# Introduction to SAR/QSAR analysis: On-line chemical database and modelling environment (OCHEM)

Dr. I.V. Tetko, BIGCHEM GmbH and Institute of Structural Biology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH)

Public platform: www.ochem.eu
Online manual: http://docs.ochem.eu/display/MAN

HelmholtzZentrum münchen
German Research Center for Environmental Health

# OCHEM tutorial handout

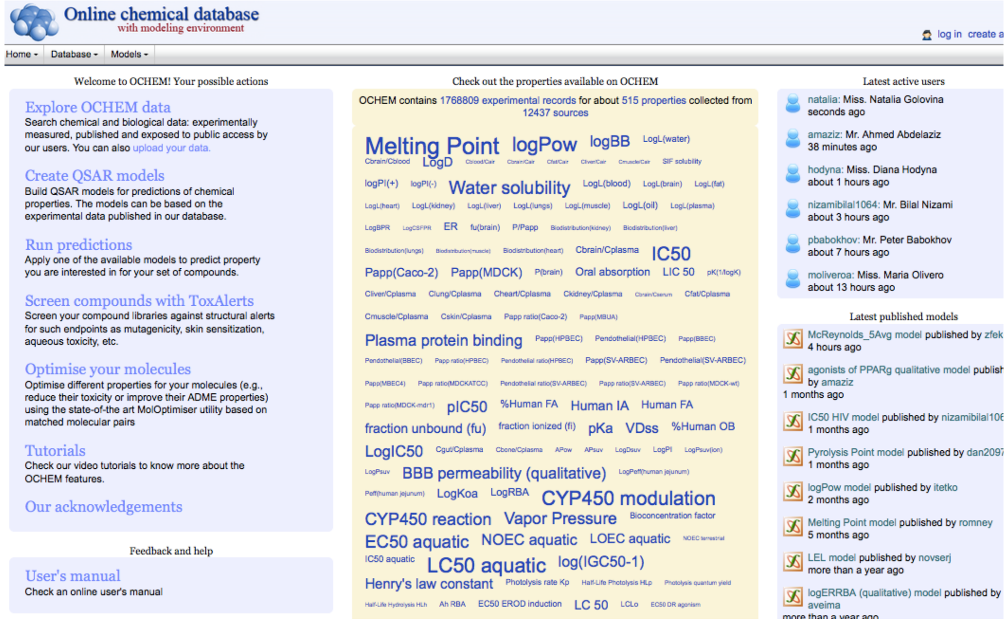
## Table of Contents

# 1. General concepts

In this chapter, we will learn the general concepts of the OCHEM interface
How (and why) to register a new account
What are the basic design components of OCHEM
What are the basic elements of OCHEM

# OCHEM tutorial handout

## 1.1 Before we start



Welcome to OCHEM! Your possible actions

**Explore OCHEM data**
Search chemical and biological data: experimentally measured, published and exposed to public access by our users. You can also upload your data.

**Create QSAR models**
Build QSAR models for predictions of chemical properties. The models can be based on the experimental data published in our database.

**Run predictions**
Apply one of the available models to predict property you are interested in for your set of compounds.

**Screen compounds with ToxAlerts**
Screen your compound libraries against structural alerts for such endpoints as mutagenicity, skin sensitization, aqueous toxicity, etc.

**Optimise your molecules**
Optimise different properties for your molecules (e.g., reduce their toxicity or improve their ADME properties) using the state-of-the art MolOptimiser utility based on matched molecular pairs

**Tutorials**
Check our video tutorials to know more about the OCHEM features.

**Our acknowledgements**

Feedback and help

**User's manual**
Check an online user's manual

Check out the properties available on OCHEM

OCHEM contains 1768809 experimental records for about 515 properties collected from 12437 sources

Melting Point logPow logBB LogL(water) Cbrain/Cblood LogD Cblood/Cair Cbrain/Cair Cfat/Cair Cliver/Cair Cmuscle/Cair SIF solubility logPI(+) logPI(-) Water solubility LogL(blood) LogL(brain) LogL(fat) LogL(heart) LogL(kidney) LogL(liver) LogL(lungs) LogL(muscle) LogL(oil) LogL(plasma) LogBPR LogCSFPR ER fu(brain) P/Papp Biodistribution(kidney) Biodistribution(liver) Biodistribution(lungs) Biodistribution(muscle) Biodistribution(heart) Cbrain/Cplasma IC50 Papp(Caco-2) Papp(MDCK) P(brain) Oral absorption LIC 50 pK(1/logK) Cliver/Cplasma Clung/Cplasma Cheart/Cplasma Ckidney/Cplasma Cbrain/Cserum Cfat/Cplasma Cmuscle/Cplasma Cskin/Cplasma Papp ratio(Caco-2) Papp(MBUA) Plasma protein binding Papp(HPBEC) Pendothelial(HPBEC) Papp(BBEC) Pendothelial(BBEC) Papp ratio(HPBEC) Pendothelial ratio(HPBEC) Papp(SV-ARBEC) Pendothelial(SV-ARBEC) Papp(MBEC4) Papp ratio(MDCKATCC) Pendothelial ratio(SV-ARBEC) Papp ratio(SV-ARBEC) Papp ratio(MDCK-wt) Papp ratio(MDCK-mdr1) pIC50 %Human FA Human IA Human FA fraction unbound (fu) fraction ionized (fi) pKa VDss %Human OB LogIC50 Cgut/Cplasma Cbone/Cplasma APow APsuv LogDsuv LogPI LogPsuv(ion) LogPsuv BBB permeability (qualitative) LogPeff(human jejunum) Peff(human jejunum) LogKoa LogRBA CYP450 modulation CYP450 reaction Vapor Pressure Bioconcentration factor EC50 aquatic NOEC aquatic LOEC aquatic NOEC terrestrial IC50 aquatic LC50 aquatic log(IGC50-1) Henry's law constant Photolysis rate Kp Half-Life Photolysis HLp Photolysis quantum yield Half-Life Hydrolysis HLh Ah RBA EC50 EROD induction LC 50 LCLo EC50 DR agonism

Latest active users

natalia: Miss. Natalia Golovina
seconds ago

amaziz: Mr. Ahmed Abdelaziz
38 minutes ago

hodyna: Miss. Diana Hodyna
about 1 hours ago

nizamibilal1064: Mr. Bilal Nizami
about 3 hours ago

pbabokhov: Mr. Peter Babokhov
about 7 hours ago

moliveroa: Miss. Maria Olivero
about 13 hours ago

Latest published models

McReynolds_5Avg model published by zfek
4 hours ago

agonists of PPARg qualitative model publish
by amaziz
1 months ago

IC50 HIV model published by nizamibilal106
1 months ago

Pyrolysis Point model published by dan2097
1 months ago

logPow model published by itetko
2 months ago

Melting Point model published by romney
5 months ago

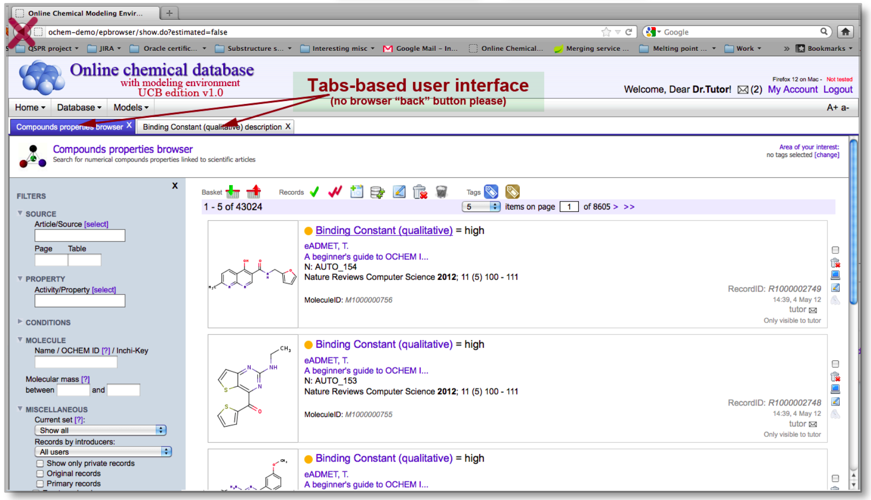
LEL model published by novserj
more than a year ago

logERRBA (qualitative) model published by
aveima
more than a year ago

The OCHEM installation is running on servers in HMGU. It can be installed on a virtual machine or on computers for better performance.

**System requirements**: For optimal performance, the host machine should have at least 8 CPU cores, 16GB RAM and about 100GB disc space.

# OCHEM tutorial handout



Tabs-based user interface

(no browser "back" button please)

Binding Constant (qualitative) = high
eADMET, T.
A beginner's guide to OCHEM I...
N: AUTO_154
Nature Reviews Computer Science **2012**; 11 (5) 100 - 111
MoleculeID: M1000000756
RecordID: R1000002749
14:39, 4 May 12
tutor
Only visible to tutor

Binding Constant (qualitative) = high
eADMET, T.
A beginner's guide to OCHEM I...
N: AUTO_153
Nature Reviews Computer Science **2012**; 11 (5) 100 - 111
MoleculeID: M1000000755
RecordID: R1000002748
14:39, 4 May 12
tutor
Only visible to tutor

Binding Constant (qualitative) = high
eADMET, T.
A beginner's guide to OCHEM I...

OCHEM is a web-based platform. Users access it with a simple web browser, similar to the way they access services like Gmail or Facebook.

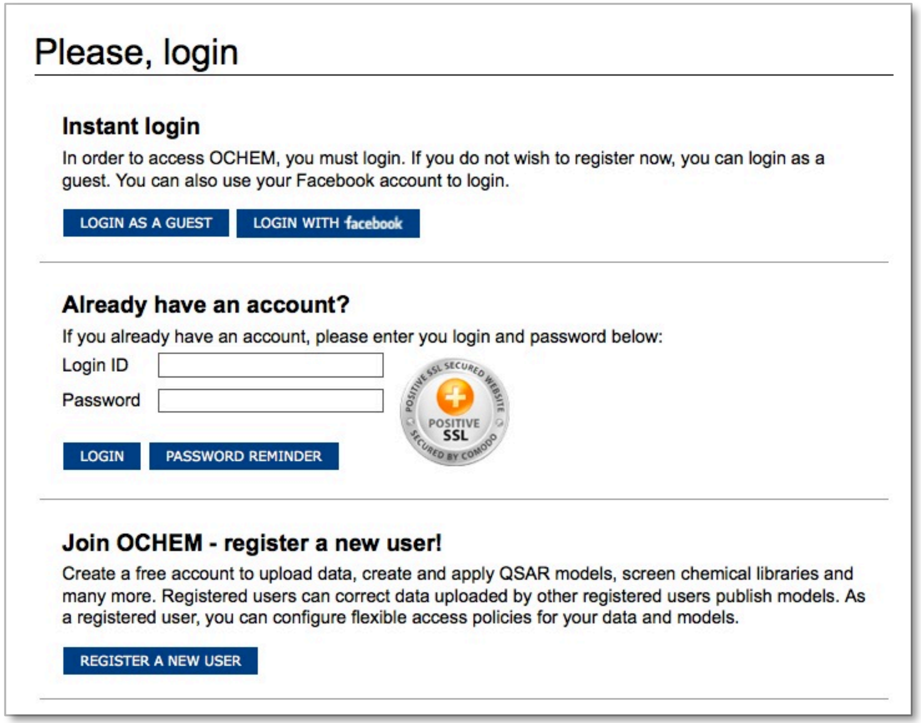
**How to access OCHEM?**
The public version of OCHEM is available at www.ochem.eu, but an in-house installation can be run inside a company/University and be accessible inside of the intranet only.

**Which browser to use?**
To get the best experience, it is recommended to use the latest Firefox browser.  Chrome and Safari are also supported. Unfortunately, at the moment OCHEM **does not fully support Internet Explorer (improves with release of new versions)** and Konqueror Web browser.

**User interface remarks.**
OCHEM uses "tabs" extensively in the user interface. New dialogues are often conveniently opened in new tabs, as it is shown on a screenshot on the left.

Please note, that the browsers "back" button is not compatible with the tabbed interface.

## 1.2 User account creation and user login



Please, login

Instant login

In order to access OCHEM, you must login. If you do not wish to register now, you can login as a guest. You can also use your Facebook account to login.

Already have an account?

If you already have an account, please enter you login and password below:

Login ID

Password

Join OCHEM - register a new user!

Create a free account to upload data, create and apply QSAR models, screen chemical libraries and many more. Registered users can correct data uploaded by other registered users publish models. As a registered user, you can configure flexible access policies for your data and models.

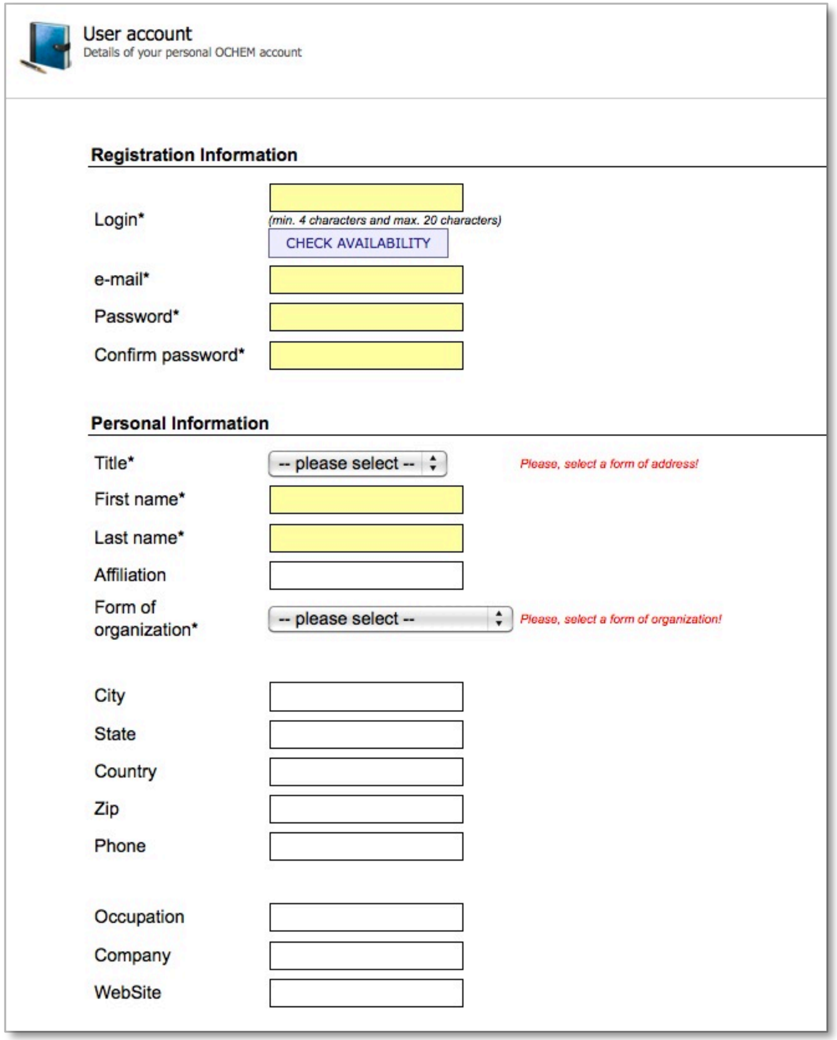
In order to use the OCHEM web platform, a user has to register an account and login using this username and password.

