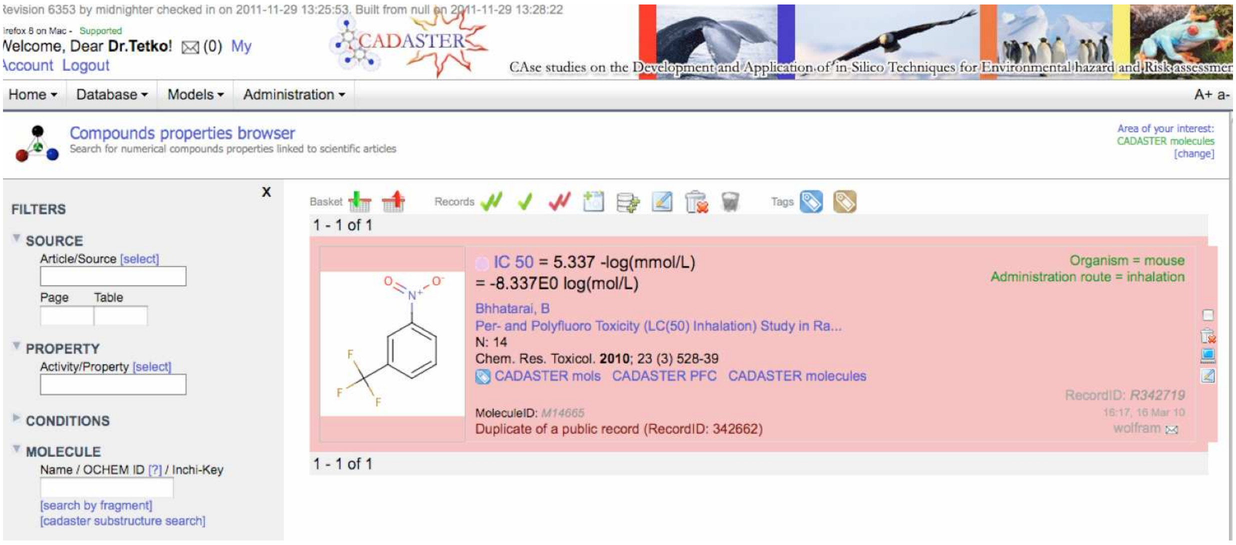
Data download. The users can download data from CADASTER web site (e.g., in Excel or SDF format). During the download user can specify which additional information (e.g., InChi, position in the article, introducer/modifier, etc.) is required. This information will be provided as tags in the sdf file or as columns in the Excel file.

Data quality and consistency

Control of errors, data origin and quality

An experimental measurement can be marked as an “error”. Such records are highlighted with a red background and indicate a possible problem. The system allows users to manually mark a record as an error if they believe there is a mistake. In this case, the user should provide an explanation of the problem in the comment or discussion field related to this record. The QSPR THESAURUS system can also automatically mark records as erroneous if they do not comply with the system rules. Namely, a record is automatically marked with red color if:

- an obligatory condition of the experiment has not been specified (for example, a boiling point measurement without specifying the pressure is ambiguous and would be marked as an error automatically)
- a duplicate of the record exists in the database (see the next section for the definition of "duplicate")



Revision 6353 by midnighter checked in on 2011-11-29 13:25:53. Built from null on 2011-11-29 13:28:22
irefox 8 on Mac - Supported
Welcome, Dear Dr.Tetko! (0) My Account Logout
CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessmen
Home
Database
Models
Administration
A+ a-
Compounds properties browser
Search for numerical compounds properties linked to scientific articles
Area of your interest: CADASTER molecules [change]
FILTERS
SOURCE
Article/Source [select]
Page
Table
PROPERTY
Activity/Property [select]
CONDITIONS
MOLECULE
Name / OCHEM ID [?] / Inchi-Key
[search by fragment]
[cadaster substructure search]
Basket
Records
Tags
1 - 1 of 1
O
N+
O-
F
F
F

IC 50 = 5.337 -log(mmol/L)
= -8.337E0 log(mol/L)
Bhhatarai, B
Per- and Polyfluoro Toxicity (LC(50) Inhalation) Study in Ra...
N: 14
Chem. Res. Toxicol. **2010**; 23 (3) 528-39
CADASTER mols CADASTER PFC CADASTER molecules
MoleculeID: M14665
Duplicate of a public record (RecordID: 342662)

Organism = mouse
Administration route = inhalation
RecordID: R342719
16:17, 16 Mar 10
wolfram

1 - 1 of 1

Figure 11. A record, which is a duplicate of a public record, is shown in red color.

