

Training on the OCHEM database
and modelling environment



www.ochem.eu

<http://docs.eadmet.com>

Table of Contents

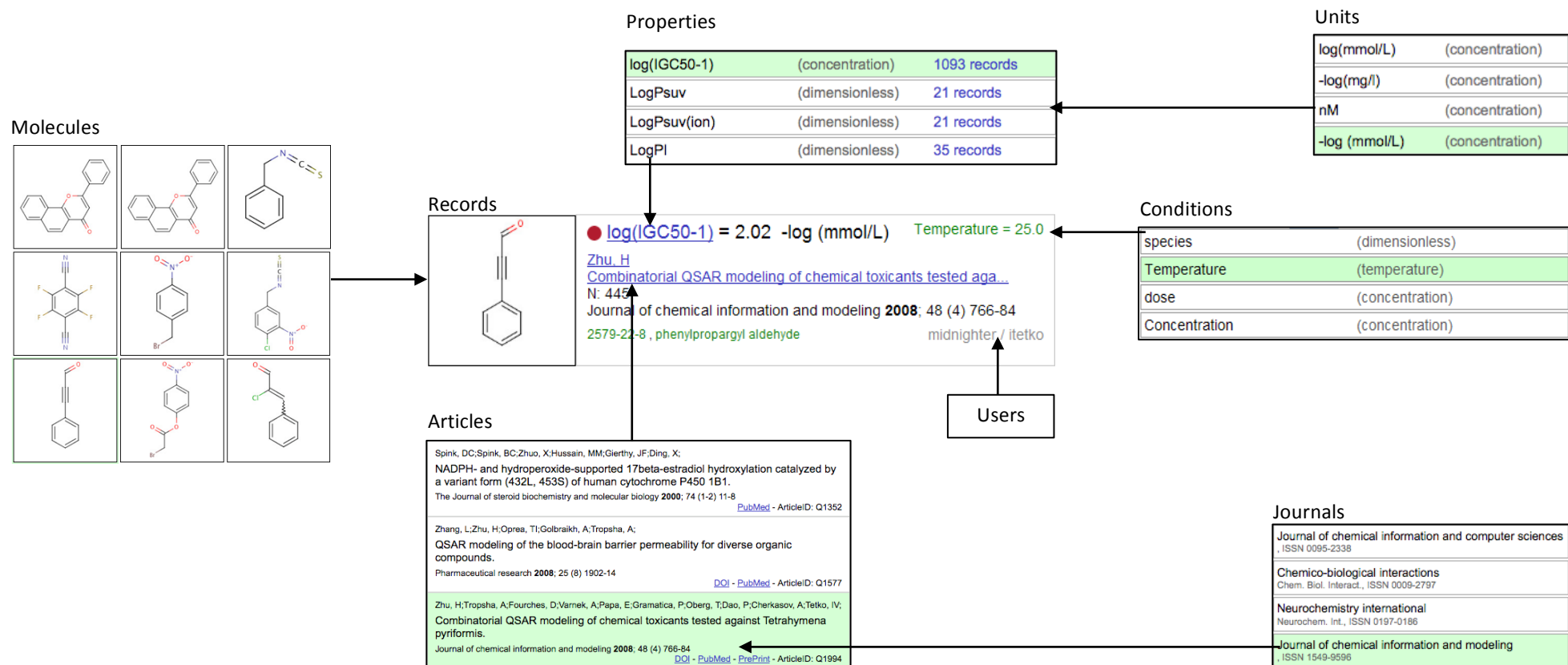
Table of Contents	3
1. General concepts	4
1.1 Simplified data structure.....	5
1.2 OCHEM Browsers: different yet similar	6
1.3 OCHEM Editors and item profiles.....	7
2. Working with data	8
2.1 Property introduction	9
2.2 Article introduction	10
2.3 Data upload	11
2.4 Review your uploaded data in default basket.....	14
2.5 Using filters to create training and test sets for QSAR	15
3. Modelling framework	16
3.1 Overview	17
3.2 Select the training sets, method and data pre-processing options	18
3.3 Configure molecular descriptors.....	19
3.4 Configure the training method and start calculations	20
3.5 Wait for the calculations to finish	21
3.6 Save your model.....	22
3.7 The model profile: review your model.....	23
3.8 Applicability domain and model export	24
3.9 Model application	25
4. Advanced features	28
4.1 Using OCHEM via web-services and KNIME	29
4.2 Comprehensive modelling	33
4.3 ToxAlert utility.....	37
4.4 Set Compare utility.....	38
4.5 Pathway Analysis.....	39

1. General concepts

In this chapter the general concepts of OCHEM are introduced. In particular these are:

- 1.1 Simplified data structure
- 1.2 OCHEM Browsers: different yet similar
- 1.3 OCHEM Editors and item profiles

1.1 Simplified data structure



Experimental property (or “record”) – a value for a property for a specific molecule published in a specific article or book.

This means that:

One molecule can have multiple records associated with it (measurements for different properties, measurements for the same property published in different articles, etc.)

One article can hold multiple records for multiple properties for multiple molecules

Most of essential OCHEM operations (such as QSAR modeling) are performed on datasets of records (and not molecules)

1.2 OCHEM Browsers: different yet similar

Properties browser
Please search property database before creating new ones

Area of your interest: no tags selected [change]

Type part of name to filter: [search] [Create new property] [Create new group]

1 - 10 of 10

W	logP Chloroform/Water	(Dimensionless / log10)	2 records	The partition coefficient between chloroform and water ...	midnighter
W	BCF	(Dimensionless / log10)	238 records	Bioconcentration factor BCF = [Concentration of X in Organism] ...	ExpDesign / Iletko
W	pKa (smiles as ob. cond.)	(Dimensionless / log10)	376 records	this pKa requires as a condition a smiles string with the i ...	Koerner
W	Aqueous Solubility	(Concentration / -log(mol/L))	8402 records	Solubility of chemical compounds in water (aqueous solubili ...	Iletko / enamine
W	AMES	(qualitative)	6542 records	This assay measures genetic damage at the single base level ...	vlad121 / Iletko
W	log(GC50-1)	(Concentration / -log(mmol/L))	1093 records	The toxic potency of chemicals, measured by their concentra ...	Iletko / mojca
W	CYP450 modulation	(qualitative)	7485 records	CYP450 modulation describes substances in terms of their sp ...	vkovallshyn / charochkina
W	logS part of Aqueous Solubility [x]	(Concentration / mg/L)	8402 records	Logarithm of intrinsic solubility in water of non-ionized m ...	Anil / Iletko
W	LogD	(Dimensionless / log10)	1 records	The distribution coefficient of octanol/water measured at s ...	mojca
W	logPow	(Dimensionless / log10)	17351 records	The partition coefficient is a ratio of concentrations of u ...	Iletko

1 - 10 of 10

Main browsers:

- Experimental property browser (record browser)
- Molecule browser
- Properties browser
- Conditions browser
- Units browser
- Article browser
- Journals browser
- Baskets browser
- Tags browser
- Models browser

Compounds properties browser
Search for numerical compounds properties linked to scientific articles

Area of your interest: no tags selected [change]

1 - 5 of 41490

5 Items on page 1 of 8298 >>

Filters:

- SOURCE: Article/Source [select]
- Page: [] Table: []
- PROPERTY: Activity/Property [select]
- CONDITIONS: []
- MOLECULE: Name / OCHEM ID [?] / Inchi-Key []
- Molecular mass [?] between [] and []
- MISCELLANEOUS:
 - Current set [7]: []
 - Show all []
 - Records by introducers: []
 - All users []
 - Show only private records []
 - Original records []
 - Primary records []
 - Not validated []
 - Error records []
 - Error in chis []
 - Mismatching names []
 - Include sterechem. []
 - Empty molecules []
- Sort by: []
- Creation time []

REFRESH RESET

1

2

3

4

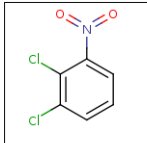
5

Main elements of a browser:

1. Items
2. Page bars – can be used to navigate items
3. Filters – can be specified to narrow down the displayed items to a specific area
4. Global toolbars – can be used either to manipulate (delete, modify) several items simultaneously, or create new items
5. Item toolbars – can be used to perform operations on a specific item

1.3 OCHEM Editors and item profiles

Compound property editor
Add and update information about the compounds properties



Names: [\[add\]](#) [\[check\]](#)
Synonyms:
0 - 0 of 0
0 - 0 of 0

Property: **BCF**

Article: **An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors**

Page:

Line:

Table:

N (Mol ID):

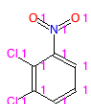
Evidence:

Comment:

Conditions: [\[add\]](#)
[Dataset \[x\]](#)

This record is public

Molecule editor
You can submit or edit Molecule here.

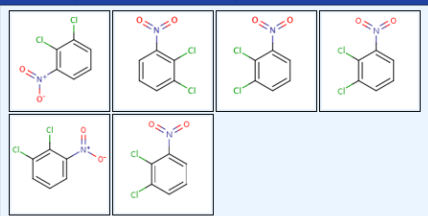


You can introduce molecule:
either A) Draw your molecule (above in JME)
either B) Paste your molecule

either C) Upload from a file

either D) Fetch a molecule structure by name from PubChem database

Names **Alternate Depictions** Properties



Basket editor
Add new basket or edit exiting basket

Name:
(min. 2 characters)

[Create a copy of this basket](#)
[Create a primary records basket](#)
[Add or delete particular records](#)
[Discretize the numerical values](#)
[Transform the basket using OScript](#)

Statistics of the basket

Properties	Records	Unique compounds
log(IGC50-1)	644 records	640 compounds
Total Compounds (ignoring stereo-chemistry)	640 (640) compounds	

Articles	Count
Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.	644

TagsCount

Main editors (mirroring the browsers):

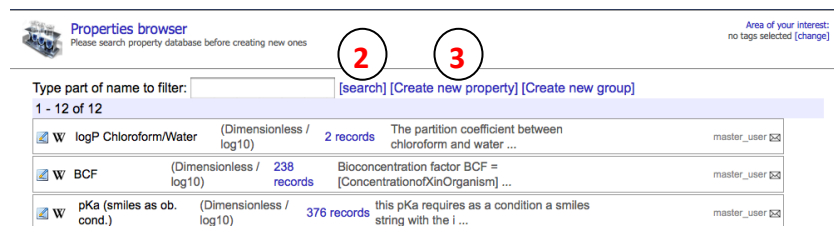
- Compound property editor(record editor)
- Molecule editor
- Property editor
- Condition editor
- Unit editor
- Article editor
- Journal editor
- Baskets editor
- Tag editor

2. Working with data

In this Chapter the prerequisites for model training are set. In particular these are:

- 2.1 Property introduction
- 2.2 Article introduction
- 2.3 Data introduction
- 2.4 Organization of uploaded data
- 2.5 Data “baskets” as training and test sets

2.1 Property introduction



4 **Property/Condition editor**
Add new property or edit existing property

Name: Type: ☐ Hidden property ☐ Make public [?]
property name is good and length is 27

System of units: Default unit:

Tags [+]: No tags selected

Obligatory Conditions [+]:

Aliases:
(comma separated list of the other names, i.e. synonyms, for this property)

Description:
This property was especially introduced for the Strasbourg Summer School hands-on session.
warning: property description length is 90 and may not be informative