Users can login as a guest user (with limited privileges) or as a registered user.

1. The first option on the login interface is the instant login (without registration). You can log in as a guest user. For this tutorial, you are encouraged to register an account.

2. If you have already created an OCHEM account, you can login using your username and password.

3. The third option is to register a new account. Please, register an account if you have not done so yet.



User account

Details of your personal OCHEM account

**Registration Information**

Login* (min. 4 characters and max. 20 characters)

CHECK AVAILABILITY

e-mail*

Password*

Confirm password*

**Personal Information**

Title* -- please select -- *Please, select a form of address!*

First name*

Last name*

Affiliation

Form of organization* -- please select -- *Please, select a form of organization!*

City

State

Country

Zip

Phone

Occupation

Company

WebSite

In order to create a new OCHEM user account, the following information is required:

4. Choose your login name and check if it is available (i.e. it is not yet used by another user). Login names should be at least 4 characters long and not longer than 20 characters.

5. A valid e-mail address is required for the automatic notification system.

6. Furthermore a password for the acount should be chosen and confirmed.

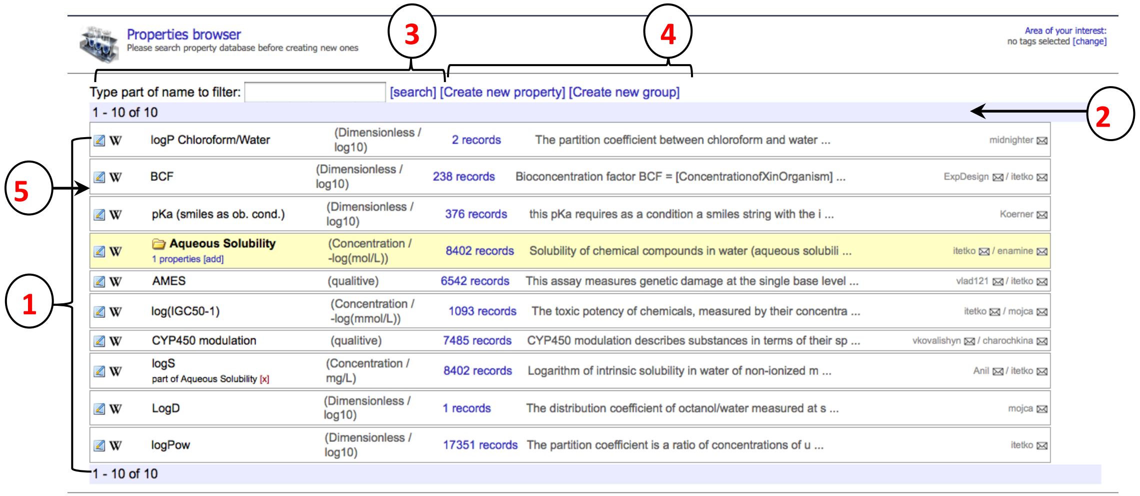
7. Additional personal information like academic title, first and last name of the user and the form of organization this person is working in is required to finish the user account creation

Note:

Registered users have access to more features than guest users (i.e., can upload data and develop models).

If users provide detailed information about themselves, their account will **be validated** by the OCHEM administrator. This will allow the users to run larger tasks, export more data and edit data of other validated users.

## 1.3 Data browsers



**Properties browser**
Please search property database before creating new ones

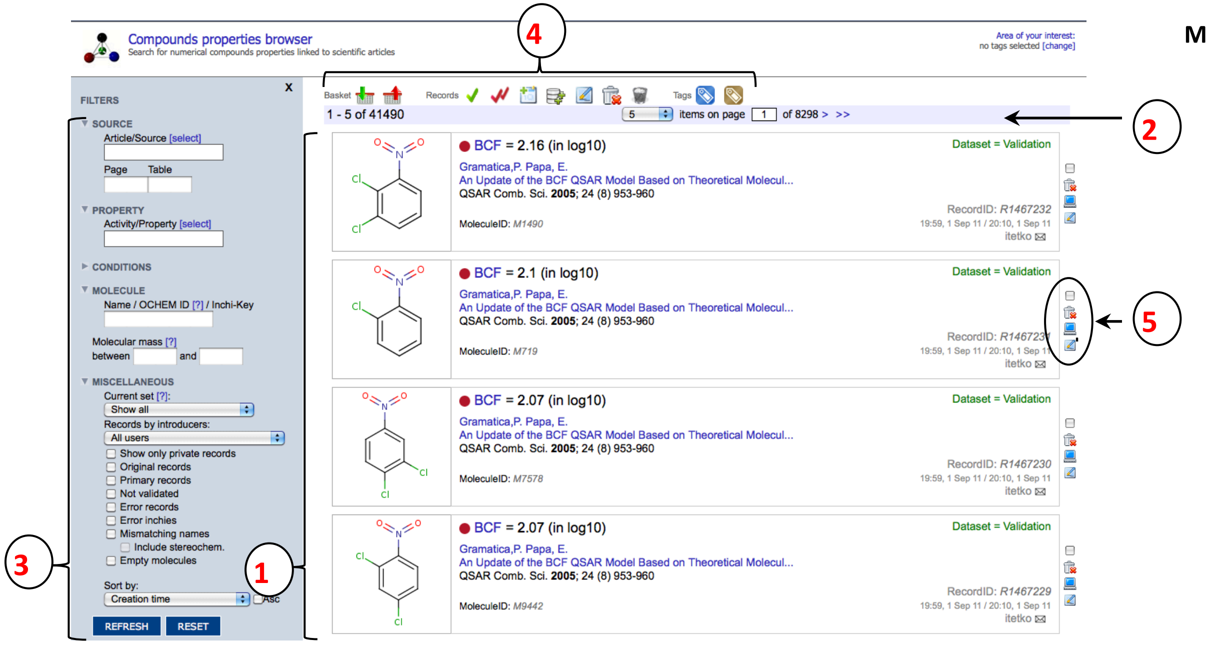
Area of your interest:
no tags selected [change]

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

1 - 10 of 10

| | Name | Unit | Records | Description | Introducers |
|---|---|---|---|---|---|
| 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 = [ConcentrationofXinOrganism] ... | ExpDesign ✉ / itetko ✉ |
| 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**<br>1 properties [add] | (Concentration / -log(mol/L)) | 8402 records | Solubility of chemical compounds in water (aqueous solubili ... | itetko ✉ / enamine ✉ |
| W | AMES | (qualitive) | 6542 records | This assay measures genetic damage at the single base level ... | vlad121 ✉ / itetko ✉ |
| W | log(IGC50-1) | (Concentration / -log(mmol/L)) | 1093 records | The toxic potency of chemicals, measured by their concentra ... | itetko ✉ / mojca ✉ |
| W | CYP450 modulation | (qualitive) | 7485 records | CYP450 modulation describes substances in terms of their sp ... | vkovalishyn ✉ / charochkina ✉ |
| W | logS<br>part of Aqueous Solubility [x] | (Concentration / mg/L) | 8402 records | Logarithm of intrinsic solubility in water of non-ionized m ... | Anil ✉ / itetko ✉ |
| 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 ... | itetko ✉ |

1 - 10 of 10



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

Area of your interest:
no tags selected [change]

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 [?]: 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

Sort by: Creation time □ Asc

1 - 5 of 41490 · 5 items on page 1 of 8298 > >>

● BCF = 2.16 (in log10) — Dataset = Validation
Gramatica,P. Papa, E.
An Update of the BCF QSAR Model Based on Theoretical Molecul...
QSAR Comb. Sci. **2005**; 24 (8) 953-960
MoleculeID: *M1490*
RecordID: *R1467232*
19:59, 1 Sep 11 / 20:10, 1 Sep 11
itetko ✉

● BCF = 2.1 (in log10) — Dataset = Validation
Gramatica,P. Papa, E.
An Update of the BCF QSAR Model Based on Theoretical Molecul...
QSAR Comb. Sci. **2005**; 24 (8) 953-960
MoleculeID: *M719*
RecordID: *R1467231*
19:59, 1 Sep 11 / 20:10, 1 Sep 11
itetko ✉

● BCF = 2.07 (in log10) — Dataset = Validation
Gramatica,P. Papa, E.
An Update of the BCF QSAR Model Based on Theoretical Molecul...
QSAR Comb. Sci. **2005**; 24 (8) 953-960
MoleculeID: *M7578*
RecordID: *R1467230*
19:59, 1 Sep 11 / 20:10, 1 Sep 11
itetko ✉

● BCF = 2.07 (in log10) — Dataset = Validation
Gramatica,P. Papa, E.
An Update of the BCF QSAR Model Based on Theoretical Molecul...
QSAR Comb. Sci. **2005**; 24 (8) 953-960
MoleculeID: *M9442*
RecordID: *R1467229*
19:59, 1 Sep 11 / 20:10, 1 Sep 11
itetko ✉

An important user interface element in OCHEM is a browser. OCHEM has various browsers for all kinds of database entities.

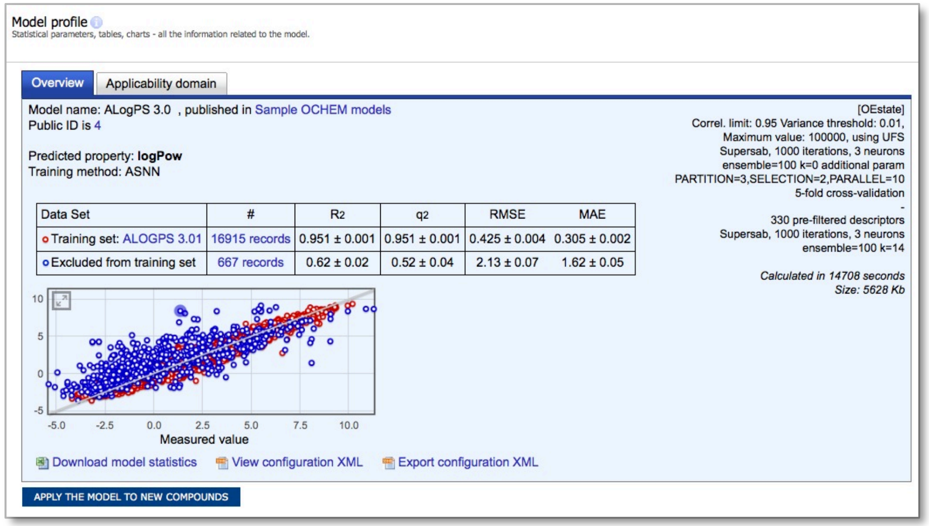
**Main browsers include:**

- Experimental property browser (record browser)
- Molecules browser
- Properties browser
- Conditions browser
- Units browser
- Articles browser
- Journals browser
- Baskets browser
- Tags browser
- Models browser

**Main elements of a browser:**

1. Items (e.g., data records, models, articles, tasks)
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

## 1.4 Item profiles



Model profile
Statistical parameters, tables, charts - all the information related to the model.

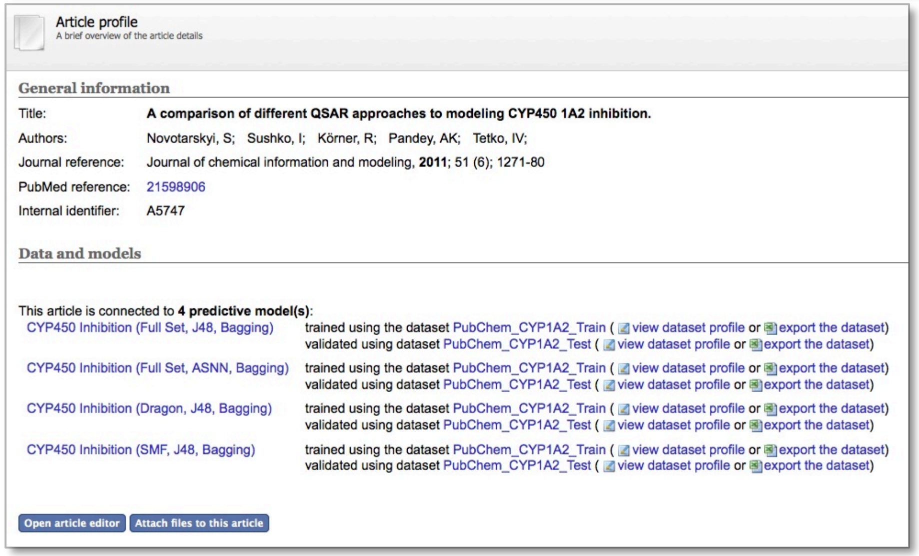
Model name: ALogPS 3.0 , published in Sample OCHEM models
Public ID is 4

Predicted property: **logPow**
Training method: ASNN

[OEstate]
Correl. limit: 0.95 Variance threshold: 0.01,
Maximum value: 100000, using UFS
Supersab, 1000 iterations, 3 neurons
ensemble=100 k=0 additional param
PARTITION=3,SELECTION=2,PARALLEL=10
5-fold cross-validation
-
330 pre-filtered descriptors
Supersab, 1000 iterations, 3 neurons
ensemble=100 k=14

*Calculated in 14708 seconds*
*Size: 5628 Kb*

| Data Set | # | R2 | q2 | RMSE | MAE |
|---|---|---|---|---|---|
| ○ Training set: ALOGPS 3.01 | 16915 records | 0.951 ± 0.001 | 0.951 ± 0.001 | 0.425 ± 0.004 | 0.305 ± 0.002 |
| ○ Excluded from training set | 667 records | 0.62 ± 0.02 | 0.52 ± 0.04 | 2.13 ± 0.07 | 1.62 ± 0.05 |



Article profile
A brief overview of the article details

**General information**

| | |
|---|---|
| Title: | **A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition.** |
| Authors: | Novotarskyi, S; Sushko, I; Körner, R; Pandey, AK; Tetko, IV; |
| Journal reference: | Journal of chemical information and modeling, **2011**; 51 (6); 1271-80 |
| PubMed reference: | 21598906 |
| Internal identifier: | A5747 |

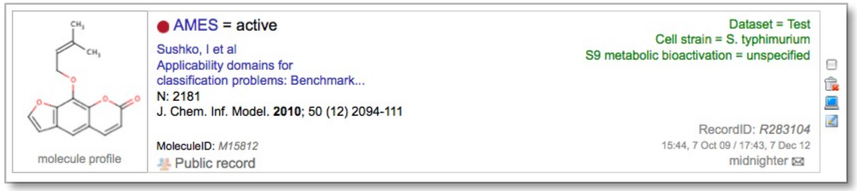
**Data and models**

This article is connected to **4 predictive model(s)**:

| | |
|---|---|
| CYP450 Inhibition (Full Set, J48, Bagging) | trained using the dataset PubChem_CYP1A2_Train ( view dataset profile or export the dataset)<br>validated using dataset PubChem_CYP1A2_Test ( view dataset profile or export the dataset) |
| CYP450 Inhibition (Full Set, ASNN, Bagging) | trained using the dataset PubChem_CYP1A2_Train ( view dataset profile or export the dataset)<br>validated using dataset PubChem_CYP1A2_Test ( view dataset profile or export the dataset) |
| CYP450 Inhibition (Dragon, J48, Bagging) | trained using the dataset PubChem_CYP1A2_Train ( view dataset profile or export the dataset)<br>validated using dataset PubChem_CYP1A2_Test ( view dataset profile or export the dataset) |
| CYP450 Inhibition (SMF, J48, Bagging) | trained using the dataset PubChem_CYP1A2_Train ( view dataset profile or export the dataset)<br>validated using dataset PubChem_CYP1A2_Test ( view dataset profile or export the dataset) |