Another quality indicator is the “to be verified” flag. This flag signals that the record has been introduced from a referencing article, e.g., benchmarking/methodological article and should be verified against the original publication. This flag can be set either manually or automatically by the system (e.g., in case of batch data upload, see the “Batch upload” section for details).

Duplicates management

To ensure data consistency, it is essential to avoid redundancy in the database. Thus, there is a need for strict rules for the definition of duplicates. In QSPR THESAURUS two experimental records of a physicochemical or biological property are considered to be duplicates if they are obtained for the same compound under the same conditions, have the same measured value (with a precision up to 3 significant digits) and are published in the same article (Figure 11). We refer to these records as *strong duplicates*, as opposed to *weak duplicates*, for which only part of the information is the same. The QSPR THESAURUS database does not forbid strong duplicates completely, but forces all the duplicates (except for the record introduced first) to be explicitly marked with red color. This ensures that there are no strong duplicates among the valid (i.e., non-error) records.

The uniqueness of chemical compounds is controlled by special molecular hashes, referred to as InChI-Keys. Namely, for the determination of duplicated experimental measurements, two chemical structures are considered the same if they have identical Inchi-keys.

QSPR THESAURUS allows weak duplicates (for example, completely identical experimental values, published in different articles) and provides facilities to find them. Moreover, in the modeling process, it is always automatically ensured that the same compounds in the training set appear only in one fold of the N-fold cross-validation process.

Experimental data origin

Each record has a colored dot indicating the origin of the data. Green dots indicate "*original records*" from publications with a description of experimental protocols; these are usually the publications where the property was originally measured (original data). The users can verify experimental conditions and experiments by reading these articles. These are the most reliable records in the database. The weak duplicates of *original records* have magenta dots. The other records have red dots and originate from articles that re-use the original data but for which the original records are not stored. These are frequently methodological QSAR/QSPR studies. The original records can be easily filtered out by checking a corresponding box in the *compound property browser*. Another filter, "*primary records*", eliminates all weak duplicates except the record with the most early publication date.

Available Data

The database contains 6838 experimental data within the four classes of chemicals (2482 chemicals). This number counts primary records, i.e. data points (value + property) without duplicates. For example, the database contains an experimental value of 4.21 log P (125225-28-7, ipconazole) from the Bhhatarai et al[2] article. This value is also reported in the US EPA database (2010) as well as in the Pesticide manual of Tomlin, 1995.[3] The primary record is the one with the oldest publication date, i.e. Tomlin, 1995. All other data points are duplicates of this primary record. For the provided statistics we considered only non-duplicated data points. They can be accessed in the database by selecting "primary records" checkbox.

Table 1. Brominated flame retardants, including BDEs (243 structures with at lest one experimental value)

Endpoints – groups	*Number of data*
Physical Chemical Properties	419
Environmental fate parameters	119
Aquatic and terrestrial ecological effects parameters	18
Other effect data	537
Total	**1093**

Table 2. Perfluoroalkylated substances (PFC) (691 structures including 454 with at least one experimental value)

Endpoints – groups	*Number of data*
Physical Chemical Properties	701
Environmental fate parameters	78
Aquatic and terrestrial ecological effects parameters	138
Other effect data	808
Total	**1725**

Table 3. Substituted musks/fragrances (532 structures including 160 with at least one experimental value)

Endpoints – groups	*Number of data*
Physical Chemical Properties	343
Environmental fate parameters	85
Aquatic and terrestrial ecological effects parameters	144
Other effect data	566
Total	**1138**

Table 4. Triazoles and Benzotriazoles (TAZ/BTAZ) (447 structures including 299 with at least one experimental value)

Endpoints – groups	*Number of data*
Physical Chemical Properties	588
Environmental fate parameters	204
Aquatic and terrestrial ecological effects parameters	933
Other effect data	1001
Total	**2726**

The data available in CADASTER database were collected from numerous articles as well as from public databases. The original source of information is provided for each record.