4 **Property/Condition editor**
Add new property or edit existing property

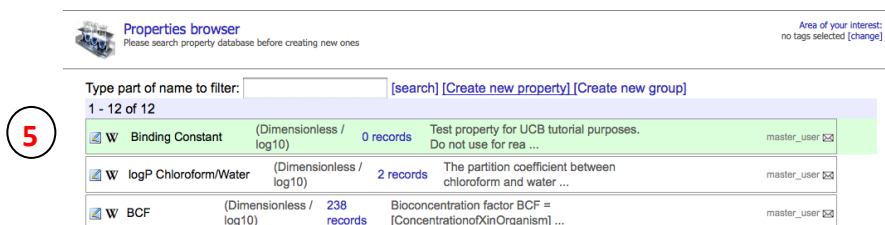
Name: Type: ☐ Hidden property ☐ Make public [?]
property name is good and length is 29

Tags [+]: No tags selected

Obligatory Conditions [+]:

Aliases:
(comma separated list of the other names, i.e. synonyms, for this property)

Description:
This property was especially introduced for the Strasbourg Summer School hands-on session.
warning: property description length is 90 and may not be informative



If the database doesn't yet contain the property to be modelled, It has to be introduced. Therefore do the following steps:

Property for regression model: "RegressionModelProperty"

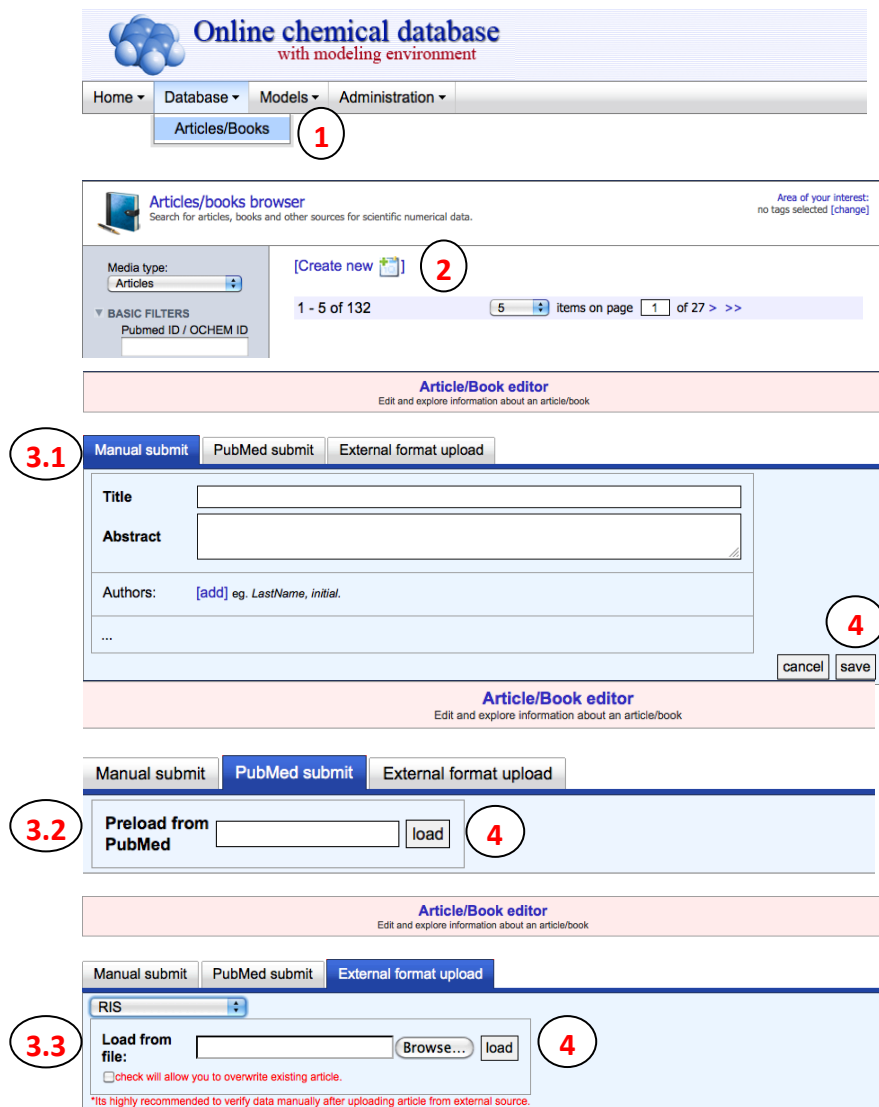
1. Select the "Properties" item in the "Database" submenu of the main OCHEM menu. You will open a properties browser with the list of existing properties.
2. Search for the property to find out if it exists in the database
3. Select "[Create new property]" item in the global toolbar of the properties browser. You will open a property editor with empty fields.
4. Fill in the information about the property
 - 4.1. Enter the name for a new property (RegressionModelProperty in our example).
 - 4.2. Define type of the property ("Numeric" for regression models).
 - 4.3. Select an appropriate system of units (Concentration) and a default unit for the property ("log(mol/L)").
 - 4.4. Provide a brief description to let other users know what this property represents.
 - 4.5. When ready click "Save".

Property for classification model: "ClassificationModelProperty"

1. Select the "Properties" item in the "Database" submenu of the main OCHEM menu. You will open a properties browser with the list of existing properties.
2. Search for the property to find out if it exists in the database
3. Select "[Create new property]" item in the global toolbar of the properties browser. You will open a property editor with empty fields.
4. Fill in the information about the property
 - 4.1. Enter the name for a new property (ClassificationModelProperty)
 - 4.2. Define type of the property ("Qualitative" for classification models)
 - 4.3. Provide a brief description to let other users know what this property represents.
 - 4.4. Add options for this property (active, inactive)
 - 4.5. When ready click "Save".
5. The newly created property appears in the properties browser.

Note: Both properties for this tutorial already exist in the database. So do not check "Make public" checkbox to avoid duplication conflicts.

2.2 Article introduction



Online chemical database
with modeling environment

Home Database Models Administration

Articles/Books **1**

Articles/books browser
Search for articles, books and other sources for scientific numerical data.

Media type: Articles

[Create new] **2**

1 - 5 of 132 5 items on page 1 of 27 > >>

BASIC FILTERS
Pubmed ID / OCHEM ID

Article/Book editor
Edit and explore information about an article/book

3.1 Manual submit PubMed submit External format upload

Title

Abstract

Authors: [add] eg. LastName, initial.

...

cancel save **4**

Article/Book editor
Edit and explore information about an article/book

Manual submit PubMed submit External format upload

3.2 Preload from PubMed

load **4**

Article/Book editor
Edit and explore information about an article/book

Manual submit PubMed submit External format upload

RIS

3.3 Load from file: Browse... load **4**

☐ check will allow you to overwrite existing article.

*its highly recommended to verify data manually after uploading article from external source.

Besides the property it is necessary to introduce an article to the database to be able to store records in the system.

1. Select the "Articles/Books" item in the "Database" submenu of the main OCHEM menu. You will open an articles/books browser with some existing articles.
2. Select "[Create new]" item in the global toolbar of the articles/books browser. You will open an article/book editor with empty article fields.
3. There are different article introduction options available:
 - 3.1. Manual submit
Fill out the relevant article data: title, abstract, list of authors, journal, publication date, issue number, pages, etc.
 - 3.2. PubMed submit
Provide a PubMed id and load the article information automatically.
 - 3.3. External format upload
Upload the article information from an external file in a certain format like RIS, EndNote, BibTex, ISI.
4. Click "save" or "load" to save the article.
5. Review the OCHEM article item in the articles/books browser.

Note: The example file article_1.ris from the supplementary webpage was already introduced to the system. The user will be informed about this fact when (s)he tries to upload this file.

5

eADMET, T.;
A beginner's guide to OCHEM I
Nature Reviews Computer Science **2012**; 11 (5) 100 - 111
article profile browse 0 records
OCHEM ID: A1000000000
master_user


With a property and an article defined, structures with their measured values can be uploaded. So use the “Batch data upload” tool to introduce records.

1. Select the “Batch data upload” item in the “Database” submenu of the main OCHEM menu. You will open first page of the “Batch upload wizard”.
2. Select your file in the “Upload file” field. The tool supports SDF and XLS file formats.
3. Click the big “Upload this file” button to continue to the next step.

4. Second page of the wizard is the “column remapping” page. Here you can preview the first few lines of your uploaded file and see which columns were recognized by the tool. On this page you also have the possibility to reassign column names and select/unselect columns for upload.

5. The column holding the data values is named “tutorial ...” in the uploaded file. We need to specify that these values represent the “RegressionModelProperty” property. Click on the red unrecognized “tutorial ...” column header and select “Property” from the popup menu.







- The “Property browser” will open. Here you have the possibility to search for and select the relevant property. Select the “RegressionModelProperty” property by clicking the green select icon.


Properties browser
Area of your interest: no tags selected [change]

Please search property database before creating new ones

Type part of name to filter:
[search] [Create new property] [Create new group]

1 - 11 of 11

		Binding Constant	(Dimensionless / log10)	153 records	Test property for UCB tutorial purposes. Do not use for rea...	master_user 52
		logP Chloroform/Water	(Dimensionless / log10)	2 records	The partition coefficient between chloroform and water ...	master_user 52
		BCF	(Dimensionless / log10)	238 records	Bioconcentration factor BCF = [ConcentrationXinOrganism] ...	master_user 52

OCHEM 1.3 Tutorial Handout

SMILES	NAME	RegressionModelProperty	UNIT (buccal LC50 aquatic)	Species	Test duration	CASRN	Database
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1alpha,2alpha,3beta,4alpha,5alpha,6beta)-1,2,3,4,5,6-hexa...	-0.422	log(μ M/L)	Danio rerio	96h	58-89-9	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	0.11	log(mmol/L)	fathead minnow	96h	107-20-9	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	0.582	log(μ M/L)	Danio rerio	96h	319-84-6	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	18.9	mg/L	fathead minnow	96h	2218-51-5	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	2.92	-log(M)	fathead minnow	96h	2218-51-5	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	-0.52870288941562	log(mmol/L)	Pimephales promelas	96h	10293-06-8	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	17.0	mg/L	fathead minnow	96h	464-48-2	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	0.89	log(mmol/L)	fathead minnow	96h	96-29-7	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	0.26	log(mmol/L)	fathead minnow	96h	1482-15-1	Training
C[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O	(1R,2R,3R,4R,5R,6R)-1,2,3,4,5,6-hexa...	5.77	mg/L	fathead minnow	96h	1126-79-0	Validation

Upload this sheet

Database entities remapping page
On this page you can remap property, condition and unit names (if they were not found in database)

Database entities remapping

Property: **RegressionModelProperty**

Values
Unit: **log(μ M/L)**, min value: -2.56, max value: 3.52
Unit: **log(mmol/L)**, min value: -6.38, max value: 2.1
Unit: **mg/L**, min value: 6.8E-5, max value: 75200.0
Unit: **-log(M)**, min value: 1.91, max value: 8.45
Unit: **-log(mol/L)**, min value: 0.84, max value: 6.52
Unit: **log(mol/L)**, min value: -5.26, max value: -0.64
Unit: **μ M**, min value: 172.0, max value: 6901.0
Unit: **-log(mmol/L)**, min value: -0.85, max value: 3.13
Unit: **-log(μ M/L)**, min value: -3.05, max value: -3.05
Unit: **ppm**, min value: 7500.0, max value: 7500.0

Condition: **Species**

Options
Danio rerio
fathead minnow
Pimephales promelas
Oncorhynchus mykiss
Salmo gairdneri
Lepomis macrochirus
Guppy

Oryzias latipes
Salmonidae
Zebrafish
Bluegill sunfish
Cyprinidae
Rasbora heteromorpha
Poecilia reticulata
Cyprinus carpio
Leiostomus xanthurus
Lepomis macrochirus
Carassius carassius
Cyprinodon variegatus
Gambusia affinis
Carassius auratus
Micropterus salmoides
Ictalurus punctatus
Anguilla japonica

Condition: **Test duration**

Options
96h

Condition: **Dataset**

Options
Training
Validation

Article: **unpublished**

Molecule set: **default**

Articles/books browser
Search for articles, books and other sources for scientific research