**Exemplary item profiles:**

- Compound property editor (record editor)
- Molecule profile
- Article profile
- Dataset (“basket”) profile
- Property editor
- Unit editor

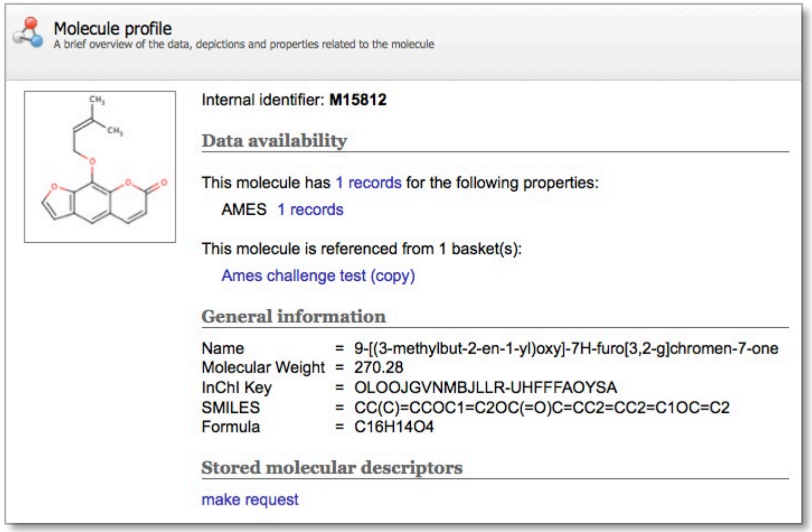


● AMES = active
Sushko, I et al
Applicability domains for
classification problems: Benchmark...
N: 2181
J. Chem. Inf. Model. **2010**; 50 (12) 2094-111

MoleculeID: *M15812*
Public record

Dataset = Test
Cell strain = S. typhimurium
S9 metabolic bioactivation = unspecified

RecordID: *R283104*
15:44, 7 Oct 09 / 17:43, 7 Dec 12
midnighter



Molecule profile
A brief overview of the data, depictions and properties related to the molecule

Internal identifier: **M15812**

**Data availability**

This molecule has 1 records for the following properties:
AMES 1 records

This molecule is referenced from 1 basket(s):
Ames challenge test (copy)

**General information**

| | | |
|---|---|---|
| Name | = | 9-[(3-methylbut-2-en-1-yl)oxy]-7H-furo[3,2-g]chromen-7-one |
| Molecular Weight | = | 270.28 |
| InChI Key | = | OLOOJGVNMBJLLR-UHFFFAOYSA |
| SMILES | = | CC(C)=CCOC1=C2OC(=O)C=CC2=CC2=C1OC=C2 |
| Formula | = | C16H14O4 |

**Stored molecular descriptors**

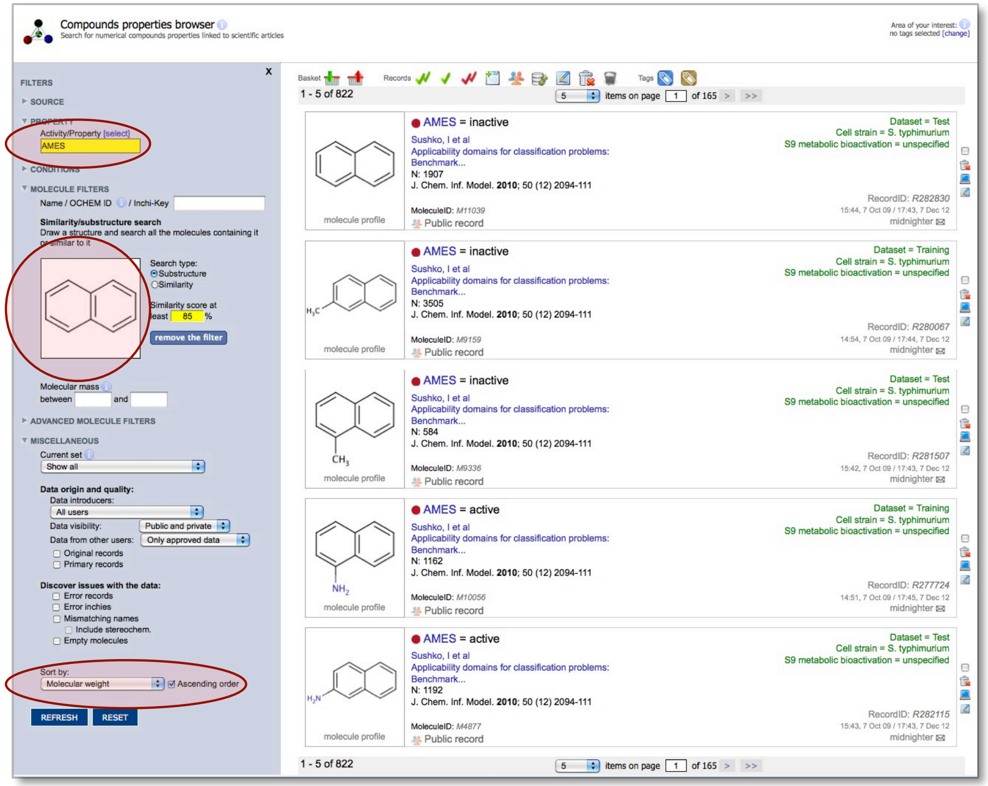
make request

# 2. Using OCHEM

In this chapter, we will how to use OCHEM for several typical scenarios, including

- Search of properties using compound properties browser
- Grouping and exporting records
- Use of ToxAlert for data exploration
- SetCompare Tool for basket comparison
- Prediction of properties for molecules using models

## 2.1 Compound properties browser



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

Area of your interest:
no tags selected [change]

FILTERS
▸ SOURCE
▾ PROPERTY
Activity/Property [select]
AMES
▸ CONDITIONS
▾ MOLECULE FILTERS
Name / OCHEM ID / Inchi-Key
Similarity/substructure search
Draw a structure and search all the molecules containing it or similar to it
Search type:
● Substructure
○ Similarity
Similarity score at least 85 %
remove the filter
Molecular mass between and
▸ ADVANCED MOLECULE FILTERS
▾ MISCELLANEOUS
Current set: Show all
Data origin and quality:
Data introducers: All users
Data visibility: Public and private
Data from other users: Only approved data
□ Original records
□ Primary records
Discover issues with the data:
□ Error records
□ Error inchies
□ Mismatching names
□ Include stereochem.
□ Empty molecules
Sort by: Molecular weight ☑ Ascending order
REFRESH RESET

Basket Records Tags
1 - 5 of 822 | 5 items on page 1 of 165 > >>

● AMES = inactive
Sushko, I et al
Applicability domains for classification problems: Benchmark...
N: 1907
J. Chem. Inf. Model. **2010**; 50 (12) 2094-111
MoleculeID: *M11039*
Public record
Dataset = Test
Cell strain = S. typhimurium
S9 metabolic bioactivation = unspecified
RecordID: *R282830*
15:44, 7 Oct 09 / 17:43, 7 Dec 12
midnighter

● AMES = inactive
Sushko, I et al
Applicability domains for classification problems: Benchmark...
N: 3505
J. Chem. Inf. Model. **2010**; 50 (12) 2094-111
MoleculeID: *M9159*
Public record
Dataset = Training
Cell strain = S. typhimurium
S9 metabolic bioactivation = unspecified
RecordID: *R280067*
14:54, 7 Oct 09 / 17:44, 7 Dec 12
midnighter

● AMES = inactive
Sushko, I et al
Applicability domains for classification problems: Benchmark...
N: 584
J. Chem. Inf. Model. **2010**; 50 (12) 2094-111
MoleculeID: *M9336*
Public record
Dataset = Test
Cell strain = S. typhimurium
S9 metabolic bioactivation = unspecified
RecordID: *R281507*
15:42, 7 Oct 09 / 17:43, 7 Dec 12
midnighter

● AMES = active
Sushko, I et al
Applicability domains for classification problems: Benchmark...
N: 1162
J. Chem. Inf. Model. **2010**; 50 (12) 2094-111
MoleculeID: *M10056*
Public record
Dataset = Training
Cell strain = S. typhimurium
S9 metabolic bioactivation = unspecified
RecordID: *R277724*
14:51, 7 Oct 09 / 17:45, 7 Dec 12
midnighter

● AMES = active
Sushko, I et al
Applicability domains for classification problems: Benchmark...
N: 1192
J. Chem. Inf. Model. **2010**; 50 (12) 2094-111
MoleculeID: *M4877*
Public record
Dataset = Test
Cell strain = S. typhimurium
S9 metabolic bioactivation = unspecified
RecordID: *R282115*
15:43, 7 Oct 09 / 17:43, 7 Dec 12
midnighter

1 - 5 of 822 | 5 items on page 1 of 165 > >>

Compound properties browser is one the main dialogues in OCHEM. It allows you to browse experimental data records using a variety of filters.

Please, try using different filters, e.g.:

1. Show only the data for the “Ames“ property
2. Filter the records by substructure
3. Sort the data by ascending molecular weight

For further use, records can be selected individually ☑ or, more commonly, it is possible to select all the records matching current filters ✓✓

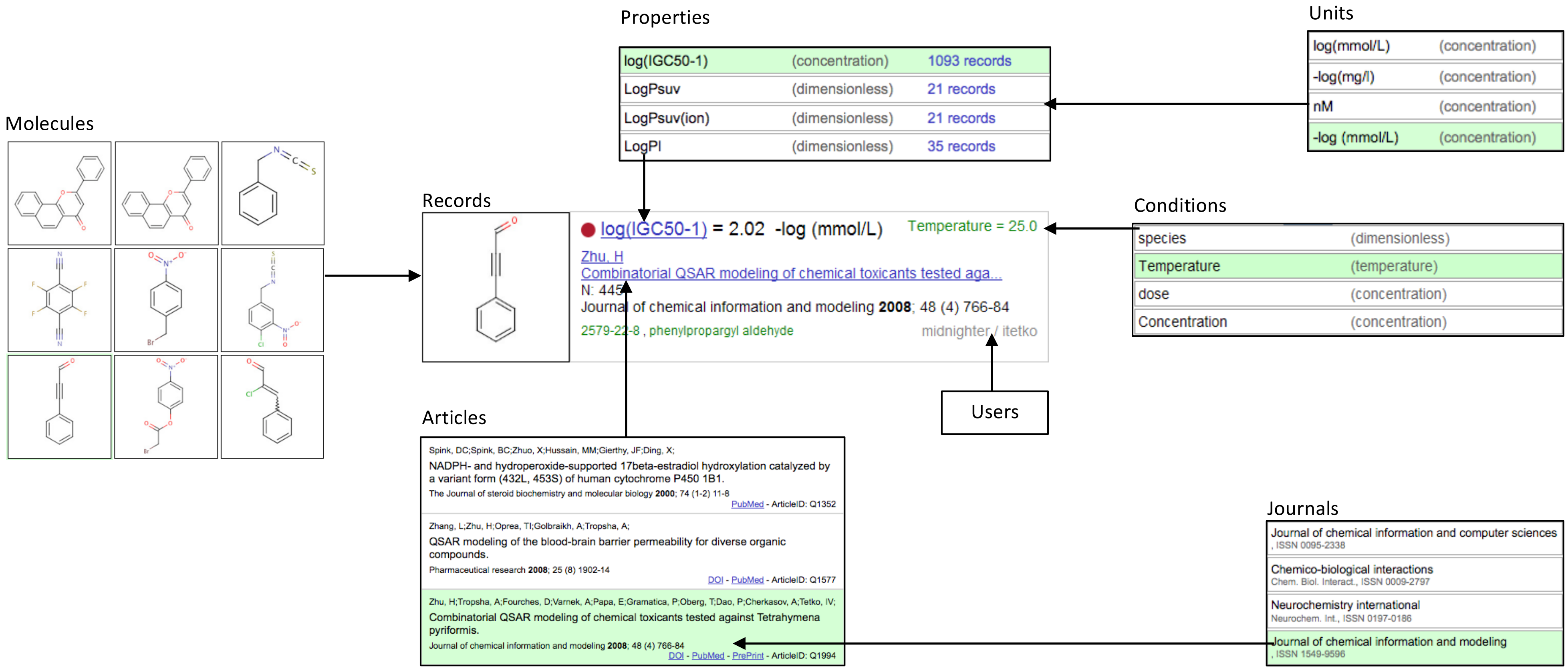
The selected records can be put into a dataset (so called “basket”), which later can be reused for any kind of tasks, e.g. for development or validation of QSAR models.

The selected records are shown at the top of the left panel and persist until they are deselected (cleared). Clear them before selecting new records unless you would like to add them to the previous records.

**Other filters of property records are available from browsers of Articles, Properties, Baskets, Models, etc.**

# OCHEM tutorial handout

## 2.2 Data structure



Properties

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

Units

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

Records

● 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

Conditions

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

Articles

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

Journals

Journal of chemical information and computer sciences
, ISSN 0095-2338

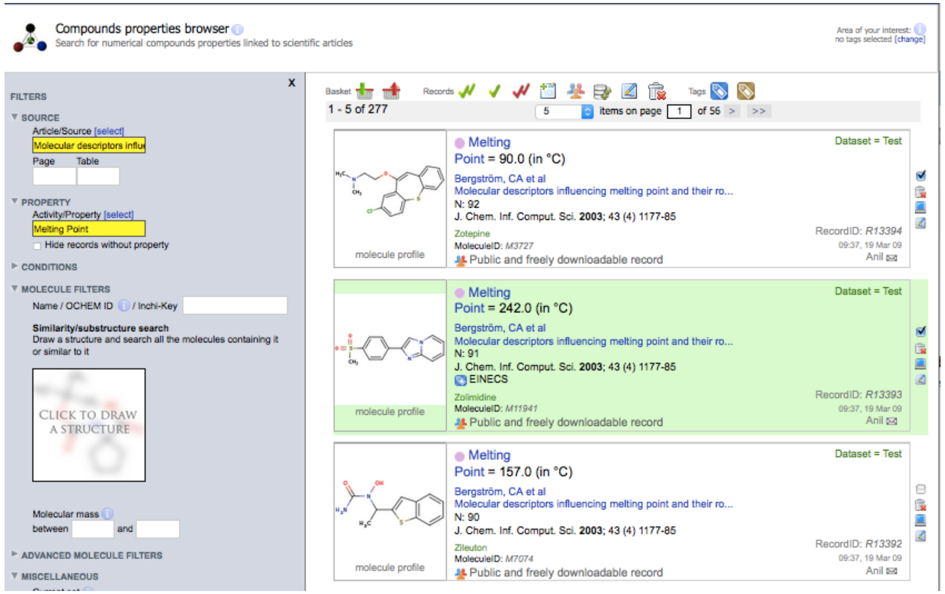
Chemico-biological interactions
Chem. Biol. Interact., ISSN 0009-2797

Neurochemistry international
Neurochem. Int., ISSN 0197-0186

Journal of chemical information and modeling
, ISSN 1549-9596

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)

## 2.3 Working with baskets and data export



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

Area of your interest:
no tags selected [change]

FILTERS

▼ SOURCE
Article/Source [select]
Molecular descriptors influ
Page Table

▼ PROPERTY
Activity/Property [select]
Melting Point
□ Hide records without property