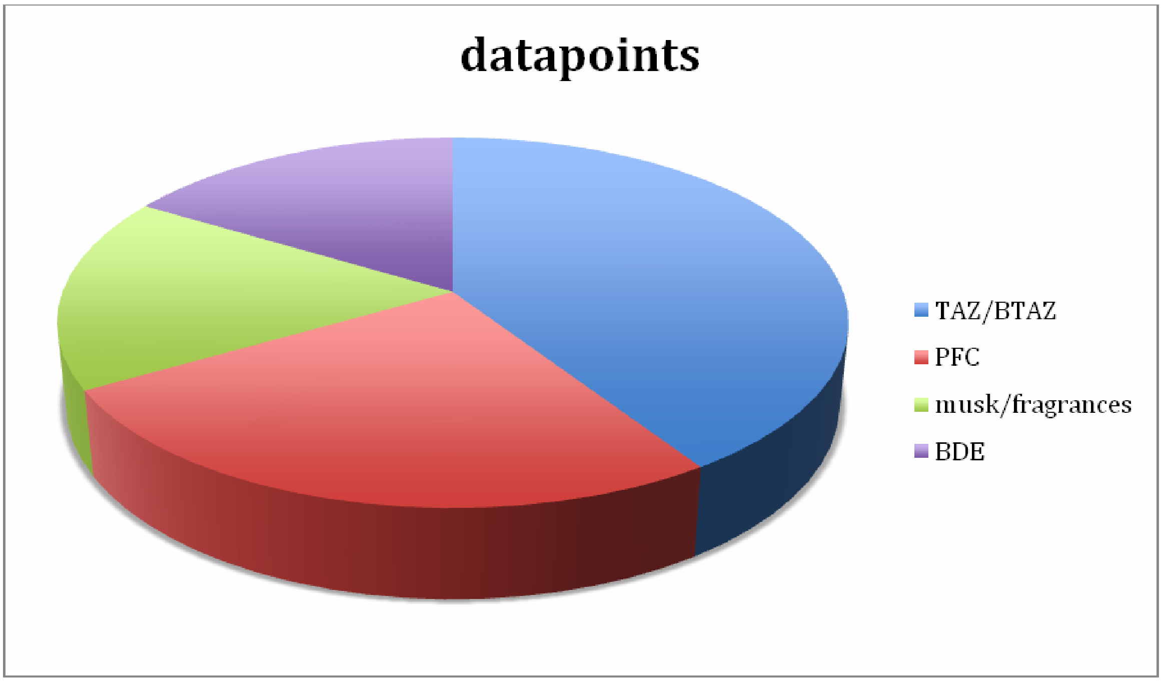