Media type: **Articles**

[Create new]

1 - 5 of 133 items on page 1 of 27 >>

Pubmed ID / OCHEM ID

Title

Authors

Journal

Volume

Year

With models

ADDITIONAL FILTERS

Area of your interest: no tags selected [change]

eADMET, T.;
A beginner's guide to OCHEM I
computer Science 2012; 11 (5) 100 - 111
browse 153 records
OCHEM ID: A100000000
master_user 82

Telko, IV;Tanchuk, VY;Villa, AE;
Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices.
J Chem Inf Comput Sci 2001; 41 (5) 1407-21
article profile browse 10118 records
DOI - PubMed ID: 11604042 - OCHEM ID: A5920
master_user 82

Article: **A32451**

Molecule set: **default**

submit

- Notice that the column header changes to dark green (recognized property), the header name is now "RegressionModelProperty", and the checkbox in the column header is checked, indicating that the column will be processed by the tool.
 - Click the big "Upload this sheet" button to proceed to page three of the wizard.
 - Third page of the wizard is the "entity remapping" page. You can review and change some aspects of the uploaded data (property, unit used for data upload, article, etc.)
 - Since no article has been specified in the data sheet, a stub "Unpublished" was put instead of the article. With "Ignore warnings" checkbox selected the data can be uploaded to the database. In this case the data will be introduced by default as hidden data and is only visible to the current user (recommended option for the tutorial exercise and continue at point 13.). To upload data originally published in an article click on the "Unpublished" link in the "Article" section of the page.
- Note: Although it is possible to upload data with an article stub, it is strongly recommended to provide an article as a published source of the data if it is available.
- An "Articles/books" browser opens. You have the possibility to search for the relevant article or even create a new one. Select the earlier created "A beginner's guide to OCHEM I" as the article by clicking the green select icon for it (✓). It might be required to filter for the title e.g.
 - The article changed to the OCHEM id of this article. The error message has disappeared. Now all the data will be uploaded to this article.
 - Click the "submit" button to continue.

Note: The upload for Classification data works in the same way. Just select the file that contains classification data and map it in step 5. to your property (or use the already existing property "ClassificationModelProperty")



14



Batch upload preview
Upload of data via CSV, SDF or Excel files.

Batch upload preview browser

Summary:
All rows in the sheet Count: **153**
Type: valid Status: preprocessed Count: **153**

Filter by row number: and row type: **all rows** Batch operations

1 - 2 of 153 2 items on page 1 of 77 > >>

<p>Row 0 eSave Skip</p> 	<p>● RegressionModelProperty = 3.12 (in %) eADMET, T. A beginner's guide to OCHEM... N: AUTO_1 Nature Reviews Computer Science 2013; 11 (5) 100 - 111 MoleculeID: M8218</p>	<p>Species = fathead minnow Test duration = 96h RecordID: R-812969 master_user \$8</p>
<p>Row 1 eSave Skip</p> 	<p>● RegressionModelProperty = 5.29 (in %) eADMET, T. A beginner's guide to OCHEM... N: AUTO_2 Nature Reviews Computer Science 2013; 11 (5) 100 - 111 MoleculeID: M3058</p>	<p>Species = fathead minnow Test duration = 96h RecordID: R-812970 master_user \$8</p>

1 - 2 of 153 5 items on page 1 of 77 > >>

Proceed with upload 16

Cancel Batch Upload Back One Step



17

Batch upload preview
Upload of data via CSV, SDF or Excel files.

Batch upload results
Batch upload is finished. You can review your [153 uploaded records](#) or look at the [detailed report of the upload](#).

Summary:
All rows in the sheet Count: **153**
Type: valid Status: uploaded Count: **153**

New Batch Upload

18

14. Depending on the size of the uploaded set, the process may take from seconds to hours to complete (>50000 data points).

15. The fourth page of the wizard is the data preview browser. Here you can review your records and determine any errors in the data upload process. The page holds information on the total number of records to be uploaded, the number of valid, and erroneous or duplicated records among them. You can select or unselect individual records from upload.

16. Since all records being uploaded are valid, continue the upload by clicking the big "Upload these records" button.

17. The upload itself is the slowest part in the process. It may take from seconds (for a hundred records) to several hours (for a large dataset of tens of thousands of records).

18. The final page of the batch data upload wizard gives some statistics about the uploaded data. You can review the uploaded data in the "Experimental property browser" or download a detailed report.

Note: Data uploaded with the batch upload tool are automatically put to a basket. Since the given data was already split into training and test sets, these default baskets can be reviewed ([chapter 2.4](#)) and used then later on for model training ([chapter 3.2](#))


2.4 Review your uploaded data in default basket

Online chemical database
with modeling environment













Home ▾ Database ▾ Models ▾

Baskets 1

Basket browser
Browse, Compare or Join molecule set


Filter by name: [Create new ] ☐ Show public sets

1 - 4 of 4



  	46014_classification_test.sdf	2474 records
  	46033_regression_test.sdf	383 records
  	46014_classification_train.sdf	2474 records
  	46014_regression_train.sdf	384 records

1 - 4 of 4

Basket editor
Add new basket or edit existing basket

Name: 

basket name is good and length is 18

 Create a copy of this basket
 Create a primary records basket
 Add or delete particular records
 Discretize the numerical values
 Transform the basket using OScript


3

Statistics of the basket

Properties	Records	Unique compounds
RegressionModelProperty	384 records	384 compounds
Total Compounds (ignoring stereo-chemistry)	384 (384) compounds	

Articles	Count
A beginner's guide to OCHEM I	384

Reviewing your baskets

1. Review a list of all your baskets by selecting the “Baskets” menu item from the Database submenu of the main OCHEM menu. This opens a basket browser.
2. Open an individual basket profile clicking the edit icon ().
3. The profile shows you brief information on the basket size, its content, articles, properties and tags. Here, you can also rename your basket or perform a number of advanced operations on it. Rename the “tutorial...” basket to e.g. “Training Set”.

2.5 Using filters to create training and test sets for QSAR

The screenshot displays the OCHEM 1.3 web interface. The top navigation bar includes 'Home', 'Database', 'Models', and 'Moderation'. The 'Database' dropdown is open, showing 'Compound properties' (circled 1), 'Molecules', and 'Properties'. The 'Compound properties' browser is active, showing a list of conditions. The left sidebar contains filters for 'SOURCE', 'PROPERTY', 'MOLECULE', and 'MISCELLANEOUS'. The 'PROPERTY' filter is set to 'ModelProperty' (circled 2). The 'MOLECULE' filter is set to 'Name / OCHEM ID' (circled 3). The 'MISCELLANEOUS' filter is set to 'Current set' (circled 3). The 'CONDITIONS' panel on the right shows a list of conditions, with 'Dataset' selected (circled 6). A dropdown menu for 'Dataset' is open, showing options like 'BioByte star', 'Test', 'Test 1', 'Test 2', 'Test 3', 'Test 4', 'Test 5', 'Test 6', 'Training', and 'Validation' (circled 7). A 'Select your molecule set' dialog box is open, showing 'Molecule set' as 'Create new set...' (circled 9) and 'New set name' as 'Regr. Training Set'.

To create training and test sets for a regression model and for a classification model, the data is already available at the system. The example property for regression is called "RegressionModelProperty" and for classification it is "ClassificationModelProperty".

Using compound properties browser filters to create training and test sets
Example for regression data:

1. Select the "Compound properties" item in the "Database" submenu of the main OCHEM menu. You will open a compound properties browser with all the data in OCHEM.
2. To filter out all data except for our property of interest (RegressionModelProperty), start typing in the property name into the Activity/Property filter. The auto-fill drop-down list will appear as you type. Select the "RegressionModelProperty" from the list and press "Return/Enter".
3. Click "refresh" button to apply the filter to the main area of the browser. The result shows a number of records, uploaded in advance to this tutorial.
4. Now we want to create a training and a test set from the data. Click "Select all" button (✓) in the global toolbar of the compound properties browser. This will put all records that match the current filters to a special "Selected records" set.
5. To separate the data into training and test set, add a filter for the condition.
6. Click on "Conditions" and "add" the condition "Dataset". Therefore the conditions browser will help to find and select this condition.
7. Select "Training" from the dropdown and click the refresh button and 384 records should be filtered.
8. Click the "Add selected records to basket" icon in the global toolbar (🛒).
9. Select the "Create new set..." item from the "Molecule set" field. Type in a new basket name. e.g. "Regr. Training Set" and click "Submit" to create the basket.
10. Change now the condition filter to "Validation" and after clicking "Refresh" button 383 records are shown.
11. Repeat step 7. – 9. by selecting "Validation" and name the basket "Regr. Validation Set".
12. Proceed in the same manner for classification data ("ClassificationModelProperty")

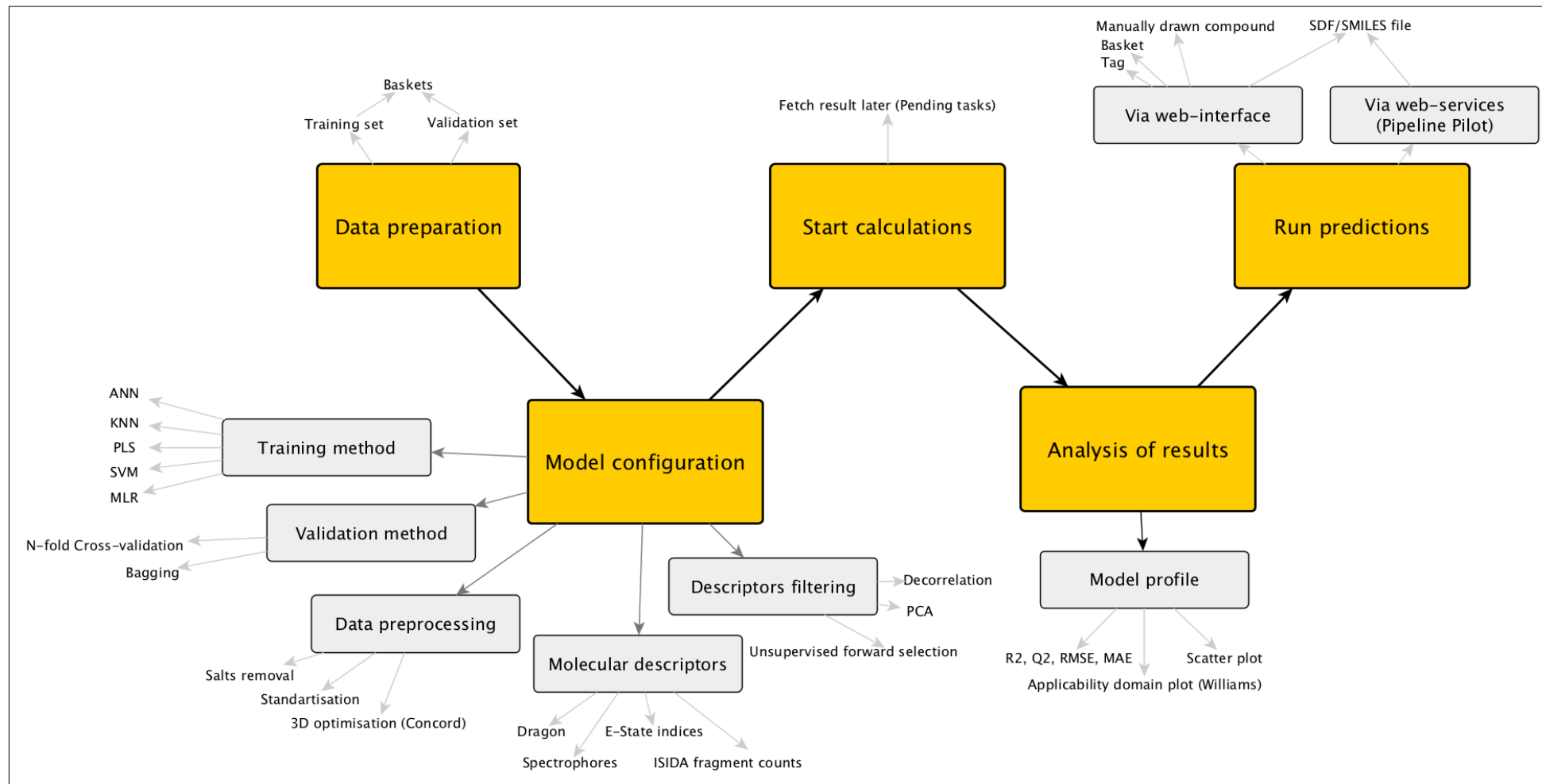
Congratulations! Now you have training sets and a test sets ready to build QSPR models.