▶ CONDITIONS

▼ MOLECULE FILTERS
Name / OCHEM ID / Inchi-Key
Similarity/substructure search
Draw a structure and search all the molecules containing it or similar to it
CLICK TO DRAW A STRUCTURE
Molecular mass
between and

▶ ADVANCED MOLECULE FILTERS

▼ MISCELLANEOUS

1 - 5 of 277 5 items on page 1 of 56 > >>

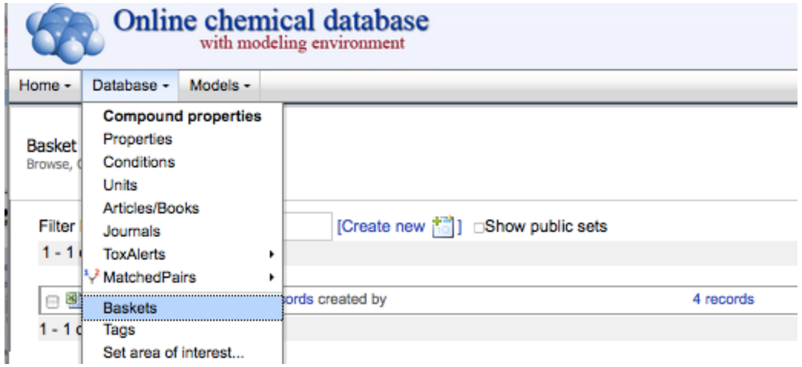
Melting Point = 90.0 (in °C)
Dataset = Test
Bergström, CA et al
Molecular descriptors influencing melting point and their ro...
N: 92
J. Chem. Inf. Comput. Sci. **2003**; 43 (4) 1177-85
Zotepine
MoleculeID: *M3727*
RecordID: *R13394*
09:37, 19 Mar 09
Anil
Public and freely downloadable record
molecule profile

Melting Point = 242.0 (in °C)
Dataset = Test
Bergström, CA et al
Molecular descriptors influencing melting point and their ro...
N: 91
J. Chem. Inf. Comput. Sci. **2003**; 43 (4) 1177-85
EINECS
Zolimidine
MoleculeID: *M11941*
RecordID: *R13393*
09:37, 19 Mar 09
Anil
Public and freely downloadable record
molecule profile

Melting Point = 157.0 (in °C)
Dataset = Test
Bergström, CA et al
Molecular descriptors influencing melting point and their ro...
N: 90
J. Chem. Inf. Comput. Sci. **2003**; 43 (4) 1177-85
Zileuton
MoleculeID: *M7074*
RecordID: *R13392*
09:37, 19 Mar 09
Anil
Public and freely downloadable record
molecule profile

### 1. Finding interesting records

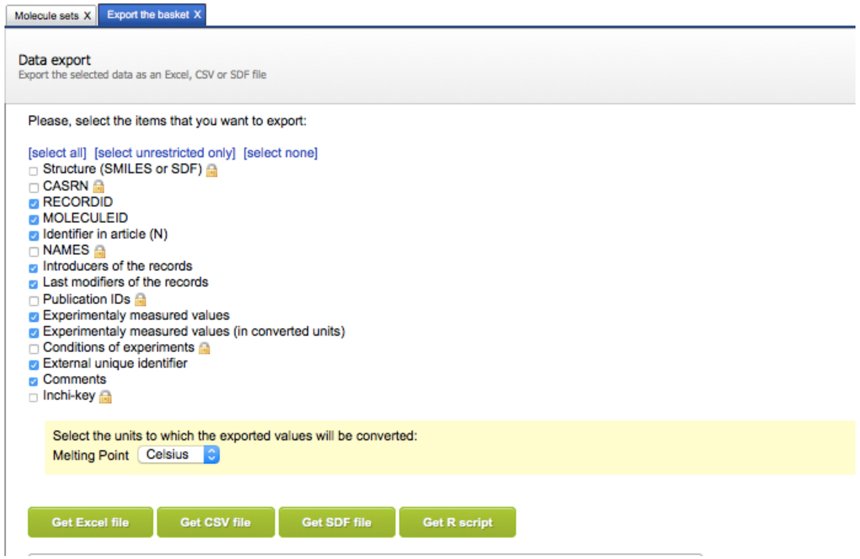
Filtering of records using filters such as article, model, property, basket or selection of records one by one, on page or all filtered



Online chemical database
with modeling environment

Basket
Browse, c
Filter
[Create new ] □Show public sets
1 - 1
ords created by
4 records
1 - 1 c

### 2. Exploring the baskets



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

Please, select the items that you want to export:

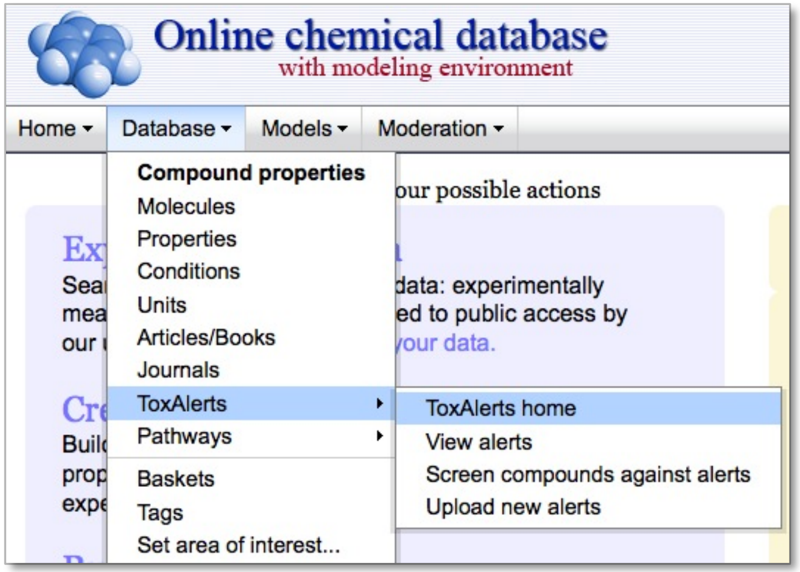
[select all] [select unrestricted only] [select none]

- □ Structure (SMILES or SDF)
- □ CASRN
- ☑ RECORDID
- ☑ MOLECULEID
- ☑ Identifier in article (N)
- □ NAMES
- ☑ Introducers of the records
- ☑ Last modifiers of the records
- □ Publication IDs
- ☑ Experimentaly measured values
- ☑ Experimentaly measured values (in converted units)
- □ Conditions of experiments
- ☑ External unique identifier
- ☑ Comments
- □ Inchi-key

Select the units to which the exported values will be converted:
Melting Point Celsius

### 3. Exporting the records

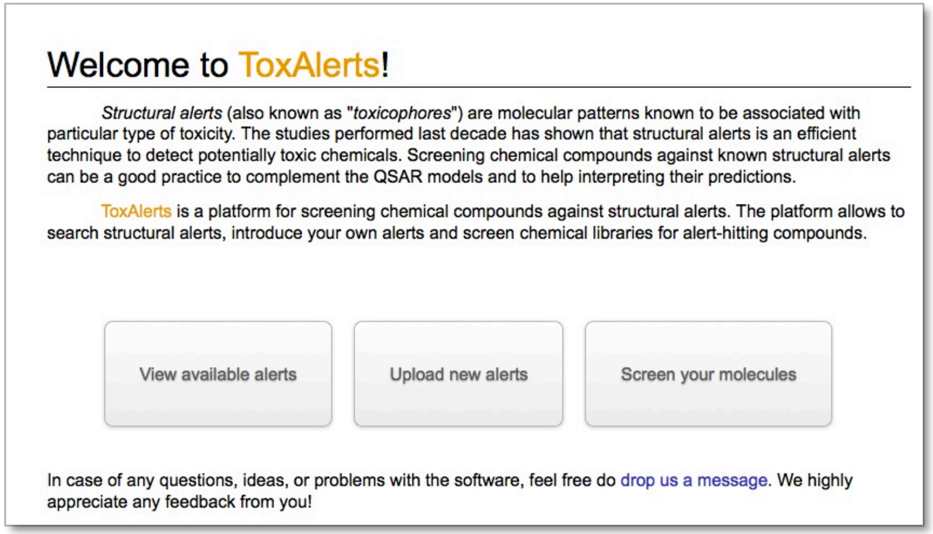
## 2.3 ToxAlert utility



Alerts are structural features that are known to be associated with a particular activity. Typical structural alerts might indicate carcinogenicity or general toxicity. Screening molecules against alerts can be simple and, most importantly, easily interpretable

The ***ToxAlert*** utility allows one to screen a set of molecules against a set of structural alerts. OCHEM comes with thousands of alerts for a number of endpoints. It is also possible to introduce your own alerts

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



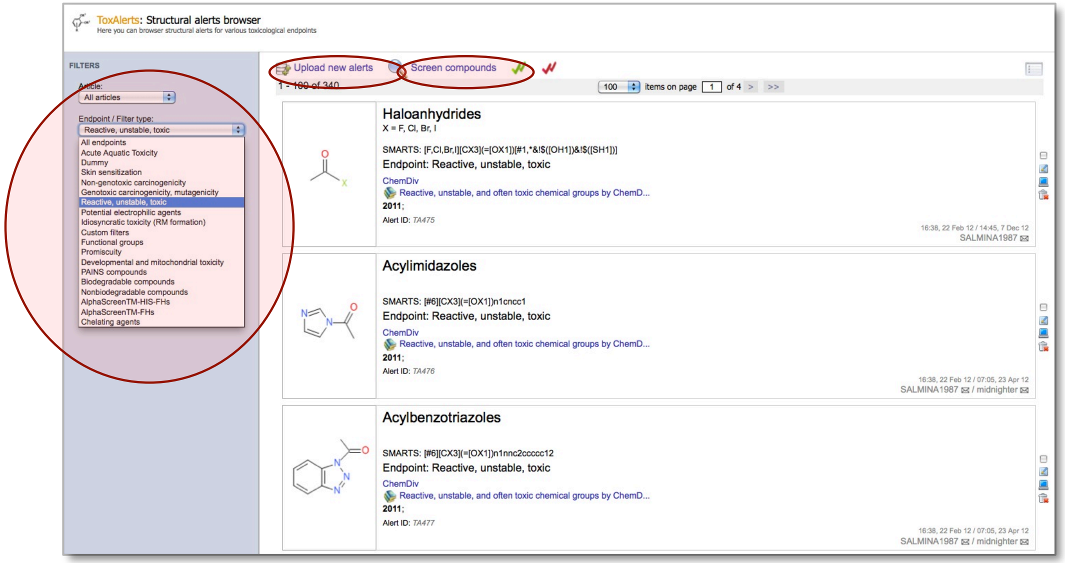
Welcome to ToxAlerts!

*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 do drop us a message. We highly appreciate any feedback from you!

2. A welcome page is the entry point for further actions, like
   - overview available alerts
   - upload new structural alerts
   - screen molecules against structural alerts



ToxAlerts: Structural alerts browser
Here you can browser structural alerts for various toxicological endpoints

FILTERS

Article: All articles

Endpoint / Filter type: Reactive, unstable, toxic

- All endpoints
- Acute Aquatic Toxicity
- Dummy
- Skin sensitization
- Non-genotoxic carcinogenicity
- Genotoxic carcinogenicity, mutagenicity
- Reactive, unstable, toxic
- Potential electrophilic agents
- Idiosyncratic toxicity (RM formation)
- Custom filters
- Functional groups
- Promiscuity
- Developmental and mitochondrial toxicity
- PAINS compounds
- Biodegradable compounds
- Nonbiodegradable compounds
- AlphaScreenTM-HIS-FHs
- AlphaScreenTM-FHs
- Chelating agents

1 - 100 of 340 | 100 items on page 1 of 4 > >>

**Haloanhydrides**
X = F, Cl, Br, I

SMARTS: [F,Cl,Br,I][CX3](=[OX1])[#1,*&!$([OH1])&!$([SH1])]
Endpoint: Reactive, unstable, toxic
ChemDiv
Reactive, unstable, and often toxic chemical groups by ChemD...
**2011**;
Alert ID: *TA475*
16:38, 22 Feb 12 / 14:45, 7 Dec 12
SALMINA1987

**Acylimidazoles**

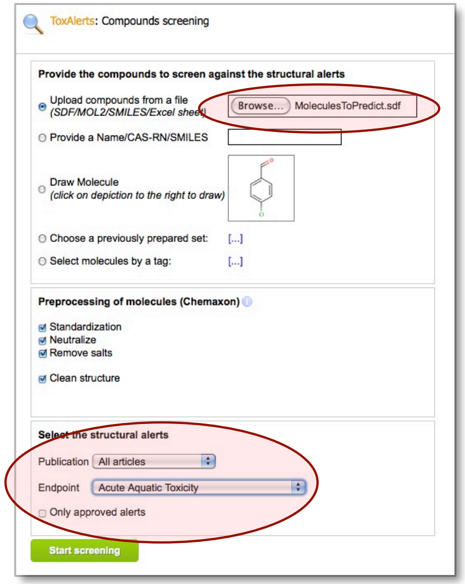
SMARTS: [#6][CX3](=[OX1])n1cncc1
Endpoint: Reactive, unstable, toxic
ChemDiv
Reactive, unstable, and often toxic chemical groups by ChemD...
**2011**;
Alert ID: *TA476*
16:38, 22 Feb 12 / 07:05, 23 Apr 12
SALMINA1987 / midnighter

**Acylbenzotriazoles**

SMARTS: [#6][CX3](=[OX1])n1nnc2ccccc12
Endpoint: Reactive, unstable, toxic
ChemDiv
Reactive, unstable, and often toxic chemical groups by ChemD...
**2011**;
Alert ID: *TA477*
16:38, 22 Feb 12 / 07:05, 23 Apr 12
SALMINA1987 / midnighter

3. The structural alerts browser gives an overview of the available alerts in the system.

   - Existing alerts can be filtered by their category
   - New alerts can be uploaded
   - Selected alerts can be used in a screening against a set of structures



ToxAlerts: Compounds screening

**Provide the compounds to screen against the structural alerts**

● Upload compounds from a file *(SDF/MOL2/SMILES/Excel sheet)* Browse... MoleculesToPredict.sdf

○ 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: [...]

**Preprocessing of molecules (Chemaxon)**

☑ Standardization
☑ Neutralize
☑ Remove salts

☑ Clean structure

**Select the structural alerts**

Publication: All articles

Endpoint: Acute Aquatic Toxicity

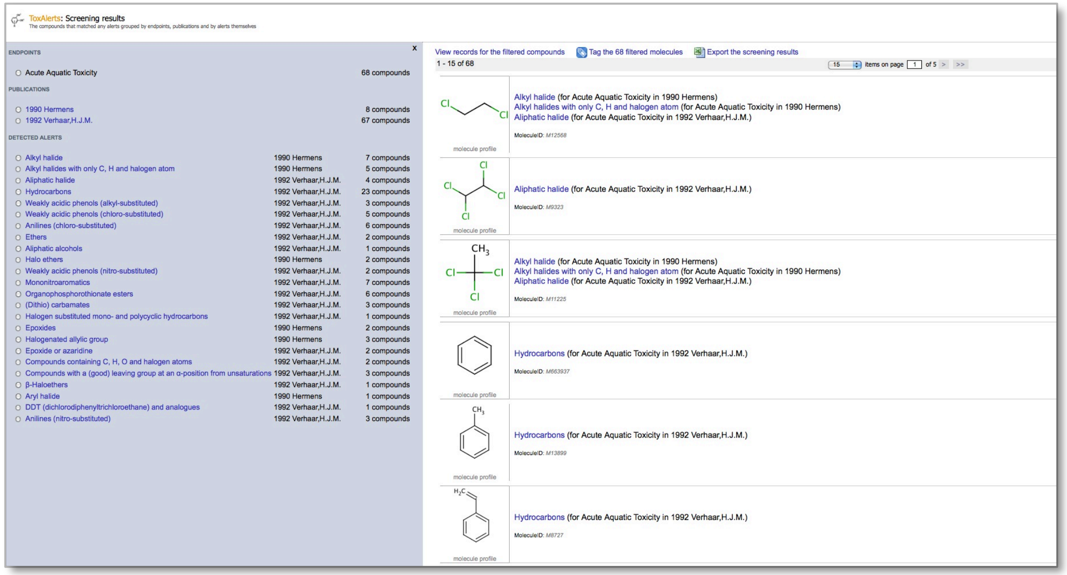
☐ Only approved alerts

**Screening compounds against alerts**

4. Select the compounds you would like to screen. This can be a prepared basket from the OCHEM platform, a single structure drawn or automatically fetched by its name, or like in this example an uploaded file (SDF, Smiles, Excel).