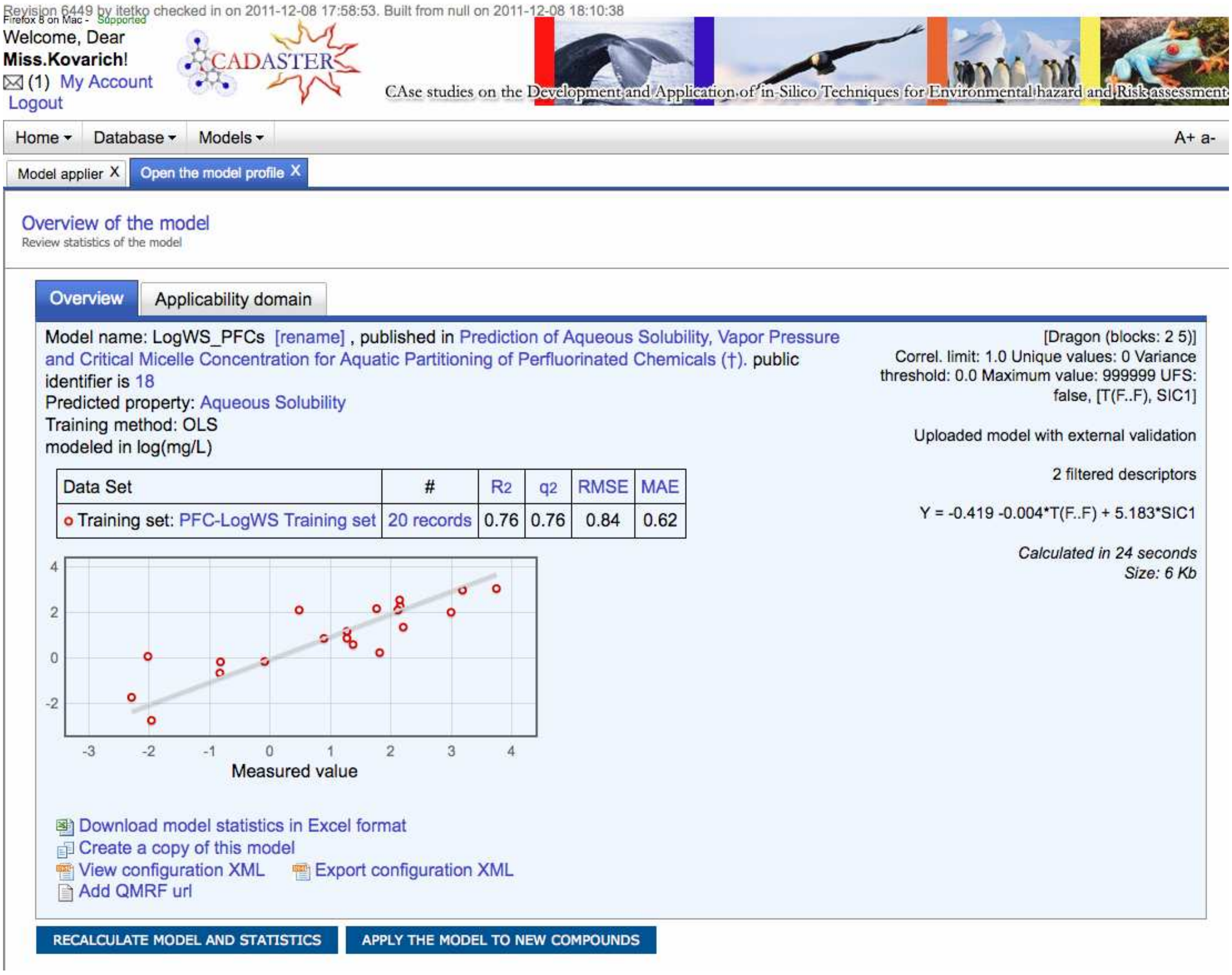