3. Modelling framework

In this chapter the modelling framework of OCHEM is introduced:

- 3.1 Overview
- 3.2 Select the training sets, method and data pre-processing options
- 3.3 Configure molecular descriptors
- 3.4 Configure the training method and start calculations
- 3.5 Wait for the calculations to finish
- 3.6 Save your model
- 3.7 The model profile: review your model
- 3.8 Applicability domain and model export
- 3.9 Model application

3.1 Overview



The basic steps of a QSAR modeling lifecycle: prepare data, configure model, train the model, analyse results and use the model to predict new compounds

3.2 Select the training sets, method and data pre-processing options

To start the model creation process, please open “Model > Create a model” menu.

Create a model
Select the training and validation sets, the machine learning method and the validation protocol

1 **Select the training and validation sets:**
Training set (required): [...]
Validation set (optional): [...]

Choose the learning method:

- ☒ ASNN (ASsociative Neural Networks) [W]
- ☐ Consensus model (experimental) [W]
- ☐ FSMLR (Fast Stagewise Multiple Linear Regression) [W]
- ☐ KNN (K-Nearest Neighbors) [W]
- ☐ KRR (Kernel Ridge Regression) [W]
- ☐ LibSVM wrapper with grid-search parameter optimisation [W]
- ☐ LogP-LIBRARY [W]
- ☐ MLR (Multiple Linear Regression) [W]
- ☐ Wolfram's Neural Gas Network (beta) [W]
- ☐ PLS (Partial Least Square) [W]
- ☐ R-RF Supervised (experimental) (R-based Random Forest supervised) [W]
- ☐ WEKA-J48 (Weka-based implementation of C4.5 decision tree) [W]
- ☐ WEKA-RF (Weka-based implementation of Random Forest) [W]

Model validation
Validation method:
Number of folds:
☐ Stratified cross-validation

You can create a model from template: [import an XML model template](#) or [use another model as a template](#)

3

1. Select the training and validation sets that you have prepared before by clicking on the [...] labels
2. We will use defaults for most of the configurable options. Thus, we will select neural networks (ASNN) to train the model and 5-fold cross-validation to validate it.
3. The model creation process is organized as a “wizard”. Click “Next” to navigate forward.

Pre-processing of the molecules includes three options: standardization of some chemical groups for consistency, neutralization of ions and removal of salts.

We will use the default recommended configuration and employ all three pre-processing options.

Model editor
Select model template and training set

Select the preferred data preprocessing options

Preprocessing of molecules (Chemaxon)

- ☒ Standardization [W]
- ☒ Neutralize [W]
- ☒ Remove salts [W]

3.3 Configure molecular descriptors

Model editor

Select model template and training set

Select the molecular descriptors:

☒ E-state [W](#)

E-State types:

☒ Atom indices

☒ Bonds indices

☐ Atom counts

☐ Bonds counts

Aromatize structures:

☐ OESate [W](#)

☒ ALogPS (2) [W](#)

☐ MolPrint [W](#)

☐ Dragon v. 5.5 (3190/3D) [W](#)

☐ MOPAC descriptors (21/3D) [W](#)

☐ ADRIANA.Code (211/3D) [W](#)

☐ CDK descriptors (246/3D) [W](#)

☐ QNPR [W](#)

☐ 'Inductive' descriptors (54/3D) [W](#)

☐ Chemaxon descriptors (499/3D) [W](#)

☐ Spectrophores (144/3D) [W](#)

☐ Experimental values of other properties [W](#)

Outputs of other models [W](#)

[\[Add a model\]](#)

[<<Back](#) [Next>>](#)

Selection of molecular descriptors is an important step that can significantly contribute to the quality of the model.

For this tutorial, we will use the default selection – E-State descriptors and ALogPS.

Model editor

Select model template and training set

Select filters of descriptors [W](#)

☒ Eliminate descriptors with less than unique values

☒ Delete descriptors that have absolute values larger than

☒ Delete descriptors that have variance smaller than

☒ Group descriptors, that have pair-wise correlations Pearson's correlation coefficient R larger than

☐ Use Unsupervised Forward Selection to delete variables using the above value of multiple correlation coefficient R

☐ Perform principal component analysis

☐ After filtering, I want to select necessary descriptors myself (advanced)

[<<Back](#) [Next>>](#)

The next dialog allows filtering out redundant and correlated descriptors.

Again, for the purpose of this tutorial, we will use the default values, which include simple filters like pairwise decorrelation.

3.4 Configure the training method and start calculations

Model editor
Select model template and training set

Configure ANN method

Training method: SuperSAB
 Number of neurons in hidden layer: 3
 Learning iterations (learning iterations): 1000
 Ensemble: 64
 Disable ASNN: ☐
 Additional Parameters (separated by comma):

<<Back Next>>

Each machine learning method (e.g., neural networks in our case, KNN, MLR, PLS, etc.) may require additional configuration options. In this case, we can configure the training algorithm, the number of neurons, learning iterations and the number of networks in the ensemble.

We will not experiment here now and will use the default options.

Model editor
Select model template and training set

Start calculation of the model

Now we are ready to start calculation.

Please provide the name for your model: TutorialModel **1**

Task priority:

- ☐ High priority (please, use for fast tasks only)
☒ Normal priority
☐ Low priority

<<Back Start calculation>> Discard **2**

Finally, the entire necessary configuration has been performed and we are ready to start calculations. The only thing that we *must* provide here is the name of the model **(1)**.

Specifying the priority of the calculations is optional and defaults to “normal”.

Please press “start calculations” **(2)**.

3.5 Wait for the calculations to finish

Model editor
Select model template and training set

Run model builder

Running the teacher - Waiting for a free server
[cancel] [fetch result later]

1

<<Back Next>>

You are forwarded to a waiting-screen that shows you the status of the calculations. The training process can take a while to complete.

Although we could have waited, we will opt to click “fetch result later” **(1)**.

Pending tasks

The overview of all running tasks and all completed tasks awaiting your action

All tasks types All tasks statuses [Refresh] ☐ Refresh every minute

1 - 1 of 1

Task type / Time started	Model	Property / Set	Method	Status	Priority	Details
Model training 2012-05-03 19:55:59	TutorialModel	RegressionModelProperty Tutorial (training)	ANN	assigned	normal	Task started terminate

1 - 1 of 1

The next screen is the list of currently pending tasks, also accessible from menu “Model > View pending tasks”. This list displays all the tasks that are currently running or finished, but not yet fetched by the user.

Here you can observe, terminate running tasks or fetch the ready tasks. Please, click “refresh” **(2)** to actualize the page.

When the task is finished, please click the green checkbox button to fetch the model **(3)**.

Pending tasks

The overview of all running tasks and all completed tasks awaiting your action

All tasks types All tasks statuses [Refresh] ☐ Refresh every minute

1 - 1 of 1

Task type / Time started	Model	Property / Set	Method	Status	Priority	Details
Model training 2012-05-03 19:55:59	TutorialModel	RegressionModelProperty Tutorial (training)	ANN	ready	normal	recalculate

1 - 1 of 1

3

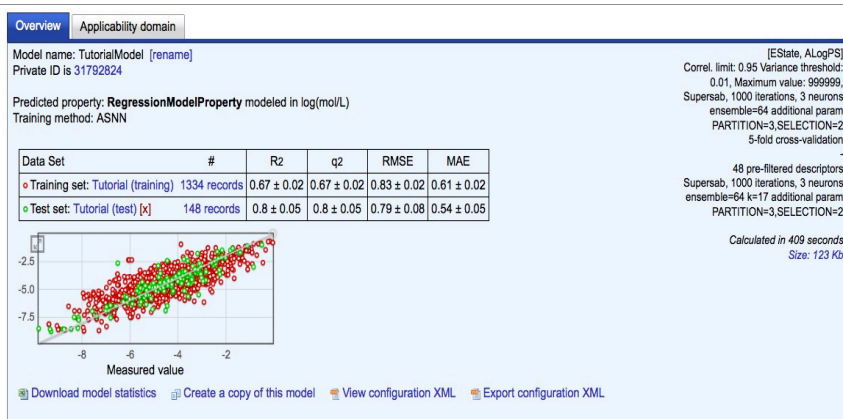
3.6 Save your model

Model editor

Select model template and training set

Save the model

Please enter your model's name:



If the calculation was successful, you can now see the profile of the ready model.

We will explain this important dialog to that later in more detail. For now please save your model **(1)**.

1 Save Discard

Models applier browser

Select a model from the list

Model name or model ID: and property name: or by article id: Models visibility:

Order by: [refresh]

1 - 15 of 18

15 items on page 1 of 2 >>

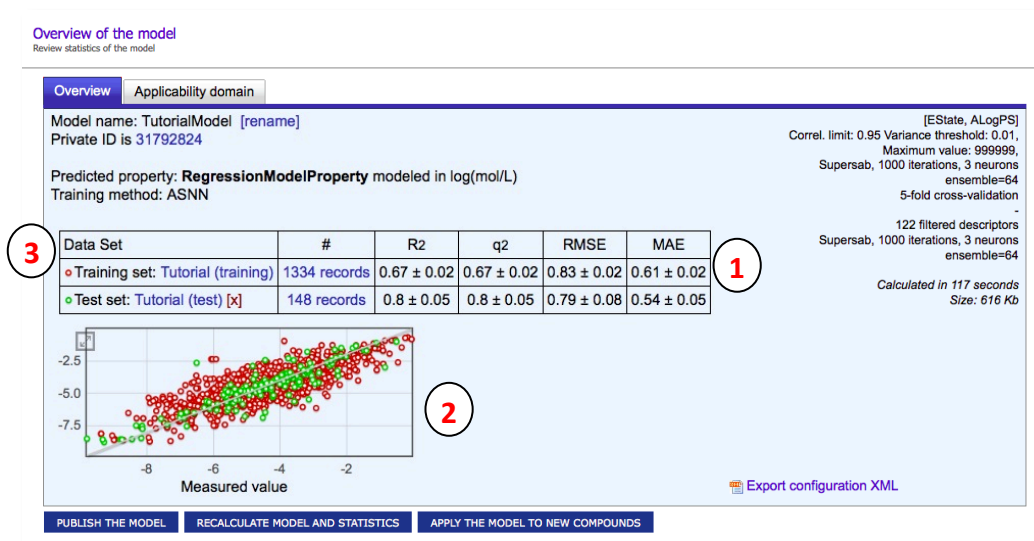
TutorialModel **2** predicts RegressionModelProperty using Tutorial (training) (1186)
validated by Tutorial (test) (296) 2012-05-03

Now that your model is saved, you can see it in the model list available at "Model > Apply a model".