5. Optionally, you can screen against alerts for a particular endpoint or alerts from a particular publication.

# OCHEM tutorial handout



ToxAlerts: Screening results

The compounds that matched any alerts grouped by endpoints, publications and by alerts themselves

ENDPOINTS

|  |  |
|---|---|
| ○ Acute Aquatic Toxicity | 68 compounds |

PUBLICATIONS

|  |  |
|---|---|
| ○ 1990 Hermens | 8 compounds |
| ○ 1992 Verhaar,H.J.M. | 67 compounds |

DETECTED ALERTS

|  |  |  |
|---|---|---|
| ○ Alkyl halide | 1990 Hermens | 7 compounds |
| ○ Alkyl halides with only C, H and halogen atom | 1990 Hermens | 5 compounds |
| ○ Aliphatic halide | 1992 Verhaar,H.J.M. | 4 compounds |
| ○ Hydrocarbons | 1992 Verhaar,H.J.M. | 23 compounds |
| ○ Weakly acidic phenols (alkyl-substituted) | 1992 Verhaar,H.J.M. | 3 compounds |
| ○ Weakly acidic phenols (chloro-substituted) | 1992 Verhaar,H.J.M. | 5 compounds |
| ○ Anilines (chloro-substituted) | 1992 Verhaar,H.J.M. | 6 compounds |
| ○ Ethers | 1992 Verhaar,H.J.M. | 2 compounds |
| ○ Aliphatic alcohols | 1992 Verhaar,H.J.M. | 1 compounds |
| ○ Halo ethers | 1990 Hermens | 2 compounds |
| ○ Weakly acidic phenols (nitro-substituted) | 1992 Verhaar,H.J.M. | 2 compounds |
| ○ Mononitroaromatics | 1992 Verhaar,H.J.M. | 7 compounds |
| ○ Organophosphorothionate esters | 1992 Verhaar,H.J.M. | 6 compounds |
| ○ (Dithio) carbamates | 1992 Verhaar,H.J.M. | 3 compounds |
| ○ Halogen substituted mono- and polycyclic hydrocarbons | 1992 Verhaar,H.J.M. | 1 compounds |
| ○ Epoxides | 1990 Hermens | 2 compounds |
| ○ Halogenated allylic group | 1990 Hermens | 3 compounds |
| ○ Epoxide or azaridine | 1992 Verhaar,H.J.M. | 2 compounds |
| ○ Compounds containing C, H, O and halogen atoms | 1992 Verhaar,H.J.M. | 2 compounds |
| ○ Compounds with a (good) leaving group at an α-position from unsaturations | 1992 Verhaar,H.J.M. | 3 compounds |
| ○ β-Haloethers | 1992 Verhaar,H.J.M. | 1 compounds |
| ○ Aryl halide | 1990 Hermens | 1 compounds |
| ○ DDT (dichlorodiphenyltrichloroethane) and analogues | 1992 Verhaar,H.J.M. | 1 compounds |
| ○ Anilines (nitro-substituted) | 1992 Verhaar,H.J.M. | 3 compounds |

View records for the filtered compounds | Tag the 68 filtered molecules | Export the screening results

1 - 15 of 68 | 15 items on page 1 of 5 > >>

Alkyl halide (for Acute Aquatic Toxicity in 1990 Hermens)
Alkyl halides with only C, H and halogen atom (for Acute Aquatic Toxicity in 1990 Hermens)
Aliphatic halide (for Acute Aquatic Toxicity in 1992 Verhaar,H.J.M.)
MoleculeID: *M12568*
molecule profile

Aliphatic halide (for Acute Aquatic Toxicity in 1992 Verhaar,H.J.M.)
MoleculeID: *M9323*
molecule profile

Alkyl halide (for Acute Aquatic Toxicity in 1990 Hermens)
Alkyl halides with only C, H and halogen atom (for Acute Aquatic Toxicity in 1990 Hermens)
Aliphatic halide (for Acute Aquatic Toxicity in 1992 Verhaar,H.J.M.)
MoleculeID: *M11225*
molecule profile

Hydrocarbons (for Acute Aquatic Toxicity in 1992 Verhaar,H.J.M.)
MoleculeID: *M663937*
molecule profile

Hydrocarbons (for Acute Aquatic Toxicity in 1992 Verhaar,H.J.M.)
MoleculeID: *M13899*
molecule profile

Hydrocarbons (for Acute Aquatic Toxicity in 1992 Verhaar,H.J.M.)
MoleculeID: *M8727*
molecule profile

6. As result you can see the structures for which a particular alert was found and from which publication this alert stems.

Further information about ToxAlert screening is available in the OCHEM documentation (ToxAlerts)

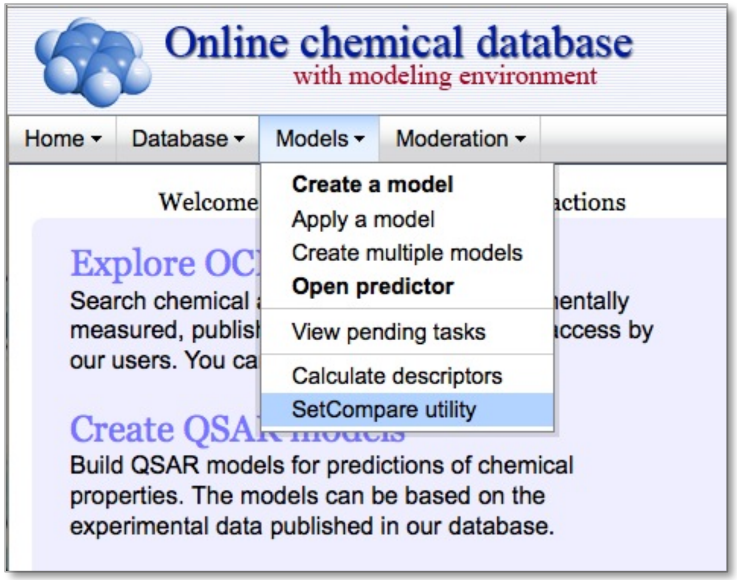
## 2.5 Set Compare utility

***The SetCompare*** utility allows juxtaposing two sets of chemical compounds and finding distinguishing features of each set. For example, you can compare active and inactive compounds for a particular property.

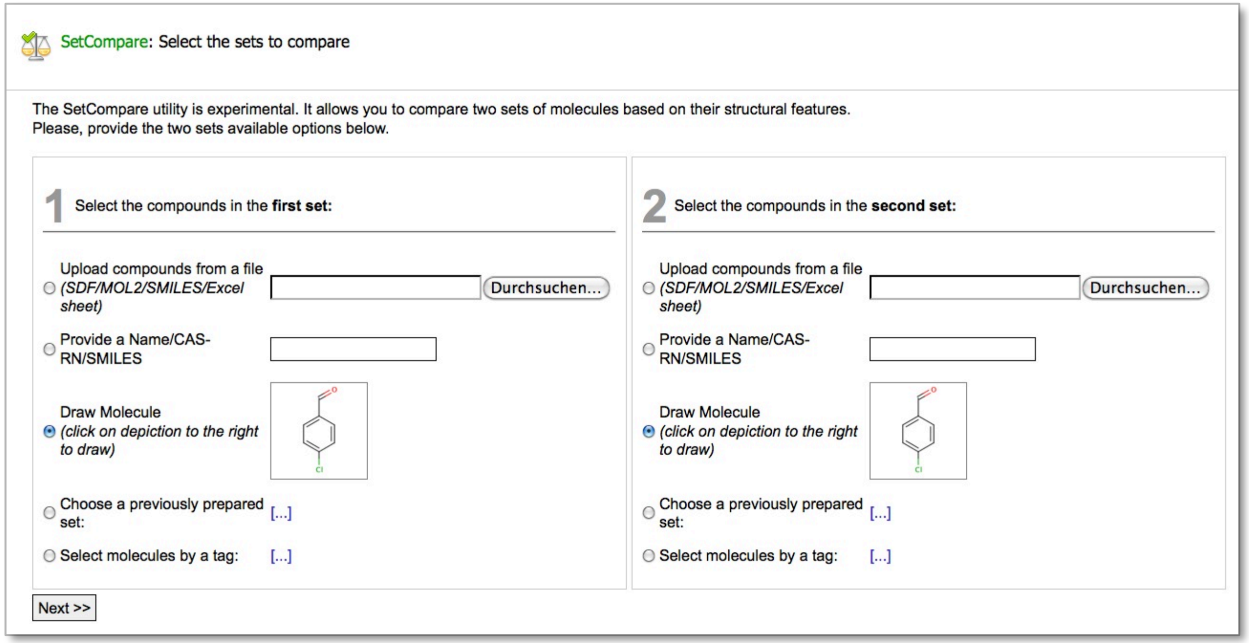
The utility helps to address the following questions:

- What are the distinguishing structural features of active compounds?
- How significant are these results statistically?
- Which are the compounds that possess a particular important structural feature?

SetCompare is accessible from the menu bar.



In the first page of the wizard two sets have to be selected. 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:**

○ Upload compounds from a file *(SDF/MOL2/SMILES/Excel sheet)* Durchsuchen...

○ 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:**

○ Upload compounds from a file *(SDF/MOL2/SMILES/Excel sheet)* Durchsuchen...

○ Provide a Name/CAS-RN/SMILES

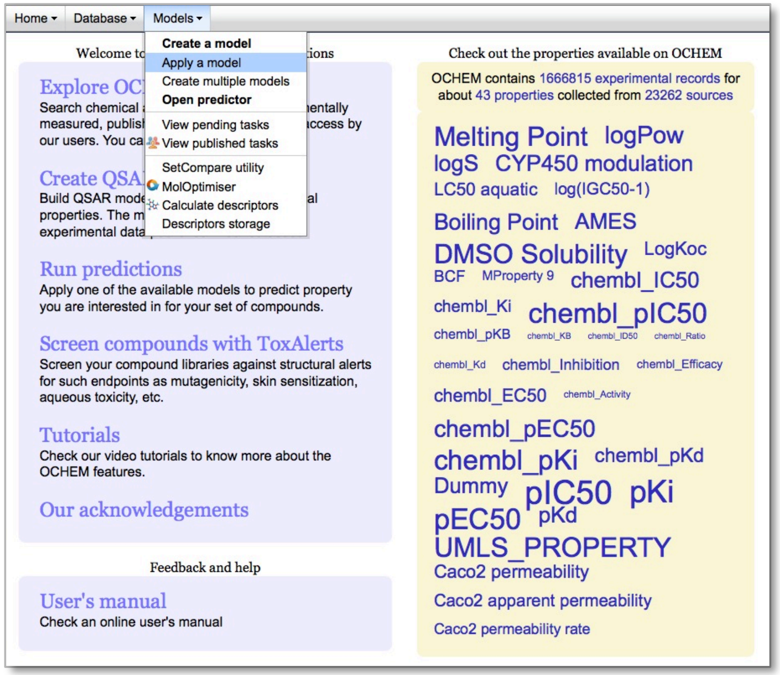
● Draw Molecule *(click on depiction to the right to draw)*

○ Choose a previously prepared set: [...]

○ Select molecules by a tag: [...]

Next >>

## 2.6 Model application



Welcome to ions

Explore OC
Search chemical entally measured, publish ccess by our users. You ca

Create QSA
Build QSAR mode al properties. The m experimental data

Run predictions
Apply one of the available models to predict property you are interested in for your set of compounds.

Screen compounds with ToxAlerts
Screen your compound libraries against structural alerts for such endpoints as mutagenicity, skin sensitization, aqueous toxicity, etc.

Tutorials
Check our video tutorials to know more about the OCHEM features.

Our acknowledgements

Feedback and help

User's manual
Check an online user's manual

Check out the properties available on OCHEM

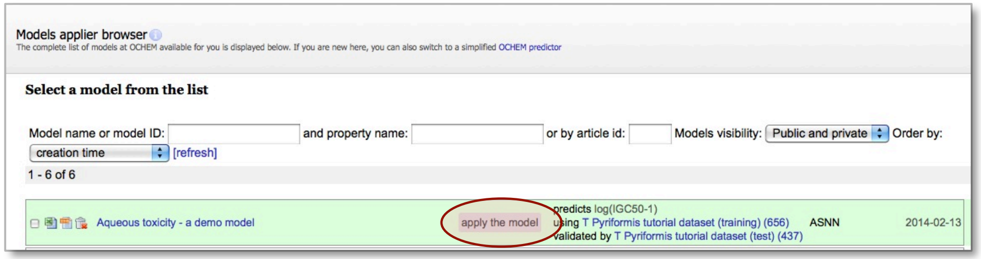
OCHEM contains 1666815 experimental records for about 43 properties collected from 23262 sources

**Getting to model application**

**1.** Open the model applier browser from Models > Apply a model.

**Models applier browser**

- The models applier browser lists all the models (public and private, developed by the user). It shows the model name, the predicted properties, used training set, used machine learning method and the creation date.
- The 📊 icon links to model export.



Models applier browser
The complete list of models at OCHEM available for you is displayed below. If you are new here, you can also switch to a simplified OCHEM predictor

Select a model from the list

Model name or model ID: [ ] and property name: [ ] or by article id: [ ] Models visibility: Public and private Order by: creation time [refresh]

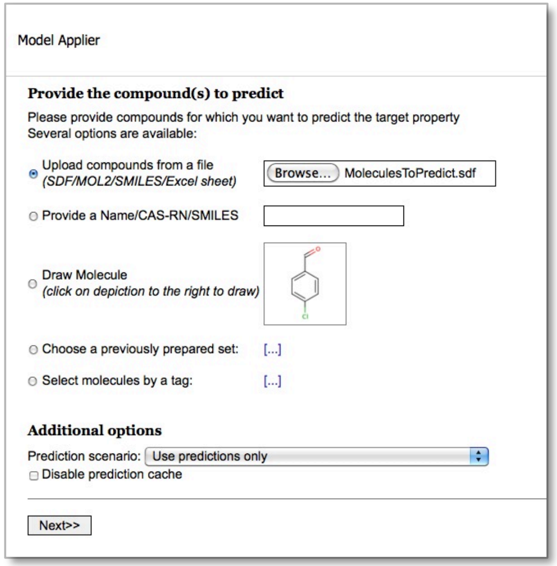
1 - 6 of 6

□ Aqueous toxicity - a demo model | apply the model | predicts log(IGC50-1) using T Pyriformis tutorial dataset (training) (656) validated by T Pyriformis tutorial dataset (test) (437) | ASNN | 2014-02-13

**2.** Please, find the model you want to apply and click “apply the model”.

Note:
You can check and apply multiple models simultaneously.