Figure 12. The distribution of data across different classes indicates that largest numbers of experimental data are available for TAZ/BTAZ.

Model Upload

The database is integrated with models. The tools available on the site of the project provide a possibility to upload models developed by project participants, which are than published on the web site of the project (Figure 13).



Revision 6449 by itetko checked in on 2011-12-08 17:58:53. Built from null on 2011-12-08 18:10:38

Firefox 8 on Mac - Supported

Welcome, Dear **Miss.Kovarich!**

✉ (1) My Account

Logout

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Overview of the model

Review statistics of the model

Model name: LogWS_PFCs [rename] , published in Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals (†). public identifier is 18

Predicted property: Aqueous Solubility

Training method: OLS

modeled in log(mg/L)

Data Set	#	R2	q2	RMSE	MAE
○ Training set: PFC-LogWS Training set	20 records	0.76	0.76	0.84	0.62

[Dragon (blocks: 2 5)]
Correl. limit: 1.0 Unique values: 0 Variance threshold: 0.0 Maximum value: 999999 UFS: false, [T(F..F), SIC1]

Uploaded model with external validation

2 filtered descriptors

Y = -0.419 -0.004*T(F..F) + 5.183*SIC1

Calculated in 24 seconds
Size: 6 Kb

Download model statistics in Excel format

Create a copy of this model

View configuration XML

Export configuration XML

Add QMRF url

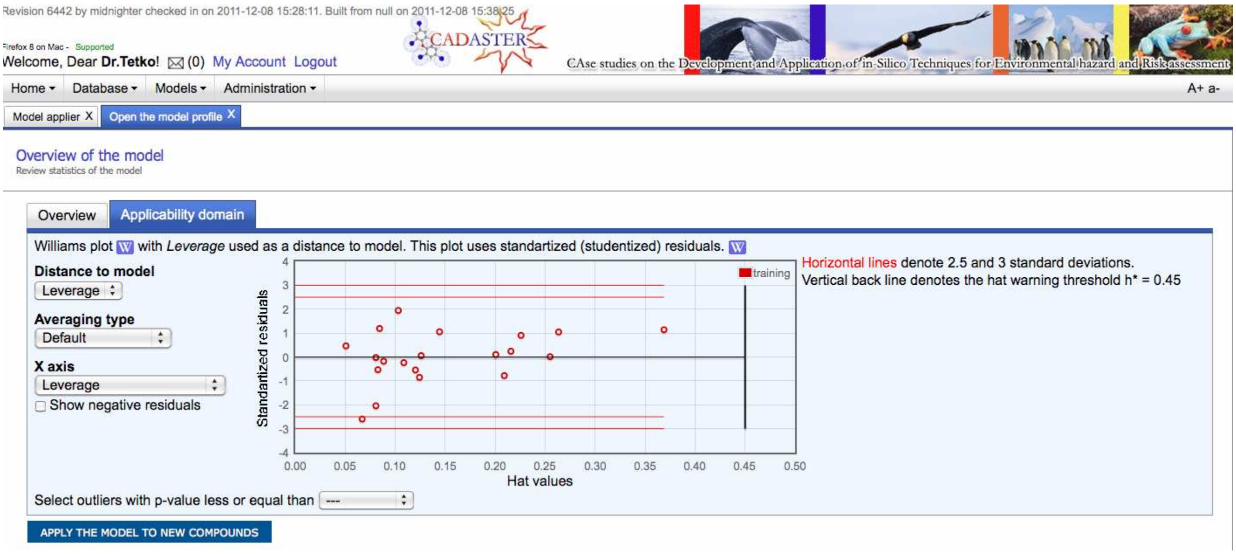
Figure 13. A model published on the CADASTER web site. A click on statistical parameters (R^2, Q^2, RMSE, and MAE) opens a wiki page with their description. All these parameters are computed using results calculated by a model. Additional statistical parameters can be specified and provided using QRMF. The user who uploaded the model has a possibility to provide it (see “Add QRMF url” link).

The upload of linear models developed using OLS and PLS approaches is currently supported. It is also possible to develop models using other algorithms on http://qspr-thesaurus.eu and publish them on the web site of the project.

Estimation of the accuracy of predictions.

Applicability domain assessment for linear models

All uploaded or developed models come with the estimation of the accuracy of prediction. For linear models, the Williams plot, which reports the standardized residuals in y-axis and the hat values (h) from the H matrix in x axis is used. For chemicals with available experimental data, the h* (leverage) threshold is used to identify predictions for chemicals that are more dissimilar, in the model space, to the majority of chemicals (x axis), while the outliers for the response are those with predicted values lower/higher than 3 standardized residuals (y axis)[4] as shown in (Figure 14). Several other measures to estimate applicability domain of models are provided by integration of CADASTER database with web services provided by Ambit software. The user can specify them when uploading his/her model and they will be calculated for the prediction of new compounds.



Revision 6442 by midnighter checked in on 2011-12-08 15:28:11. Built from null on 2011-12-08 15:38:25

Firefox 8 on Mac - Supported

Welcome, Dear **Dr.Tetko**! ✉ (0) My Account Logout

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Overview of the model

Review statistics of the model

Williams plot with *Leverage* used as a distance to model. This plot uses standartized (studentized) residuals.