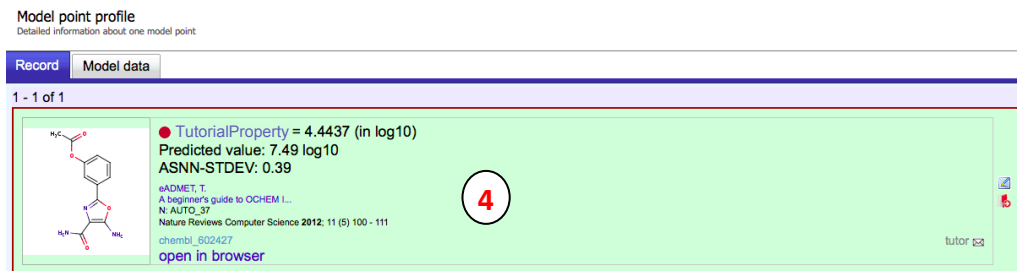
Here you can see and search through all your saved models (and the models published by other users).

To open the profile of your model, please click the model name **(2)**.

3.7 The model profile: review your model



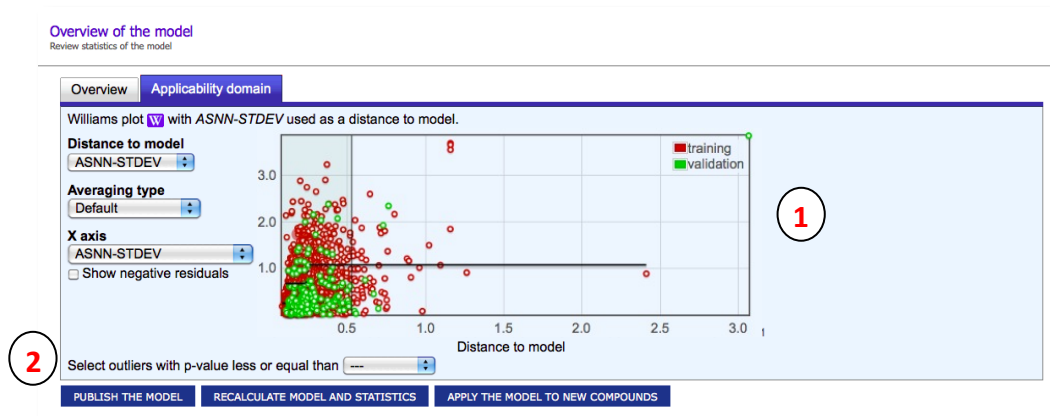
The model profile is an important dialog that contains all the information related to the performance of the model: the statistical parameters **(1)**, the scatter plot **(2)**, links to the sets **(3)** and various operations, like export of the model, application of the model to new compounds, etc.



Each point on the scatter plot **(2)** is clickable and will open the “model point profile” **(4)** containing the details of the respective compound from the training or the validation set.

This is a powerful feature that allows you to investigate outliers “under microscope”. What are the prediction values, molecular descriptor values, the respective publication, the user who introduced this record? You can track this individually for each compound.

3.8 Applicability domain and model export



The “applicability domain” tab shows the dependency of the prediction accuracy from the “distance to model” concept. This plot (1) is also referred to as “Williams plot”.

Similarly to the scatter plot, each point can be clicked and tracked back to the original compound. Moreover, outliers can be automatically selected based on p-value (2).

Data export

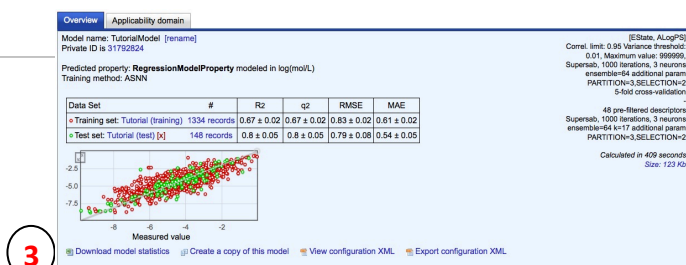
Export the selected data as an Excel, CSV or SDF file

Please, select the items that you want to export:

[select all] [select none]

- ☐ Structure (SMILES or SDF)
- ☐ CASRN
- ☒ RECORDID
- ☒ MOLECULEID
- ☒ Identifier in article (N)
- ☐ NAMES
- ☒ Introducers of the records
- ☒ Last modifiers of the records
- ☐ Publication IDs
- ☒ Error messages
- ☒ Predicted values
- ☒ Experimentally measured values
- ☒ DM (distance to model) values
- ☐ Conditions of experiments
- ☐ DESCRIPTORS
- ☒ External unique identifier
- ☒ Comments
- ☐ Inchi-key

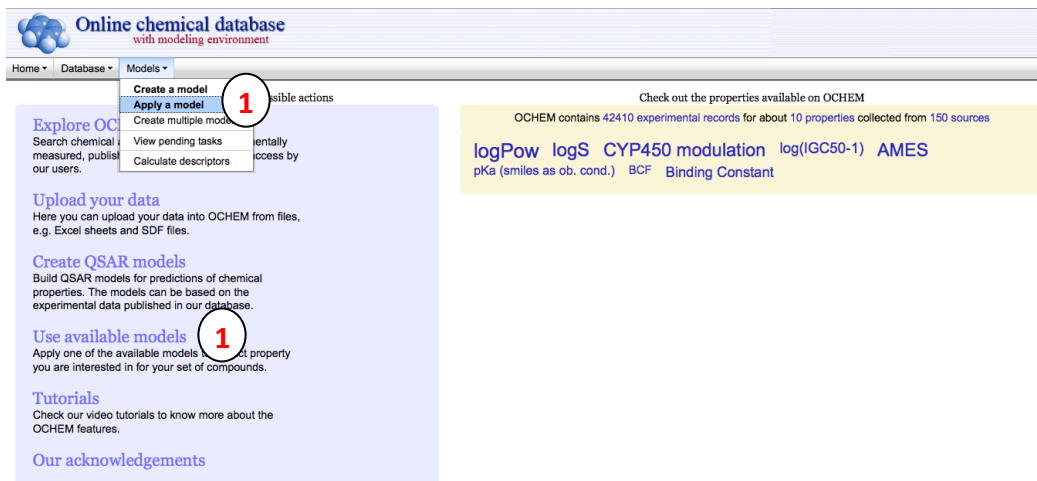
Get Excel file Get CSV file Get SDF file Get R script



It is possible to export the data related to your model by clicking “Download model statistics in Excel format” (3). The appearing dialog allows you to select detailed info for the training and validation set – the molecular structures, identifiers, predicted and measured values, prediction accuracies, etc.

You can export this data in Excel, CSV, SDF or R formats. For this tutorial, please try to export an Excel file (4).



3.9 Model application

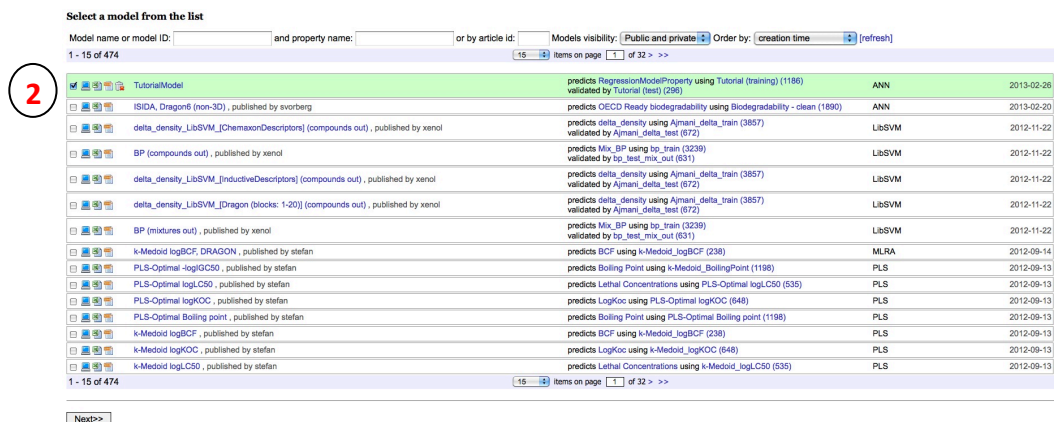

















Getting to model application

1. To apply a model, there are two ways to get there. First directly from the home page (link "Use available models") or from the menu bar (link "Apply a model"). Both links lead you to the Model applier browser.

Model applier browser

- In the model applier browser all models are listed that are available for the user. On a first glance at a particular entry, the user can see the model name, the predicted properties, used training set, used machine learning method and the creation date.
- Clicking on the computer icon  or the model name shows the model profile plot and statistics.
- The  icon links to model export.



Model name or model ID	and property name	or by article id	Models visibility	Order by	creation time	[refresh]
1 - 15 of 474			Public and private	creation time		
 TutorialModel	predicts RegressionModelProperty using Tutorial (training) (1186)	validated by Tutorial (test) (296)	ANN	2013-02-26		
 ISIDA, Dragon6 (non-3D), published by svorberg	predicts OECD Ready biodegradability using Biodegradability - clean (1690)	ANN	2013-02-20			
 delta_density_LibSVM_ChemaxonDescriptors (compounds out), published by xenol	predicts delta_density using Ajmani_delta_train (3857)	validated by Ajmani_delta_test (872)	LibSVM	2012-11-22		
 BP (compounds out), published by xenol	predicts Mix_BP using bp_train (3239)	validated by bp_test_mix_out (831)	LibSVM	2012-11-22		
 delta_density_LibSVM_InductiveDescriptors (compounds out), published by xenol	predicts delta_density using Ajmani_delta_train (3857)	validated by Ajmani_delta_test (872)	LibSVM	2012-11-22		
 delta_density_LibSVM_Dragon (blocks: 1-20) (compounds out), published by xenol	predicts delta_density using Ajmani_delta_train (3857)	validated by Ajmani_delta_test (872)	LibSVM	2012-11-22		
 BP (mixtures out), published by xenol	predicts Mix_BP using bp_train (3239)	validated by bp_test_mix_out (831)	LibSVM	2012-11-22		
 k-Medoid logBCF, DRAGON, published by stefan	predicts BCF using k-Medoid_logBCF (238)	MLRA	2012-09-14			
 PLS-Optimal -logIC50, published by stefan	predicts Boiling Point using k-Medoid_BoilingPoint (1198)	PLS	2012-09-13			
 PLS-Optimal logIC50, published by stefan	predicts Lethal Concentrations using PLS-Optimal logIC50 (535)	PLS	2012-09-13			
 PLS-Optimal logKOC, published by stefan	predicts LogKoc using PLS-Optimal logKOC (648)	PLS	2012-09-13			
 PLS-Optimal Boiling point, published by stefan	predicts Boiling Point using PLS-Optimal Boiling point (1198)	PLS	2012-09-13			
 k-Medoid logBCF, published by stefan	predicts BCF using k-Medoid_logBCF (238)	PLS	2012-09-13			
 k-Medoid logKOC, published by stefan	predicts LogKoc using k-Medoid_logKOC (648)	PLS	2012-09-13			
 k-Medoid logIC50, published by stefan	predicts Lethal Concentrations using k-Medoid_logIC50 (535)	PLS	2012-09-13			
1 - 15 of 474						

2. Next step is to select the model you want to apply, so check the box and click next button (in the down)

- Note:** Multiple selections of models are supported; also regression and classification models can be mixed. Multitask models give several predictions by default.

Apply the model

No compounds selected

Provide the compound(s)

Please provide compounds for which you want to predict the target property
Several options are available:

3 ☒ Choose a previously prepared set: [Selected records](#)

☐ Select molecules by a tag: [...]

☐ Upload compounds from a file
(SDF/MOL2/SMILES/Excel sheet)

Durchsuchen...