## Application of regression model:



Model Applier

**Provide the compound(s) to predict**

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

● Upload compounds from a file *(SDF/MOL2/SMILES/Excel sheet)* Browse... MoleculesToPredict.sdf

○ 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: [...]

**Additional options**

Prediction scenario: Use predictions only

□ Disable prediction cache

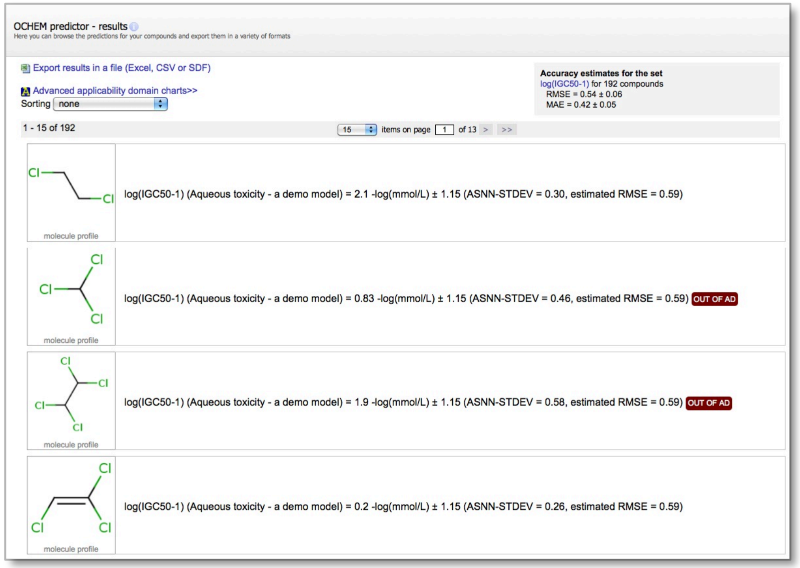
Next>>

**3.** Now it is time to provide the compounds you would like to predict. There are several possibilities:

- Upload structures (e.g., in SD-format)
- Provide SMILES for a single molecule
- Draw a structure in a visual structure editor
- Use an earlier created basket
- Select a certain set of records / molecules by a tag

Selecting the prediction scenario and disabling the prediction cache are additional options for the prediction process.

**4.** Click on next button to start the application
Wait until the calculations are completed. Again, if the task is taking long time, it is possible to fetch results anytime later from the pending tasks browser (Models > Pending tasks menu).



OCHEM predictor - results

Here you can browse the predictions for your compounds and export them in a variety of formats

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

Advanced applicability domain charts>>

Sorting none

Accuracy estimates for the set
log(IGC50-1) for 192 compounds
RMSE = 0.54 ± 0.06
MAE = 0.42 ± 0.05

1 - 15 of 192 — 15 items on page 1 of 13 > >>

log(IGC50-1) (Aqueous toxicity - a demo model) = 2.1 -log(mmol/L) ± 1.15 (ASNN-STDEV = 0.30, estimated RMSE = 0.59)

log(IGC50-1) (Aqueous toxicity - a demo model) = 0.83 -log(mmol/L) ± 1.15 (ASNN-STDEV = 0.46, estimated RMSE = 0.59) OUT OF AD

log(IGC50-1) (Aqueous toxicity - a demo model) = 1.9 -log(mmol/L) ± 1.15 (ASNN-STDEV = 0.58, estimated RMSE = 0.59) OUT OF AD

log(IGC50-1) (Aqueous toxicity - a demo model) = 0.2 -log(mmol/L) ± 1.15 (ASNN-STDEV = 0.26, estimated RMSE = 0.59)

**5.** Now the calculations have finished. The results include prediction values, and accuracy estimates for each predicted compound. Additionally, there is an estimate for the overall prediction accuracy for the set.

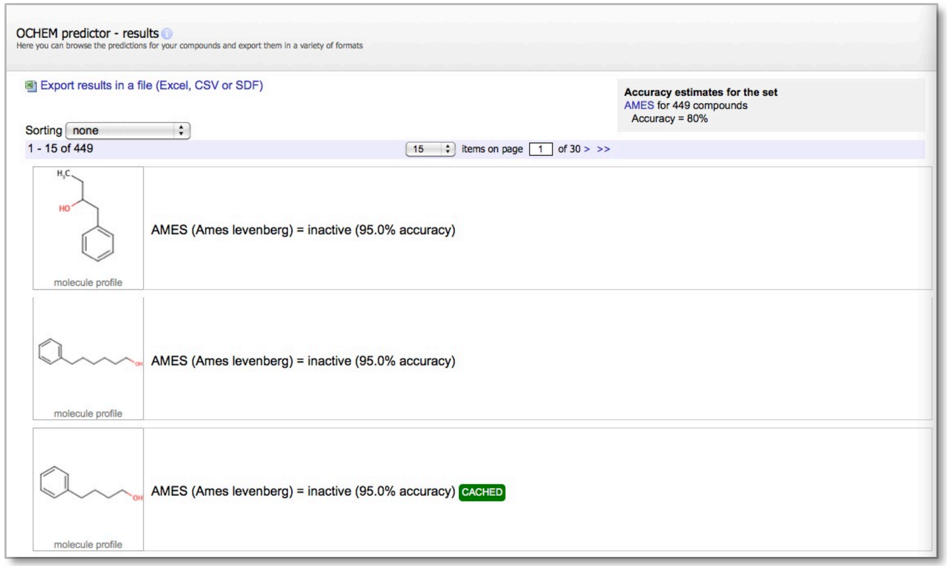
**6.** Predictions can be exported to an Excel or CSV sheet, SDF files or R scripts.

**7.** Prediction results of each single structure are listed in the browser. There are the predicted values themselves, distance to model value and an estimation of the accuracy (RMSE), as well as the originally measured value if it is known.

### Application of classification models

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

8. 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.



OCHEM predictor - results

Here you can browse the predictions for your compounds and export them in a variety of formats

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

Accuracy estimates for the set
AMES for 449 compounds
Accuracy = 80%

Sorting: none

1 - 15 of 449 — 15 items on page 1 of 30 > >>

AMES (Ames levenberg) = inactive (95.0% accuracy)

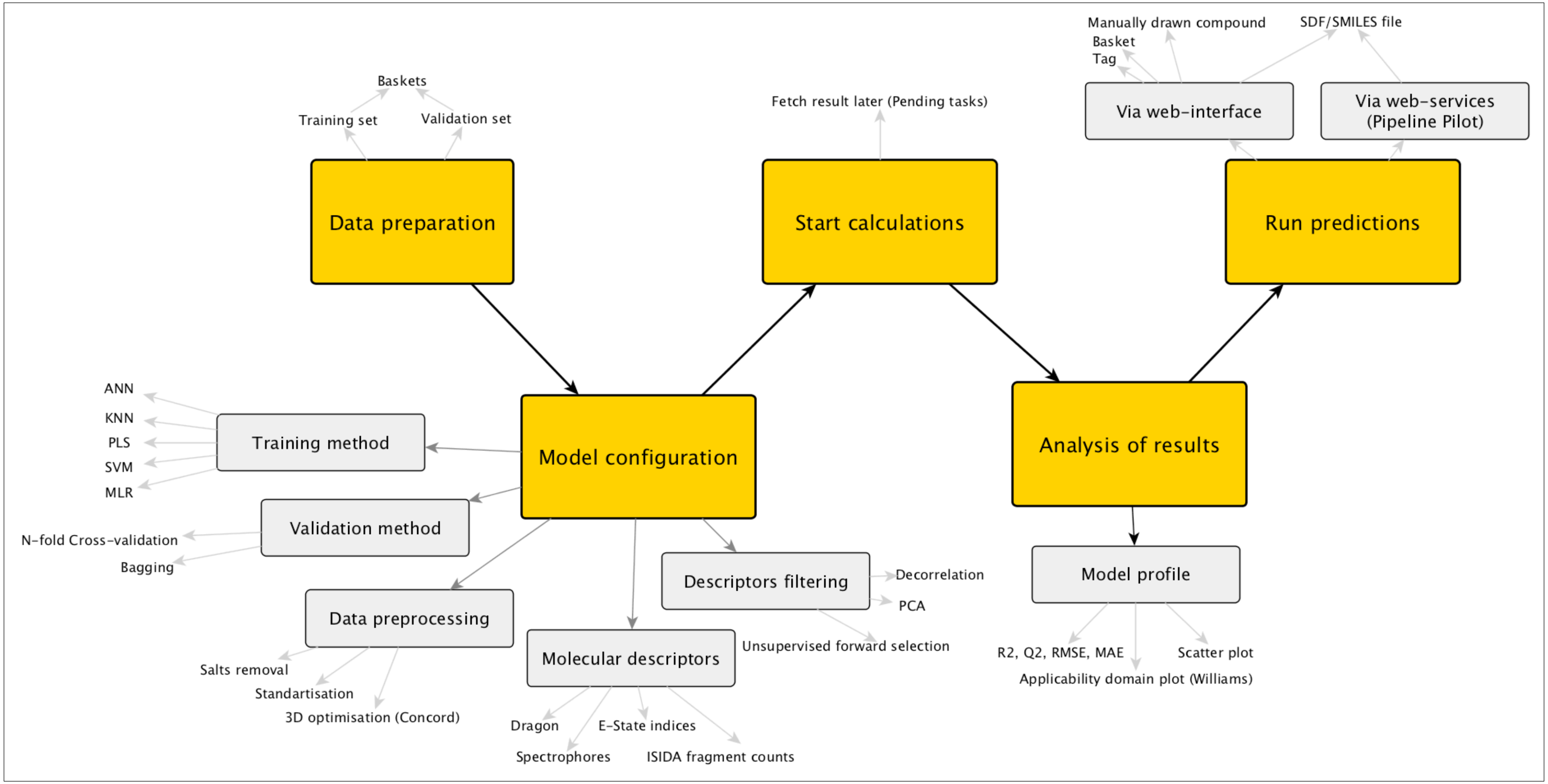
AMES (Ames levenberg) = inactive (95.0% accuracy)

AMES (Ames levenberg) = inactive (95.0% accuracy) CACHED

# 3 Modeling framework

In this chapter, we learn how to upload data and develop QSAR models

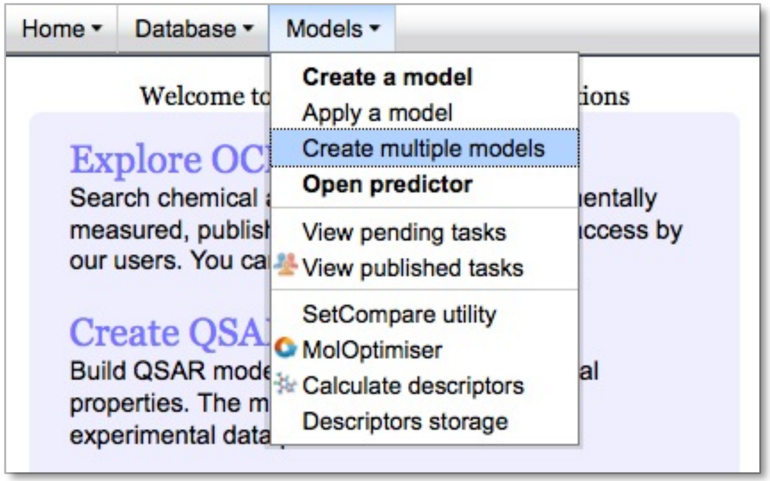
Development of multiple models using Comprehensive Modeling (CM)
Development of single models



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

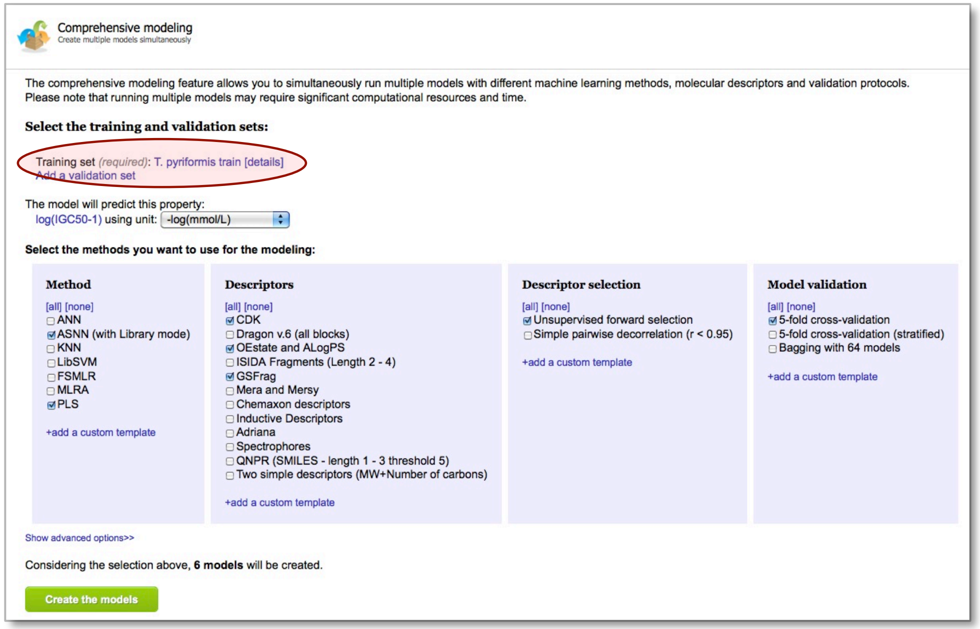
# OCHEM tutorial handout

## 3.1 Comprehensive modeling



**1.** The "comprehensive modelling" feature accessible via the "Models" menu 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.



Comprehensive modeling
Create multiple models simultaneously

The comprehensive modeling feature allows you to simultaneously run multiple models with different machine learning methods, molecular descriptors and validation protocols. Please note that running multiple models may require significant computational resources and time.

**Select the training and validation sets:**

Training set *(required)*: T. pyriformis train [details]
Add a validation set

The model will predict this property:
log(IGC50-1) using unit: -log(mmol/L)

**Select the methods you want to use for the modeling:**

**Method**
[all] [none]
□ ANN
☑ ASNN (with Library mode)
□ KNN
□ LibSVM
□ FSMLR
□ MLRA
☑ PLS
+add a custom template

**Descriptors**
[all] [none]
☑ CDK
□ Dragon v.6 (all blocks)
☑ OEstate and ALogPS
□ ISIDA Fragments (Length 2 - 4)
☑ GSFrag
□ Mera and Mersy
□ Chemaxon descriptors
□ Inductive Descriptors
□ Adriana
□ Spectrophores
□ QNPR (SMILES - length 1 - 3 threshold 5)
□ Two simple descriptors (MW+Number of carbons)
+add a custom template

**Descriptor selection**
[all] [none]
☑ Unsupervised forward selection
□ Simple pairwise decorrelation (r < 0.95)
+add a custom template

**Model validation**
[all] [none]
☑ 5-fold cross-validation
□ 5-fold cross-validation (stratified)
□ Bagging with 64 models
+add a custom template

Show advanced options>>

Considering the selection above, **6 models** will be created.