Distance to model Leverage

Averaging type Default

X axis Leverage

□ Show negative residuals

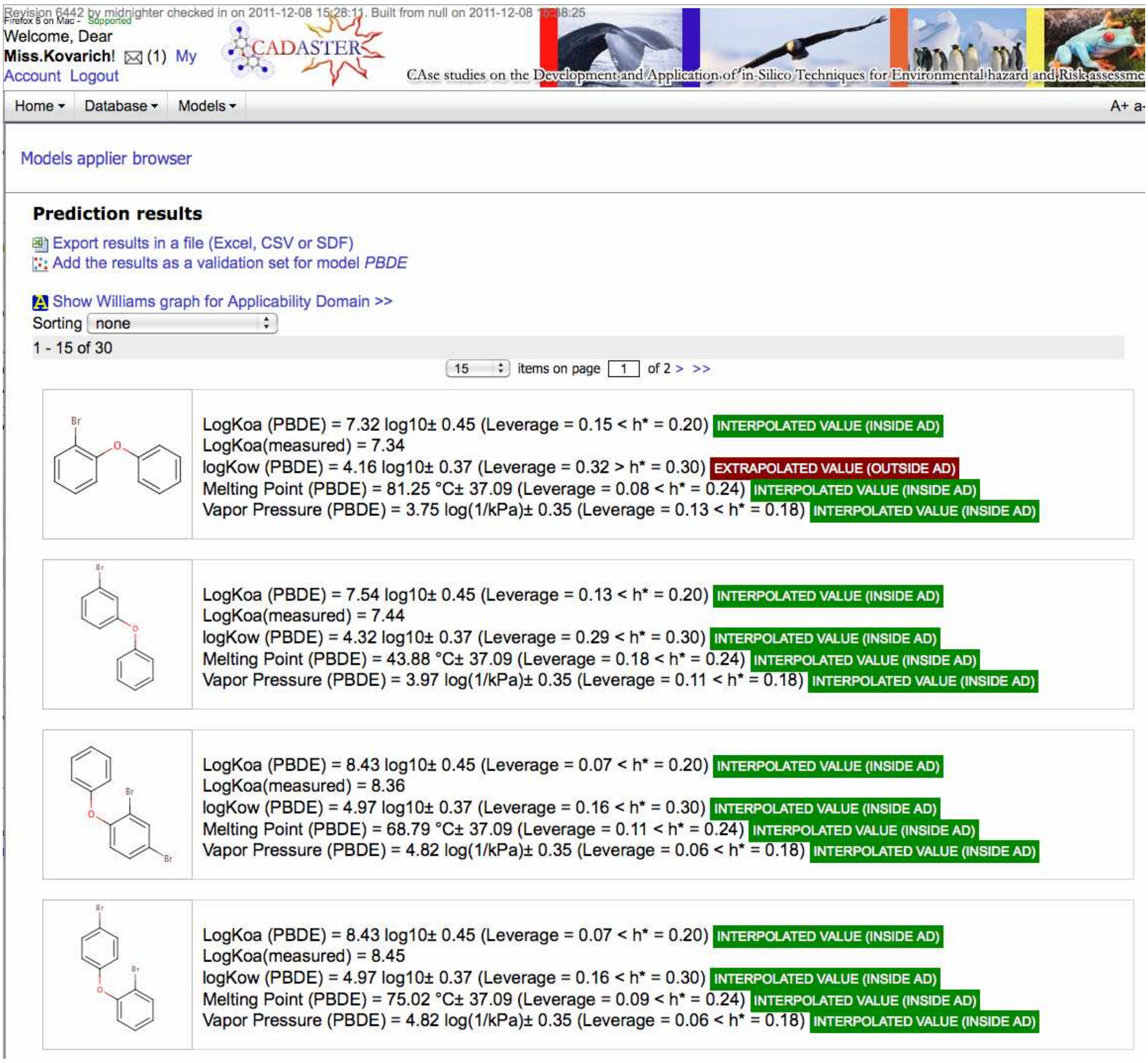
Horizontal lines denote 2.5 and 3 standard deviations.
Vertical back line denotes the hat warning threshold h* = 0.45

Select outliers with p-value less or equal than ---

Figure 14. Williams plot for a linear model.

For the prediction of new compounds, the h value can be used for verifying if the model can be structurally applicable to a new chemical[4-6]. If the leverage h value is lower of the cut-off value of h* the chemical is inside the structural AD of the model and its predicted data is interpolated, while if its h value is higher than h* its prediction is less reliable, being extrapolated (Figure 15).

This information on the possibility of interpolation vs. extrapolation is highly important and may help the external user to decide whether he or she is going to use the provided predictions for the estimation of the given molecule.



Revision 6442 by midnighter checked in on 2011-12-08 15:28:11. Built from null on 2011-12-08 15:38:25

Firefox 8 on Mac - Supported

Welcome, Dear **Miss.Kovarich!** ✉ (1) My Account Logout

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessme

Models applier browser

Prediction results

Export results in a file (Excel, CSV or SDF)

Add the results as a validation set for model *PBDE*

Show Williams graph for Applicability Domain >>

Sorting: none

1 - 15 of 30

15 items on page 1 of 2 > >>

LogKoa (PBDE) = 7.32 log10± 0.45 (Leverage = 0.15 < h* = 0.20) INTERPOLATED VALUE (INSIDE AD)
LogKoa(measured) = 7.34
logKow (PBDE) = 4.16 log10± 0.37 (Leverage = 0.32 > h* = 0.30) EXTRAPOLATED VALUE (OUTSIDE AD)
Melting Point (PBDE) = 81.25 °C± 37.09 (Leverage = 0.08 < h* = 0.24) INTERPOLATED VALUE (INSIDE AD)
Vapor Pressure (PBDE) = 3.75 log(1/kPa)± 0.35 (Leverage = 0.13 < h* = 0.18) INTERPOLATED VALUE (INSIDE AD)