☐ Provide a Name/CAS-RN/SMILES

c1ccccc1N

☐ Draw Molecule

(click on depiction to the right to draw)



Prediction scenario: Use predictions only

4 Next>>

Application of regression model:

3. Now a target has to be selected. There are several possibilities (3):

- ☐ Using an earlier created basket
- ☐ Select a certain set of records / molecules by a tag
- ☐ Upload structures (by known file formats)
- ☐ Provide smiles for a single molecule
- ☐ Draw a structure

☐ Select the prediction scenario

4. Click on next button to start the application

Wait until the application is done or click on fetch results later to get to the pending tasks browser

5. When the model is finished, results are shown

6. There is the prediction value itself, distance to the model and an estimation of the accuracy (RMSE)

7. Predicted results can be exported to Excel or CSV sheet, SDF files or R scripts. Therefore different properties can be chosen.

Furthermore the applicability domain can be shown, as well as the predicted basket can be integrated into the model as test set. To integrate the predicted basket it must contain the same property as the model was built on. Then the link to the model statistics is available ([Add the results as a validation set for model ...](#)).

Models applicer browser

Prediction results

[Export results in a file \(Excel, CSV or SDF\)](#)

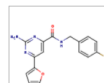
[Add the results as a validation set for model Binding Constant, 1000000006](#)

Show applicability domain >>

Sorting: none

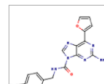
1 - 15 of 100

15 Items on page 1



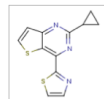
Binding Constant (Binding Constant, 1000000006) = $8.44 \log_{10} \pm 1.38$ (ASNN-STDEV = 0.14, estimate Binding Constant(measured) = 7.99897)

6

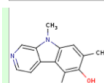


Binding Constant (Binding Constant, 1000000006) = $8.7 \log_{10} \pm 1.39$ (ASNN-STDEV = 0.23, estimate Binding Constant(measured) = 9.0)

5



Binding Constant (Binding Constant, 1000000006) = $7.51 \log_{10} \pm 1.78$ (ASNN-STDEV = 0.45, estimated RMSE = 0.91) Binding Constant(measured) = 8.48149



Binding Constant (Binding Constant, 1000000006) = $5.35 \log_{10} \pm 1.78$ (ASNN-STDEV = 0.34, estimated RMSE = 0.91) Binding Constant(measured) = 4.09367

Models applicer browser

Prediction results

[Export results in a file \(Excel, CSV or SDF\)](#)

[Add the results as a validation set for model Binding Constant, 1000000006](#)

Show applicability domain >>

Sorting: none

1 - 15 of 100

15 Items on page 1

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

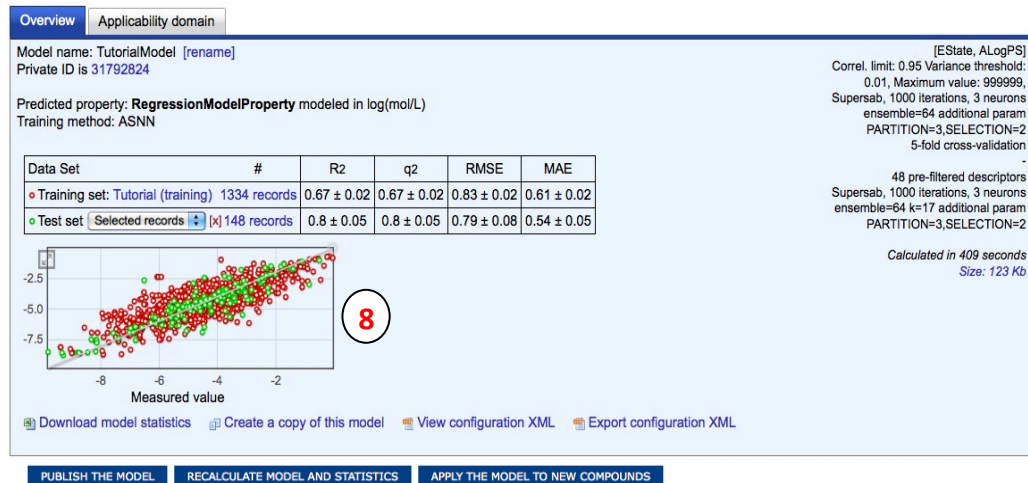
1 - 15 of 100

1 - 15 of 100

1 - 15 of 100

Overview of the model

Review statistics of the model



8. Predicted basket as test set

The predicted values are shown in the statistics plot

- **Note:**
Another way to start the application of a model is directly from the model profile.

Models applier browser

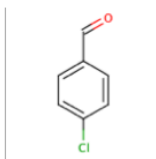
Prediction results

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

A Show applicability domain >>

Sorting none

1 - 1 of 1



Binding Constant (qualitative) (Binding constant qualitative model) = low (61.0% accuracy)

1 - 1 of 1

[<<Back](#)

Application of classification models

For application of classification models, the same steps have to be done as for a regression model.

9. The result browser shows the predicted class together with an estimation of the accuracy. In this example case the model was applied to a single drawn structure.

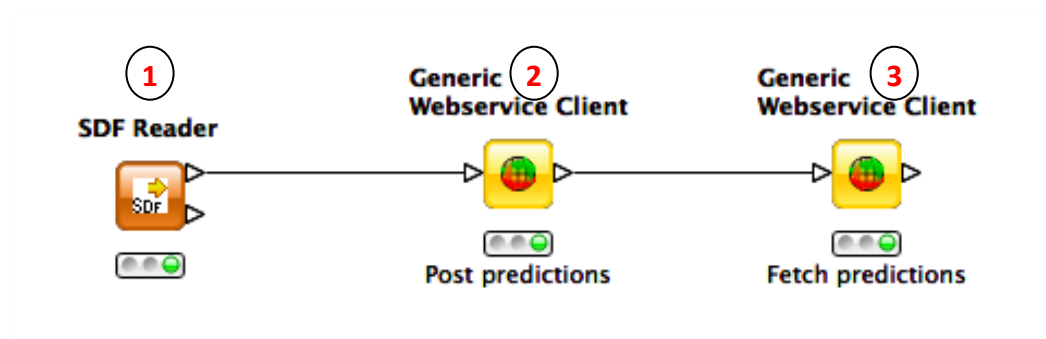
[illegible]

4. Advanced features

In this chapter the advanced features of OCHEM are introduced. In particular these are:

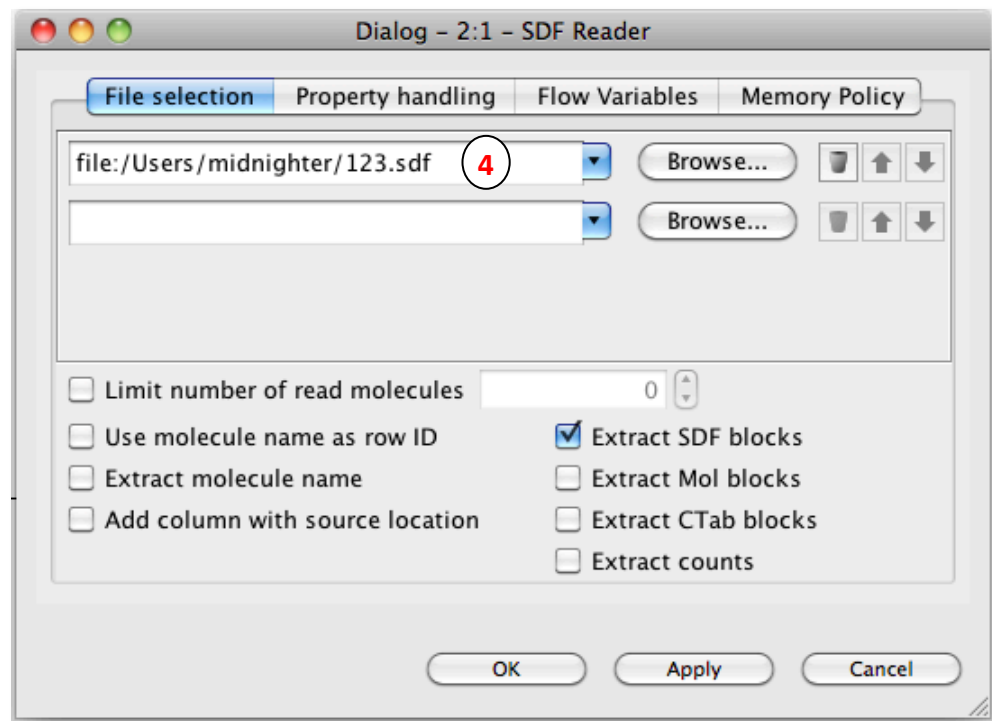
- 4.1 Using OCHEM via web-services and KNIME
- 4.2 Comprehensive modelling
- 4.3 ToxAlert utility
- 4.4 Set Compare utility
- 4.5 Pathway Analysis

4.1 Using OCHEM via web-services and KNIME



OCHEM exposes a lot of its functionality via SOAP web-services, which makes it possible to integrate OCHEM features into different workflow utilities such as KNIME or pipeline pilot.

This exemplary workflow shows how to run OCHEM predictions via KNIME. The workflow contains of 3 simple nodes - SDF file reader (1), a web-service to post predictions (2) and a web-service to fetch predictions when the tasks are ready (3).



First, select the file with molecules that you want to predict (4) in the configuration dialog for the SDF reader (1).

Dialog - 2:3 - Generic Webservice Client (Post predictions)

Webservice Description | Advanced | Flow Variables | Memory Policy

WSDL Location: **1**

Services:

Ports:

Operations: **2**

Input

Parameter	Type	Constant?	Constant value	Mapped column
modelId	Long	<input checked="" type="checkbox"/>	15810085 3	
sdf	String	<input type="checkbox"/>		sdf Molecule 4

Output

Parameter	Type	Include	Output Column	All in I...
return	ModelResponse			
metaserverTaskId	Long	<input checked="" type="checkbox"/>	\$ metaserverTaskId	
modelDescriptionUrl	JAXBElement			
modelDescriptionUrl	String	<input checked="" type="checkbox"/>	\$ modelDescriptionUrl	
predictions	Prediction			