**2.** In the following dialog, first you should select your training set**.**

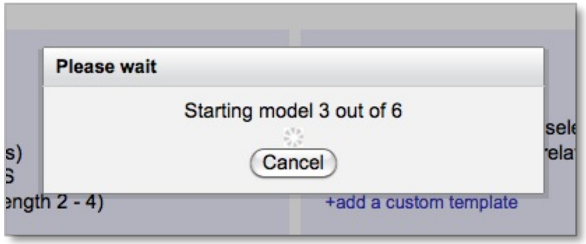
**3.** You can see a set of predefined configuration templates for several training methods, molecular descriptors, descriptor selection methods and model validation.

The checked methods will be applied using the "all against all" principle. On the following screenshot, we selected methods, descriptor sets, descriptor selection method and validation method, which results into six models.

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

**4.** Now we are ready to launch all models.

# OCHEM tutorial handout



**Please wait**

Starting model 3 out of 6

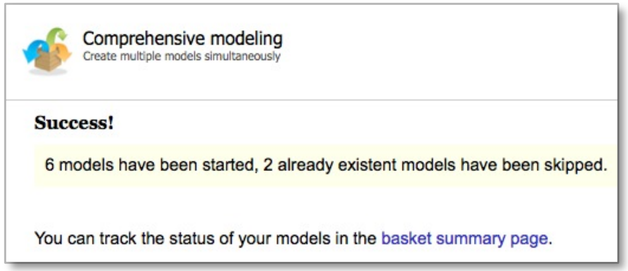
Cancel

s)
s
ength 2 - 4)

sele
rela

+add a custom template

5. Please, wait until OCHEM starts the necessary calculation tasks.



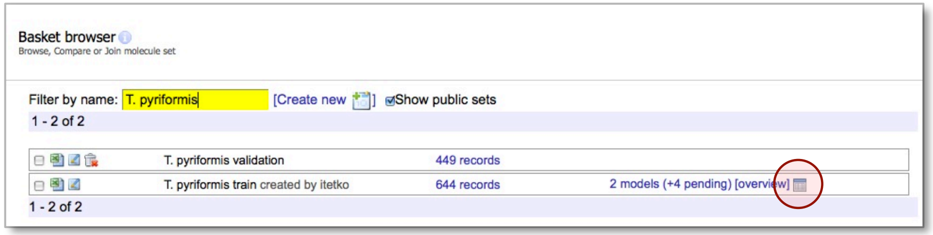
Comprehensive modeling
Create multiple models simultaneously

**Success!**

6 models have been started, 2 already existent models have been skipped.

You can track the status of your models in the basket summary page.

6. When done, you are forwarded to the success page, from which you can directly go to the **models summary page**.



Basket browser
Browse, Compare or Join molecule set

Filter by name: T. pyriformis [Create new ] ☑Show public sets

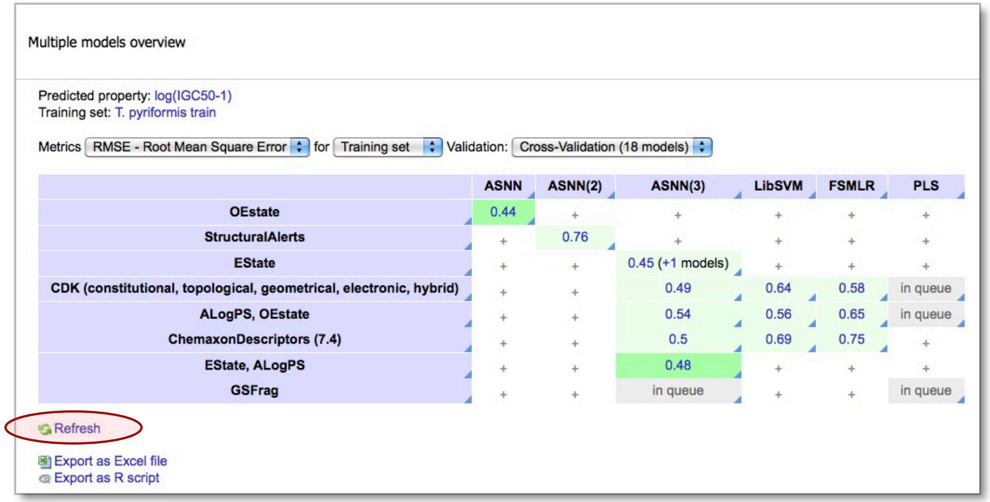
1 - 2 of 2

| | | | |
|---|---|---|---|
| □ | T. pyriformis validation | 449 records | |
| □ | T. pyriformis train created by itetko | 644 records | 2 models (+4 pending) [overview] ▦ |

1 - 2 of 2

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

# OCHEM tutorial handout



Multiple models overview

Predicted property: log(IGC50-1)
Training set: T. pyriformis train

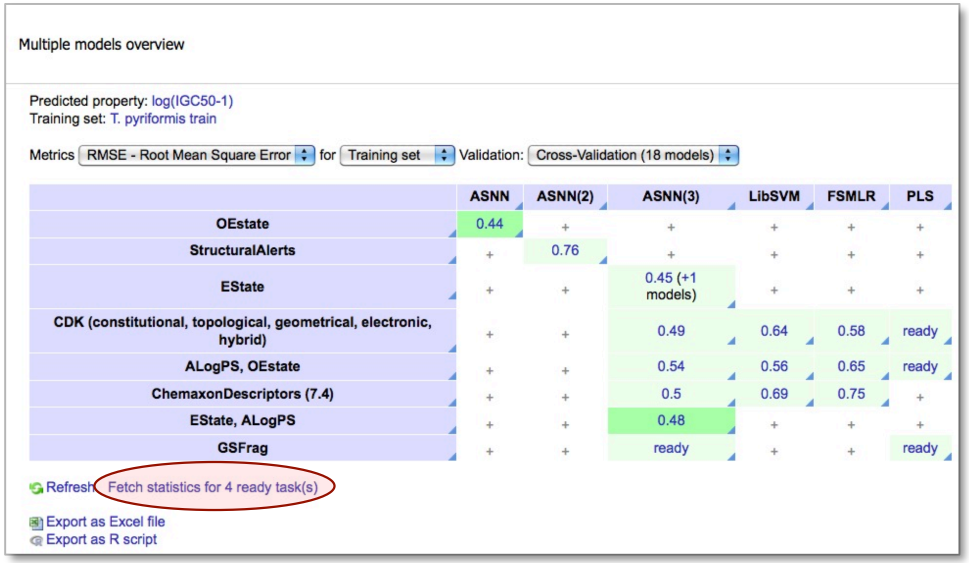
Metrics: RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (18 models)

| | ASNN | ASNN(2) | ASNN(3) | LibSVM | FSMLR | PLS |
|---|---|---|---|---|---|---|
| OEstate | 0.44 | + | + | + | + | + |
| StructuralAlerts | + | 0.76 | + | + | + | + |
| EState | + | + | 0.45 (+1 models) | + | + | + |
| CDK (constitutional, topological, geometrical, electronic, hybrid) | + | + | 0.49 | 0.64 | 0.58 | in queue |
| ALogPS, OEstate | + | + | 0.54 | 0.56 | 0.65 | in queue |
| ChemaxonDescriptors (7.4) | + | + | 0.5 | 0.69 | 0.75 | + |
| EState, ALogPS | + | + | 0.48 | + | + | + |
| GSFrag | + | + | in queue | + | + | in queue |

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

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

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



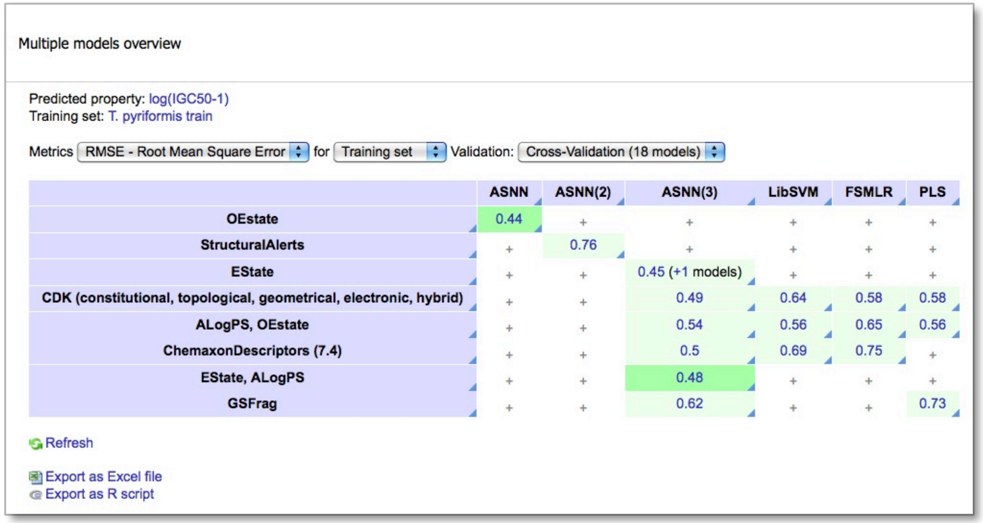
Multiple models overview

Predicted property: log(IGC50-1)
Training set: T. pyriformis train

Metrics: RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (18 models)

| | ASNN | ASNN(2) | ASNN(3) | LibSVM | FSMLR | PLS |
|---|---|---|---|---|---|---|
| OEstate | 0.44 | + | + | + | + | + |
| StructuralAlerts | + | 0.76 | + | + | + | + |
| EState | + | + | 0.45 (+1 models) | + | + | + |
| CDK (constitutional, topological, geometrical, electronic, hybrid) | + | + | 0.49 | 0.64 | 0.58 | ready |
| ALogPS, OEstate | + | + | 0.54 | 0.56 | 0.65 | ready |
| ChemaxonDescriptors (7.4) | + | + | 0.5 | 0.69 | 0.75 | + |
| EState, ALogPS | + | + | 0.48 | + | + | + |
| GSFrag | + | + | ready | + | + | ready |

**9.** To calculate statistics for all the completed models, press the "fetch statistics for ready models".



Multiple models overview

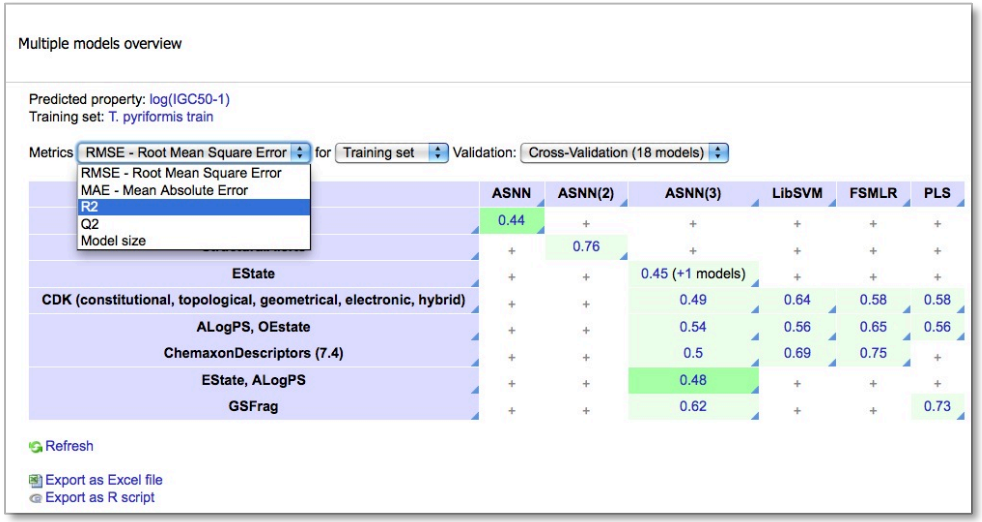
Predicted property: log(IGC50-1)
Training set: T. pyriformis train

Metrics RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (18 models)

| | ASNN | ASNN(2) | ASNN(3) | LibSVM | FSMLR | PLS |
|---|---|---|---|---|---|---|
| OEstate | 0.44 | + | + | + | + | + |
| StructuralAlerts | + | 0.76 | + | + | + | + |
| EState | + | + | 0.45 (+1 models) | + | + | + |
| CDK (constitutional, topological, geometrical, electronic, hybrid) | + | + | 0.49 | 0.64 | 0.58 | 0.58 |
| ALogPS, OEstate | + | + | 0.54 | 0.56 | 0.65 | 0.56 |
| ChemaxonDescriptors (7.4) | + | + | 0.5 | 0.69 | 0.75 | + |
| EState, ALogPS | + | + | 0.48 | + | + | + |
| GSFrag | + | + | 0.62 | + | + | 0.73 |

**10.** We can see all four of our models are 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 (RMSE) than e.g. the PLS models.



Multiple models overview

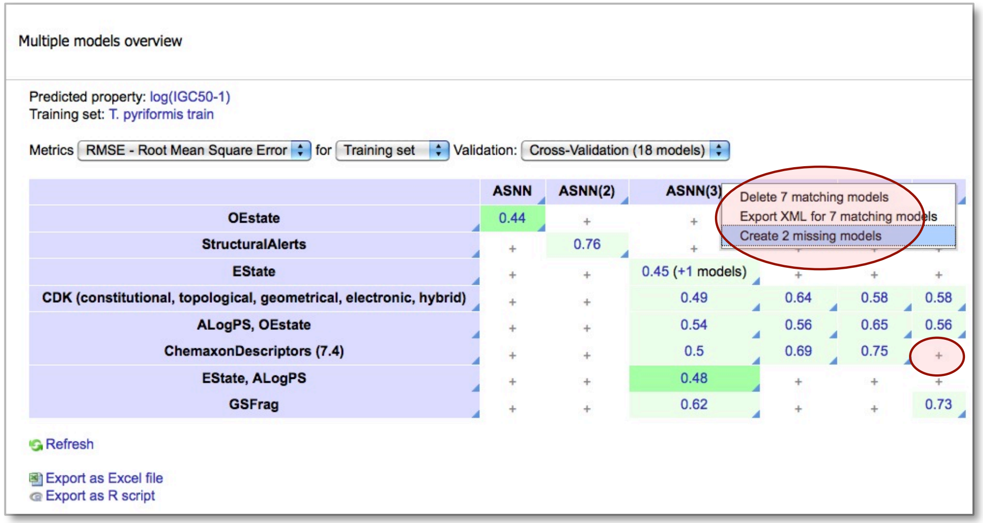
Predicted property: log(IGC50-1)
Training set: T. pyriformis train

Metrics RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (18 models)

| | ASNN | ASNN(2) | ASNN(3) | LibSVM | FSMLR | PLS |
|---|---|---|---|---|---|---|
| | 0.44 | + | + | + | + | + |
| | + | 0.76 | + | + | + | + |
| EState | + | + | 0.45 (+1 models) | + | + | + |
| CDK (constitutional, topological, geometrical, electronic, hybrid) | + | + | 0.49 | 0.64 | 0.58 | 0.58 |
| ALogPS, OEstate | + | + | 0.54 | 0.56 | 0.65 | 0.56 |
| ChemaxonDescriptors (7.4) | + | + | 0.5 | 0.69 | 0.75 | + |
| EState, ALogPS | + | + | 0.48 | + | + | + |
| GSFrag | + | + | 0.62 | + | + | 0.73 |

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

# OCHEM tutorial handout



Multiple models overview

Predicted property: log(IGC50-1)
Training set: T. pyriformis train

Metrics: RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (18 models)