LogKoa (PBDE) = 7.54 log10± 0.45 (Leverage = 0.13 < h* = 0.20) INTERPOLATED VALUE (INSIDE AD)
LogKoa(measured) = 7.44
logKow (PBDE) = 4.32 log10± 0.37 (Leverage = 0.29 < h* = 0.30) INTERPOLATED VALUE (INSIDE AD)
Melting Point (PBDE) = 43.88 °C± 37.09 (Leverage = 0.18 < h* = 0.24) INTERPOLATED VALUE (INSIDE AD)
Vapor Pressure (PBDE) = 3.97 log(1/kPa)± 0.35 (Leverage = 0.11 < h* = 0.18) INTERPOLATED VALUE (INSIDE AD)

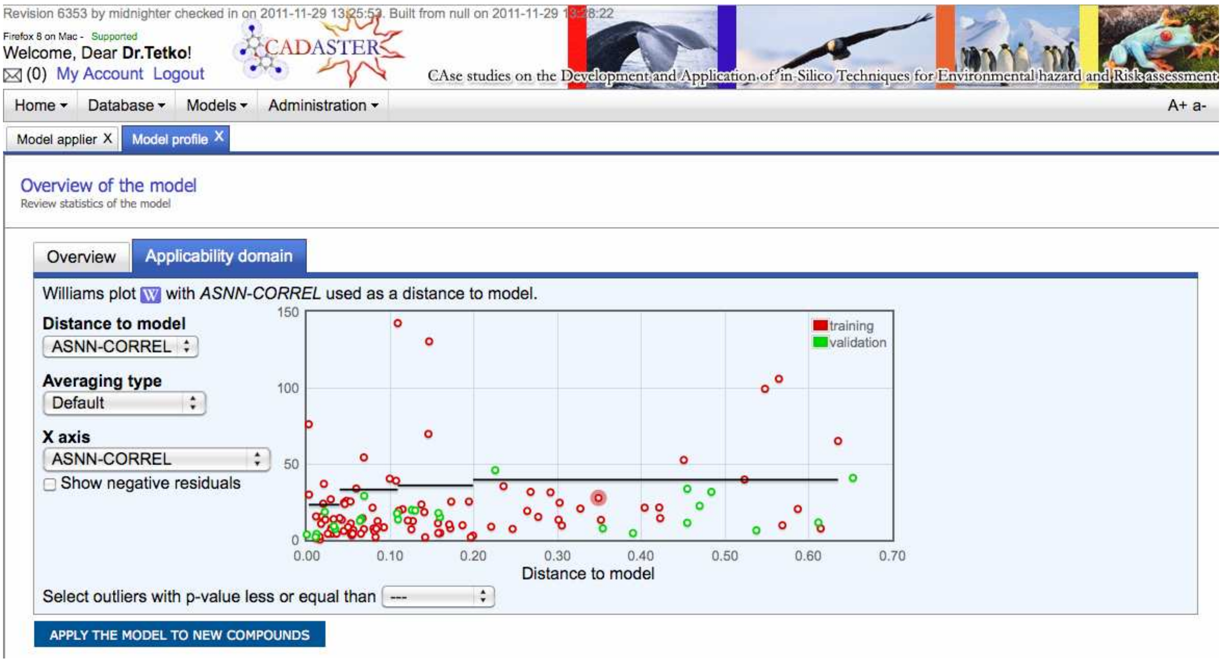
LogKoa (PBDE) = 8.43 log10± 0.45 (Leverage = 0.07 < h* = 0.20) INTERPOLATED VALUE (INSIDE AD)
LogKoa(measured) = 8.36
logKow (PBDE) = 4.97 log10± 0.37 (Leverage = 0.16 < h* = 0.30) INTERPOLATED VALUE (INSIDE AD)
Melting Point (PBDE) = 68.79 °C± 37.09 (Leverage = 0.11 < h* = 0.24) INTERPOLATED VALUE (INSIDE AD)
Vapor Pressure (PBDE) = 4.82 log(1/kPa)± 0.35 (Leverage = 0.06 < h* = 0.18) INTERPOLATED VALUE (INSIDE AD)