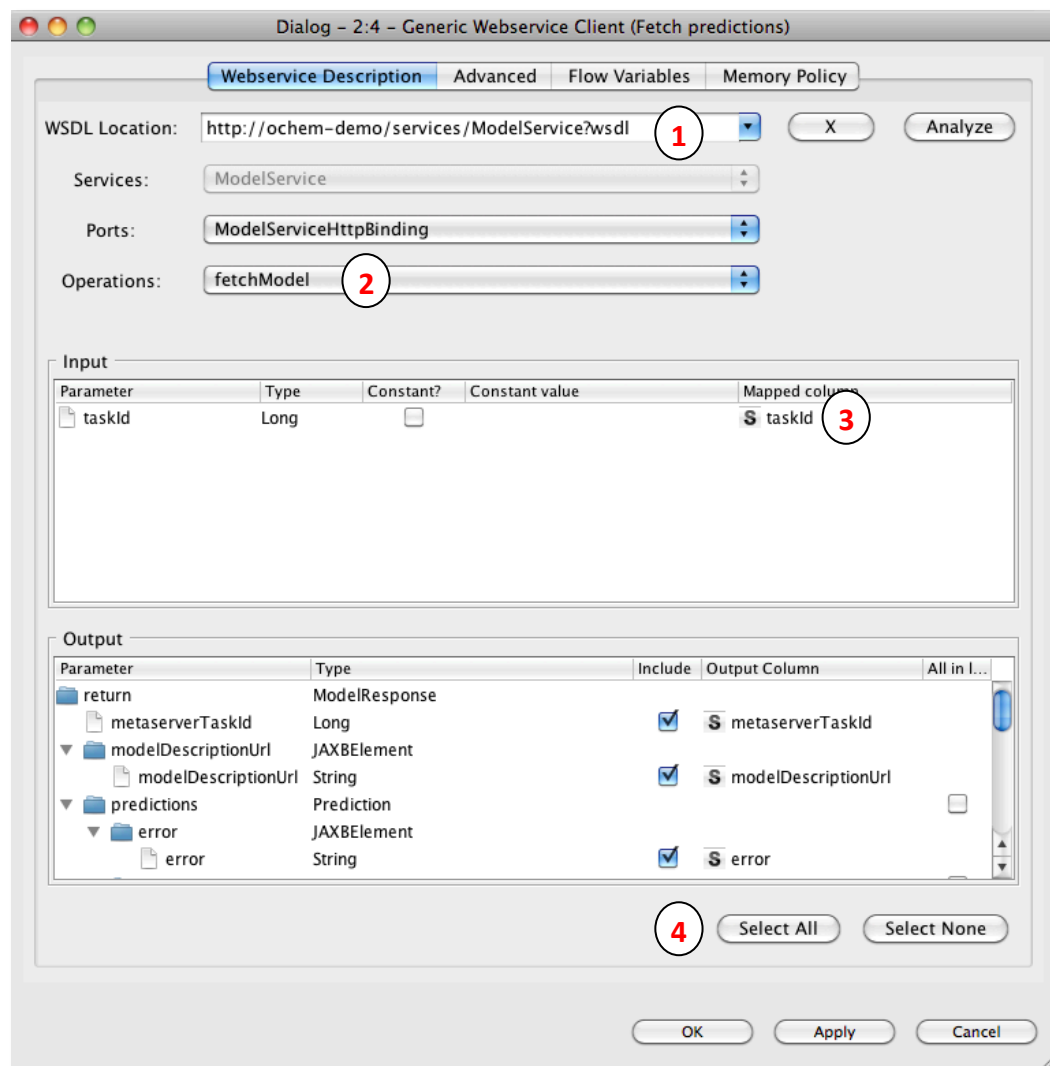
5

Then, you must configure the second web-service node that posts the predictions.

Here, provide the WSDL address of the OCHEM web-service endpoint **(1)**. This address is <http://ochem.eu/services/ModelService?wsdl>. Paste it to the "WSDL Location" field and click "Analyze".

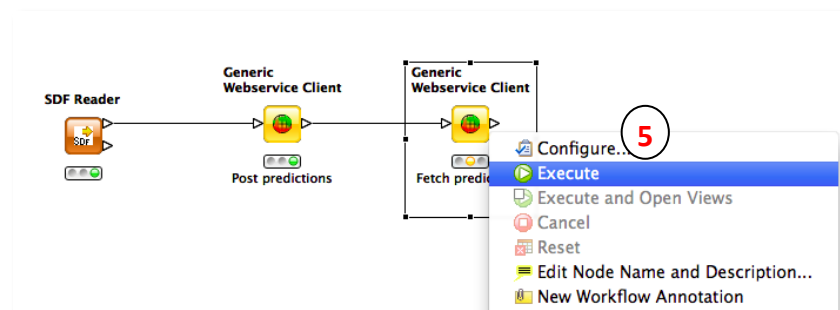
Please remember or copy this URL, since you will need it in every place where you want to access OCHEM features via web-services.

Then, select the postModelSingleSDF operation **(2)**, provide the identifier **(3)** of the model that you want to run prediction with, map the molecule column **(4)** and instruct KNIME to include all the available columns in the result **(5)**.



The configuration of the last node is similar to the previous one: you must specify the same address (1), select the „fetchModel“ operation (2), remark the taskID column (3) and instruct KNIME to include all the available columns in output (4).

After all the nodes have been configured, we are ready to start the KNIME workflow (5).



Output table - 2:4 - Generic Webservice Client (Fetch predictions)

File

Table "default" - Rows: 2 Spec - Columns: 25 Properties Flow Variables

Row ID	rl(# 1)	\$	D accura...	\$ error(...	\$ molec...	\$ property(# 1)	D realVal...	\$ u	D value...	\$ status(...	\$ taskid(...
Row0	lel/158100... ?	3	0.908	?	?	Binding Const...	0	log10	6.132	success	1
Row1	lel/158100... ?		0.709	?	?	Binding Const...	0	log10	5.568	success	0

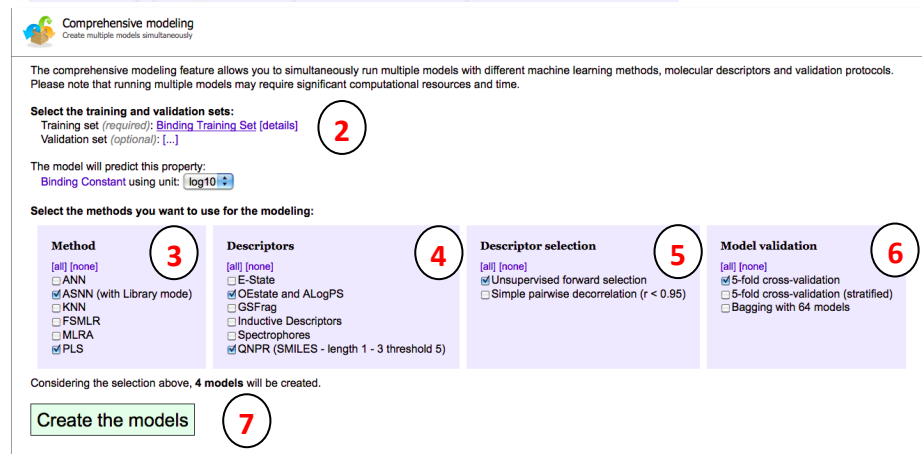
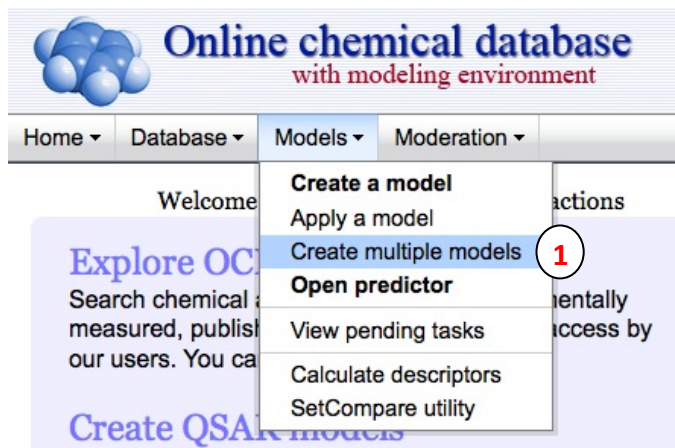
The output of the last node will show you the status of the submitted calculation tasks. You may need to re-run the node multiple times until the status **(1)** is „success“ for all the molecules.

After that, you should be able to access the prediction values **(2)** and prediction accuracies **(3)**.

Similarly, you can access the OCHEM functionality from any other tool that supports SOAP web-services:

The important thing that you always need to have at hand when using OCHEM via web-services is the V
<http://ochemeu/services/ModelService?wsdl>.

4.2 Comprehensive modelling



The “comprehensive modelling” feature accessible via menu (1) is an advanced feature that allows you to easily create multiple models based on different descriptor sets and training methods.

With this feature, you can create dozens of models simultaneously and directly compare their performance.

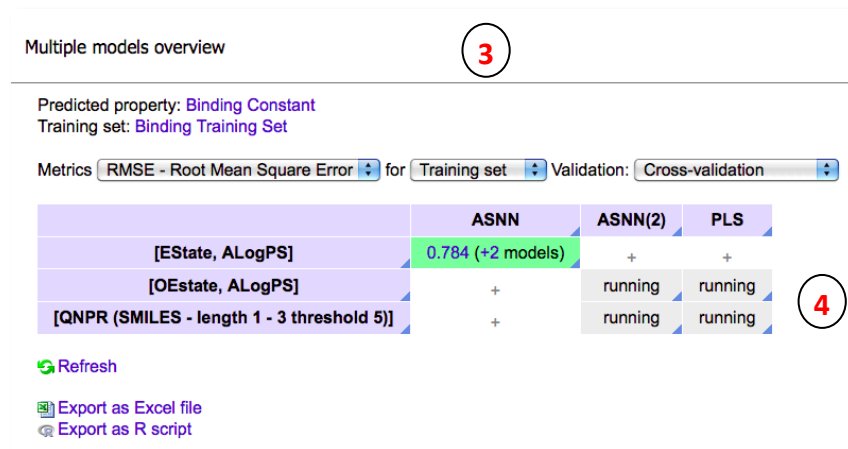
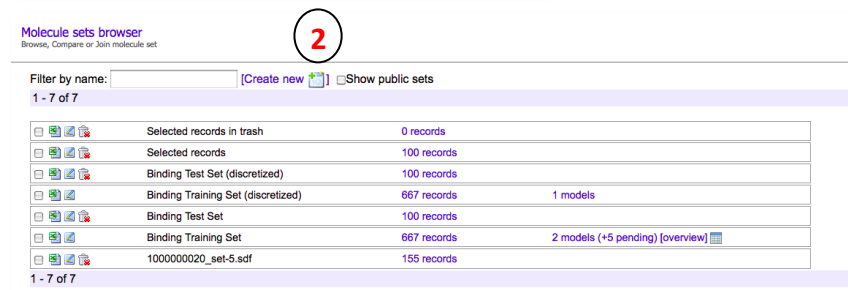
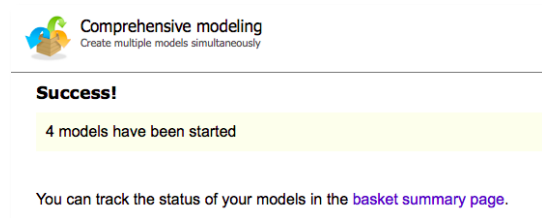
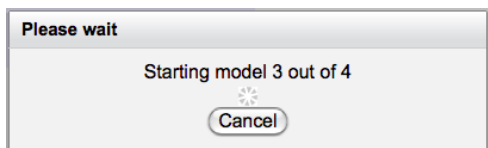
In the following dialog, at first you should select your training set (2).

You can see a set of predefined configuration templates for several training methods (3), molecular descriptors (4), descriptor selection methods (5) and model validation (6).

The checked methods will be applied using “all against all” principle. On the following screenshot, we selected 2 methods (3), 2 descriptor sets (4), one selection method and one validation method, which results into 4 models.


We selected only 4 models for speed. Normally, you can run dozens or hundreds of models, depending on available calculation resources.

Now we are ready to launch all the four models (7).



Please, wait until OCHEM starts the necessary calculation tasks.

When done, you are forwarded to the success page, from where you can directly go to the **models summary page (1)**.

The models summary page built for a particular basket is also available via Basket browser (menu Database > Baskets), by clicking  icon for your basket **(2)**.

The models summary page shows all the models (ready and pending) for the selected basket **(3)**.

The models are grouped by methods, descriptors and validation protocols. Currently, we see that our four models are still running **(4)**.

You can return to this page at any time to check the status of your models or click "refresh" to update the dialog. Normally, creation of multiple models takes a while.

Multiple models overview

Predicted property: Binding Constant
Training set: Binding Training Set

Metrics: RMSE - Root Mean Square Error for Training set Validation: Cross-validation

	ASNN	ASNN(2)	PLS
[EState, ALogPS]	0.784 (+2 models)	+	+
[OEstimate, ALogPS]	+	ready	ready
[QNPR (SMILES - length 1 - 3 threshold 5)]	+	ready	ready

Refresh Fetch statistics for 4 ready task(s)

Export as Excel file
Export as R script

1

Multiple models overview

Predicted property: Binding Constant
Training set: Binding Training Set

Metrics: RMSE - Root Mean Square Error for Training set Validation: Cross-validation

	ASNN	ASNN(2)	PLS
[EState, ALogPS]	0.784 (+2 models)	+	+
[OEstimate, ALogPS]	+	0.782	0.847
[QNPR (SMILES - length 1 - 3 threshold 5)]	+	0.779	0.837

Refresh

Export as Excel file
Export as R script

Metrics: R2 for Training set Validation: Cross-validation

RMSE - Root Mean Square Error
MAE - Mean Absolute Error
R2
Q2
Model size

	ASNN	ASNN(2)	PLS
[EState, ALogPS]	0.55 (+2 models)	+	+
[OEstimate, ALogPS]	+	0.55	0.47
[QNPR (SMILES - length 1 - 3 threshold 5)]	+	0.56	0.48

2

	ASNN	ASNN(2)	PLS
[EState, ALogPS]	0.55 (+2 models)	+	+
[OEstimate, ALogPS]	+	0.55	0.47
[QNPR (SMILES - length 1 - 3 threshold 5)]	+	0.56	0.48

3

To calculate statistics for all the completed models, press the “fetch statistics for ready models” (1).

We can see all four our models ready. The numbers in the cells (“metrics”) show the root mean square error.

In this particular case, we can immediately observe that neural network models (ASNN) have lower errors (0.782 and 0.779) than PLS models.

It is also possible to display other statistical parameters, such as R2 or Q2, using the drop-down box (2).

You can perform row-wise or column-wise batch operations, e.g., delete the models or create new models.

You also can create new models individually by pushing “+” sign in the “missing” cells (3).

Comprehensive modelling can be a very powerful feature. Which descriptors are the best? How do the models evolve when outliers are excluded? Which training method performs best for this property? All these questions require deep analysis made possible using comprehensive modelling.

The screenshot below shows an intermediate result of a real on-going study – prediction of **melting point** based on more than **30,000** experimental measurements. More than **150 models** have been built. Using the comprehensive modelling feature, it was possible to identify the best methods and to gradually improve models by removing outliers and reducing noise.

Multiple models overview

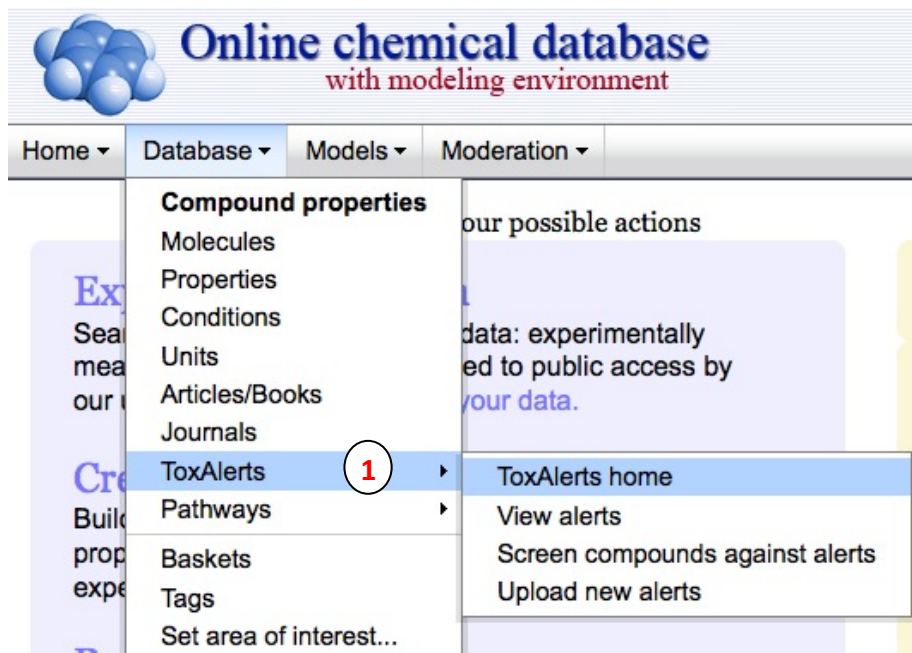
Predicted property: **Melting Point**

Training set: **MP clean**

Metrics: **RMSE - Root Mean Square Error** for Training set Validation: **Cross-validation**

	ANN	ASNN	KNN	LibSVM	FSMLR	MLRA	PLS	ASNN(2)	LibSVM(2)	ASNN(3)	ASNN(4)	ASNN(5)	ASNN(6)
CDK	46.6	41.3	51.9	57.8	57.7	56.6	60.1	39.3	56.9	39.4	+	38.3	39.7
Dragon6 (blocks: 1-29)	42.7	40.3	54.6	78.7	51.7	49.0	64.9	38.1	78.5	37.6	37.2	38.1	38.1
OEstate, ALogPS	48.7	43.0	53.2	70.8	58.8	58.3	61.4	41.4	62.9	41.0	+	41.7	41.8
Fragmentor (Length 2 - 4)	46.6	42.8	60.7	65.8	58.1	56.1	+	41.2	65.1	40.7	40.6	38.6	38.7
GSFrag	56.8	51.2	59.6	70.8	69.1	68.5	74.7	51.1	69.7	49.4	+	43.2	49.9
Mera, Mersy	49.7	44.7	56.9	63.3	69.3	56.4	79.9	43.1	62.4	42.7	+	41.3	43.0
ChemaxonDescriptors (7.4)	48.2	42.7	50.7	58.5	68.1	58.3	60.1	40.8	57.6	40.5	+	38.8	41.0
InductiveDescriptors	59.6	50.2	59.5	73.1	98.8	68.8	70.4	48.7	72.5	48.3	+	47.9	48.7
Adriana	50.1	43.9	52.4	60.2	60.5	59.4	65.6	42.6	60.2	42.0	+	40.0	42.2
Spectrophores (accuracy=20 Stereospecificity=0 resolution=3.0)	72.5	68.2	71.1	77.8	78.3	77.6	78.1	67.2	77.1	67.1	+	64.8	65.8
ShapeSignatures	70.9	68.0	70.9	79.3	165.0	76.0	78.1	67.1	79.0	67.2	+	66.8	66.6
QNPR (SMILES - length 1 - 3 threshold 5)	48.7	45.0	64.3	64.8	60.2	58.3	60.6	43.3	63.9	42.7	42.7	40.4	43.0
Dragon6 (blocks: 1 28)	76.6	76.1	76.6	82.4	82.2	82.2	82.2	75.5	82.0	75.2	+	75.4	75.6
OEstate, ALogPS	+	+	+	+	55.5	+	+	+	+	+	+	38.8	38.8

4.3 ToxAlert utility



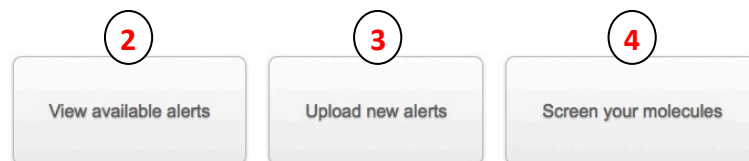
To get to the ToxAlert utility just select it from the menu bar **(1)**.

The ToxAlert utility allows one to screen a set of molecules against a set of structural alerts. Typical structural alerts might indicate carcinogenicity or general toxicity.

Welcome to ToxAlerts!

Structural alerts (also known as "*toxicophores*") are molecular patterns known to be associated with particular type of toxicity. The studies performed last decade has shown that structural alerts is an efficient technique to detect potentially toxic chemicals. Screening chemical compounds against known structural alerts can be a good practice to complement the QSAR models and to help interpreting their predictions.

ToxAlerts is a platform for screening chemical compounds against structural alerts. The platform allows to search structural alerts, introduce your own alerts and screen chemical libraries for alert-hitting compounds.

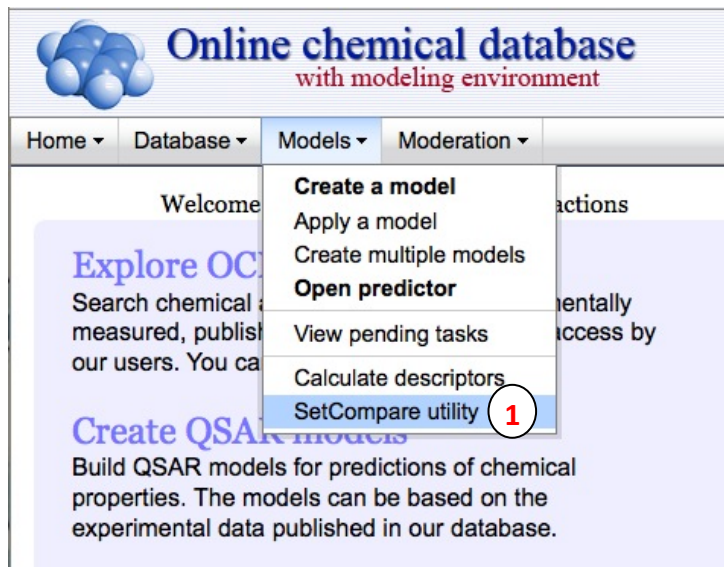


In case of any questions, ideas, or problems with the software, feel free to [drop us a message](#). We highly appreciate any feedback from you!

A welcome page is the entry point for further actions, like


- overview available alerts **(2)**
- upload new structural alerts **(3)**
- screen models against structural alerts **(4)**

4.4 Set Compare utility



The utility for set comparison can be selected from the menu bar **(1)**.

In the first page of the wizard two sets have to be selected **(2)** & **(3)**. With the set comparison utility two sets can be examined with respect to common structural alerts and common descriptors.

 **SetCompare:** Select the sets to compare

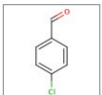
The SetCompare utility is experimental. It allows you to compare two sets of molecules based on their structural features. Please, provide the two sets available options below.

1 Select the compounds in the **first set:** **2**

Upload compounds from a file
(SDF/MOL2/SMILES/Excel sheet)

☐ Provide a Name/CAS-RN/SMILES

Draw Molecule
☒ (click on depiction to the right to draw)



☐ Choose a previously prepared set:

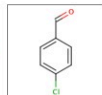
☐ Select molecules by a tag:

2 Select the compounds in the **second set:** **3**

Upload compounds from a file
(SDF/MOL2/SMILES/Excel sheet)

☐ Provide a Name/CAS-RN/SMILES

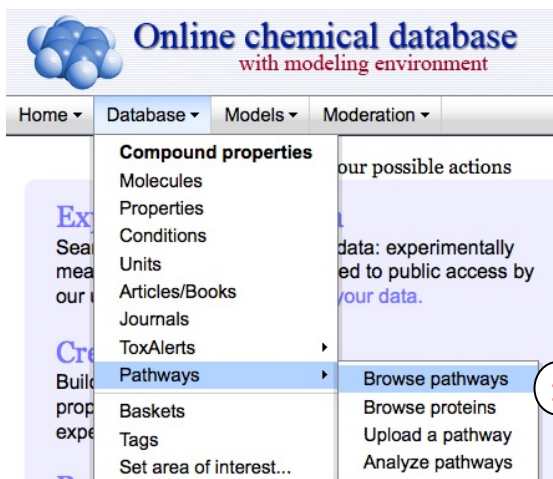
Draw Molecule
☒ (click on depiction to the right to draw)



☐ Choose a previously prepared set:

☐ Select molecules by a tag:

4.5 Pathway Analysis



A rudimentary feature for the analysis of biological pathways was implemented (1).

Pathways can be uploaded in the BioPax level 3 format via webservice or directly on the user interface.

Uploaded pathways can be overviewed (2) and are linking to their containing proteins (3).

If a protein is, e.g. a toxicological endpoint and a predictive model is available therefor, the pathway can be linked to this OCHEM property. Therefore other pathways can be analysed for this connection.