LogKoa (PBDE) = 8.43 log10± 0.45 (Leverage = 0.07 < h* = 0.20) INTERPOLATED VALUE (INSIDE AD)
LogKoa(measured) = 8.45
logKow (PBDE) = 4.97 log10± 0.37 (Leverage = 0.16 < h* = 0.30) INTERPOLATED VALUE (INSIDE AD)
Melting Point (PBDE) = 75.02 °C± 37.09 (Leverage = 0.09 < h* = 0.24) INTERPOLATED VALUE (INSIDE AD)
Vapor Pressure (PBDE) = 4.82 log(1/kPa)± 0.35 (Leverage = 0.06 < h* = 0.18) INTERPOLATED VALUE (INSIDE AD)

Figure 15. Prediction of properties using CADASTER models. Interpolated and extrapolated values are indicated.

Applicability domain assessment of the ASNN models

The estimation of the accuracy of neural network models, which are also published on the web site of the project, is based on the concept of "distance to a model" (DM)[7,8], i.e., some numerical value estimated solely from molecular structures and experimental conditions, which correlates with the average model performance (Figure 16). Currently several DMs are supported: the standard deviation of an ensemble of models (STDEV), the correlation in the space of models (CORREL). The DMs are calibrated against the accuracy of models for the training set using N-fold cross-validation as described elsewhere[4]. The estimated accuracy of predictions as a function of the respective DM is visualized on the *accuracy averaging plot* (see Figure 16), which shows the absolute values of the prediction residuals versus the respective DM. The DMs are used to estimate the prediction accuracies for new molecules. The same methodology has been recently extended for classification models.[9]



Revision 6353 by midnighter checked in on 2011-11-29 13:25:53. Built from null on 2011-11-29 13:28:22

Firefox 8 on Mac - Supported

Welcome, Dear **Dr.Tetko**!

Overview of the model

Review statistics of the model

Williams plot with *ASNN-CORREL* used as a distance to model.

Distance to model: ASNN-CORREL

Averaging type: Default

X axis: ASNN-CORREL

☐ Show negative residuals

Select outliers with p-value less or equal than ---

Figure 16. The accuracy plot for neural network model. The solid black line show average prediction error as function of a distance to model, that is standard deviation of ensemble predictions. It is used to provide a quantitative estimation of the accuracy of predictions for new molecules.

Summary

The database contains convenient tools for upload, search, editing, curation, tracking and download of experimental data. It is integrated with models that were developed during the CADASTER project. It provides an easy and professional way to store and share information about the environmental toxicity of chemical compounds. The estimation of the accuracy of predictions and applicability domain of models allows external user to make a proper choice on whether the calculated values of a particular models can be used to avoid experimental measurements.

References

1. Sushko, I. et al., Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011,** *25* (6), 533-54.

2. Bhhatarai, B.; Gramatica, P., Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Res* **2011,** *45* (3), 1463-71.

3. Tomlin, C., *The Pesticide Manual: Incorporating The Agrochemicals Handbook*. Blackwell Science Inc: 1995.

4. Roy, P.P. et al. QSAR model reproducibility and applicability: a case study of rate constants of hydroxy radical reaction models applied to Polybrominated Diphenyl Ethers and (Benzo) Triazoles, *J. Comput. Chem.,* **2011**, 32, 2386-2396.

5. Papa,E. et al Development, Validation and Inspection of the Applicability Domain of QSPR Models for physico-chemical properties of Polybrominated DiphenylEthers *QSAR Comb. Sci.*, **2009,** 28 (8), 790-796.

6. Bhatarai, B. and Gramatica P.Per- and Poly-fluoro Toxicity (LC50 inhalation) Study in Rat and Mouse using QSAR Modeling. *Chem. Res. Toxicol.,* **2010**, 23(3), 528-539.

7. Tetko, I. V. et al, Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008,** *48* (9), 1733-46.

8. Sushko, I. Applicability domain of QSAR models. Technical University of Munich, Munich, 2011.

9. Sushko, I. et al, Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *J. Chem. Inf. Model.* **2010,** *50* (12), 2094-111.