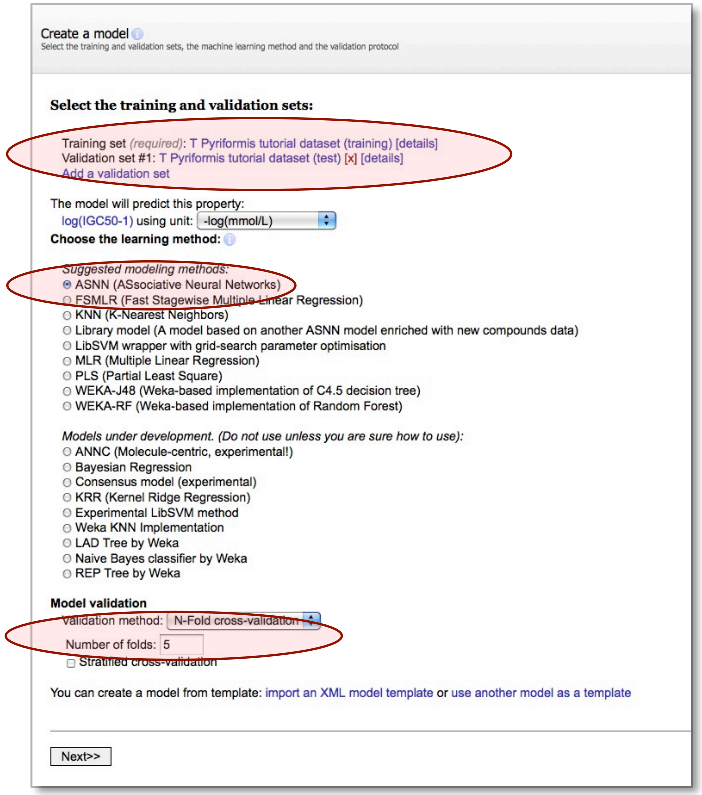
| | ASNN | ASNN(2) | ASNN(3) | | | |
|---|---|---|---|---|---|---|
| OEstate | 0.44 | + | + | | | |
| StructuralAlerts | + | 0.76 | + | | | |
| EState | + | + | 0.45 (+1 models) | + | + | + |
| CDK (constitutional, topological, geometrical, electronic, hybrid) | + | + | 0.49 | 0.64 | 0.58 | 0.58 |
| ALogPS, OEstate | + | + | 0.54 | 0.56 | 0.65 | 0.56 |
| ChemaxonDescriptors (7.4) | + | + | 0.5 | 0.69 | 0.75 | + |
| EState, ALogPS | + | + | 0.48 | + | + | + |
| GSFrag | + | + | 0.62 | + | + | 0.73 |

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

**13.** You also can create new models individually by pushing "+" sign in the "missing" cells.

## 3.2 Development of a single model

### Select training set, machine learning method and internal validation options



Create a model

Select the training and validation sets, the machine learning method and the validation protocol

**Select the training and validation sets:**

Training set *(required)*: T Pyriformis tutorial dataset (training) [details]
Validation set #1: T Pyriformis tutorial dataset (test) [x] [details]
Add a validation set

The model will predict this property:
log(IGC50-1) using unit: -log(mmol/L)

**Choose the learning method:**

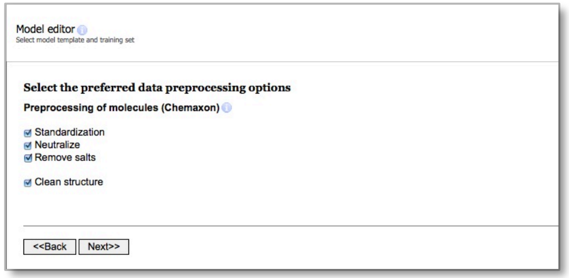
*Suggested modeling methods:*
● ASNN (ASsociative Neural Networks)
○ FSMLR (Fast Stagewise Multiple Linear Regression)
○ KNN (K-Nearest Neighbors)
○ Library model (A model based on another ASNN model enriched with new compounds data)
○ LibSVM wrapper with grid-search parameter optimisation
○ MLR (Multiple Linear Regression)
○ PLS (Partial Least Square)
○ WEKA-J48 (Weka-based implementation of C4.5 decision tree)
○ WEKA-RF (Weka-based implementation of Random Forest)

*Models under development. (Do not use unless you are sure how to use):*
○ ANNC (Molecule-centric, experimental!)
○ Bayesian Regression
○ Consensus model (experimental)
○ KRR (Kernel Ridge Regression)
○ Experimental LibSVM method
○ Weka KNN Implementation
○ LAD Tree by Weka
○ Naive Bayes classifier by Weka
○ REP Tree by Weka

**Model validation**

Validation method: N-Fold cross-validation
Number of folds: 5
□ Stratified cross-validation

You can create a model from template: import an XML model template or use another model as a template



Model editor

Select model template and training set

**Select the preferred data preprocessing options**

**Preprocessing of molecules (Chemaxon)**

☑ Standardization
☑ Neutralize
☑ Remove salts

☑ Clean structure

To start the model development process, open “Model > Create a model” from the menu panel. The first page of the model creation “wizard” asks you to select a training set, external validation sets (optional), a machine learning algorithm and an internal model validation technique.

1. Select the training set and optionally one or more external validation sets that you have prepared before by clicking on the [...] label and the “Add a validation set” link.

2. OCHEM supports two dozen state of the art machine-learning methods. For this tutorial, we will use defaults for most of the configurable options. Thus, we will select associative neural networks (ASNN) to train the model.

3. You can choose between n-fold cross-validation, bagging and no validation at all. A 5-fold cross-validation is most commonly used.

4. Model configurations can also be imported from earlier model building processes to follow the same protocol.

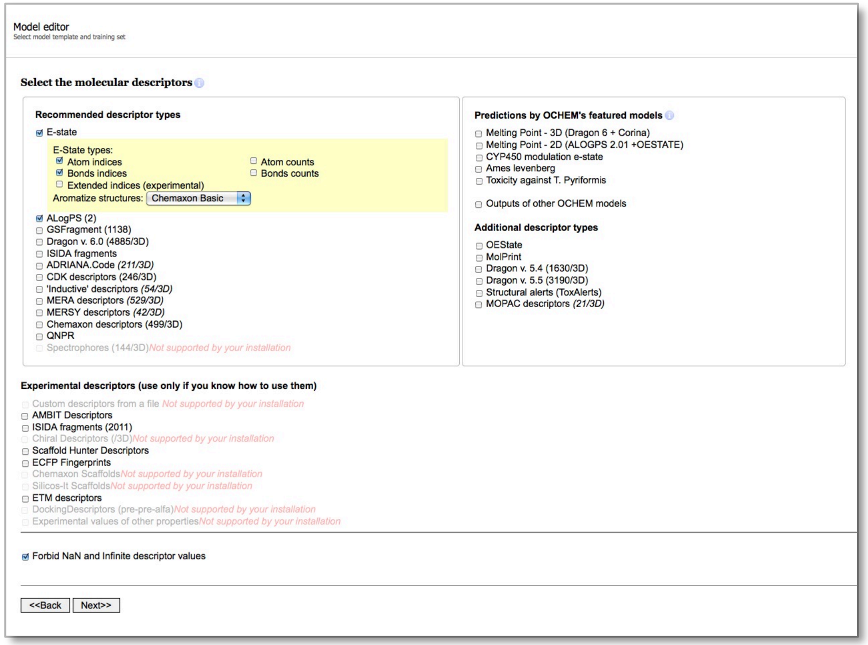
The model creation process is organized as a “wizard” guiding you through the model configuration process. So click “Next” to navigate forward.

Pre-processing of the molecules includes four options: standardization of some chemical groups for consistency, neutralization of ions, removal of salts and cleaning of given meta-information in the structure file.

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

## Configure molecular descriptors

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



Model editor
Select model template and training set

**Select the molecular descriptors**

**Recommended descriptor types**

☑ E-state

E-State types:
☑ Atom indices
☑ Bonds indices
□ Extended indices (experimental)
□ Atom counts
□ Bonds counts
Aromatize structures: Chemaxon Basic

☑ ALogPS (2)
□ GSFragment (1138)
□ Dragon v. 6.0 (4885/3D)
□ ISIDA fragments
□ ADRIANA.Code *(211/3D)*
□ CDK descriptors (246/3D)
□ 'Inductive' descriptors *(54/3D)*
□ MERA descriptors *(529/3D)*
□ MERSY descriptors *(42/3D)*
□ Chemaxon descriptors (499/3D)
□ QNPR
□ Spectrophores (144/3D)*Not supported by your installation*

**Predictions by OCHEM's featured models**

□ Melting Point - 3D (Dragon 6 + Corina)
□ Melting Point - 2D (ALOGPS 2.01 +OESTATE)
□ CYP450 modulation e-state
□ Ames levenberg
□ Toxicity against T. Pyriformis

□ Outputs of other OCHEM models

**Additional descriptor types**

□ OEState
□ MolPrint
□ Dragon v. 5.4 (1630/3D)
□ Dragon v. 5.5 (3190/3D)
□ Structural alerts (ToxAlerts)
□ MOPAC descriptors *(21/3D)*

**Experimental descriptors (use only if you know how to use them)**

□ Custom descriptors from a file *Not supported by your installation*
□ AMBIT Descriptors
□ ISIDA fragments (2011)
□ Chiral Descriptors (/3D)*Not supported by your installation*
□ Scaffold Hunter Descriptors
□ ECFP Fingerprints
□ Chemaxon Scaffolds*Not supported by your installation*
□ Silicos-It Scaffolds*Not supported by your installation*
□ ETM descriptors
□ DockingDescriptors (pre-pre-alfa)*Not supported by your installation*
□ Experimental values of other properties*Not supported by your installation*

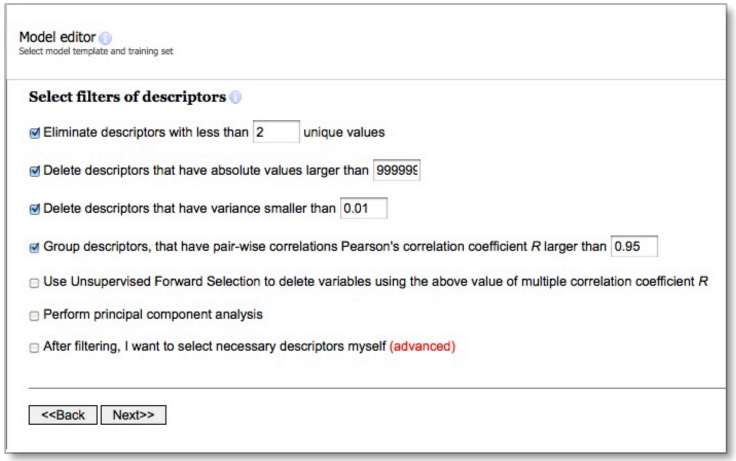
☑ Forbid NaN and Infinite descriptor values

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

Several descriptors and descriptor packages are available on the OCHEM platform, ranging from simple 1D to sophisticated 3D descriptors. If a 3D descriptor is selected, a structure optimization method can be selected in the next step (not shown here).

Furthermore, the output of already existing models can be used as input for the new model to train. I.e. a predicted logP value (if not available) can be used as a molecular descriptor. This functionality is referred to as “feature nets”.

The next dialog allows filtering out redundant and correlated descriptors.



Model editor
Select model template and training set

**Select filters of descriptors**

☑ Eliminate descriptors with less than 2 unique values

☑ Delete descriptors that have absolute values larger than 999999

☑ Delete descriptors that have variance smaller than 0.01

☑ Group descriptors, that have pair-wise correlations Pearson's correlation coefficient *R* larger than 0.95

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

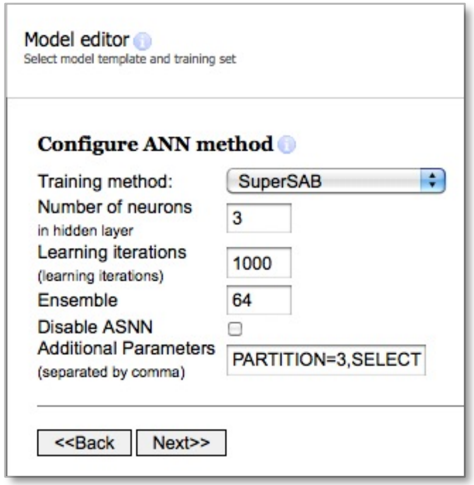
□ Perform principal component analysis

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

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

It is also possible to select the list of desired descriptors manually (advanced).

## Configure the training method and start calculations



Model editor

Select model template and training set

**Configure ANN method**

Training method: SuperSAB

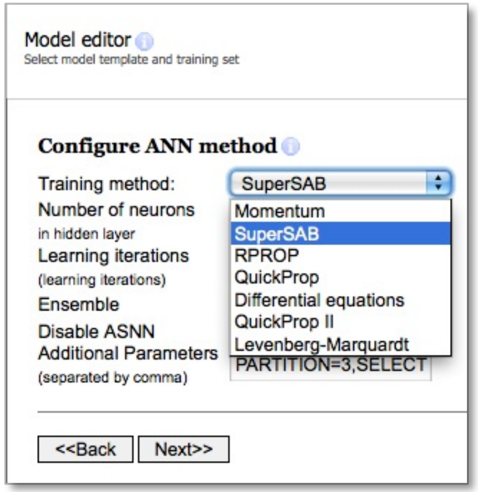
Number of neurons in hidden layer: 3

Learning iterations (learning iterations): 1000

Ensemble: 64

Disable ASNN: □

Additional Parameters (separated by comma): PARTITION=3,SELECT



Model editor

Select model template and training set

**Configure ANN method**

Training method: SuperSAB

Number of neurons in hidden layer

Learning iterations (learning iterations)

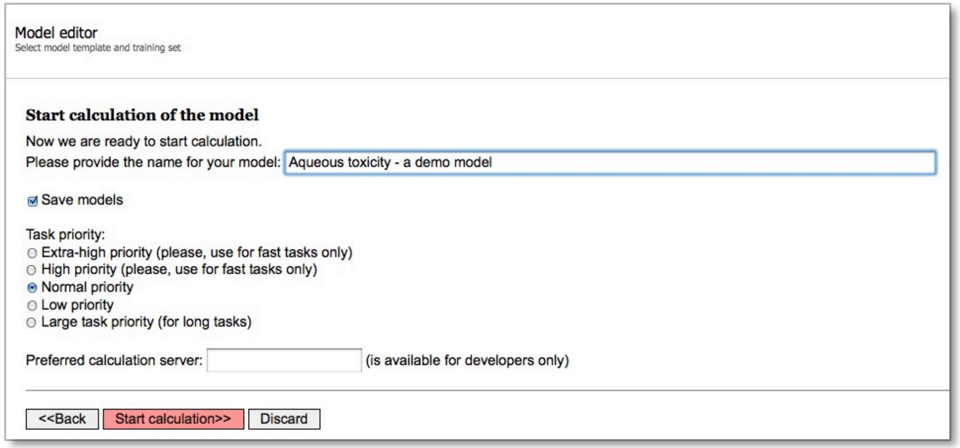
Ensemble

Disable ASNN

Additional Parameters (separated by comma): PARTITION=3,SELECT

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

8. We will not experiment here now and will use the default options, which are often a good starting point.



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: Aqueous toxicity - a demo model

☑ Save models

Task priority:

○ Extra-high priority (please, use for fast tasks only)
○ High priority (please, use for fast tasks only)
● Normal priority
○ Low priority
○ Large task priority (for long tasks)

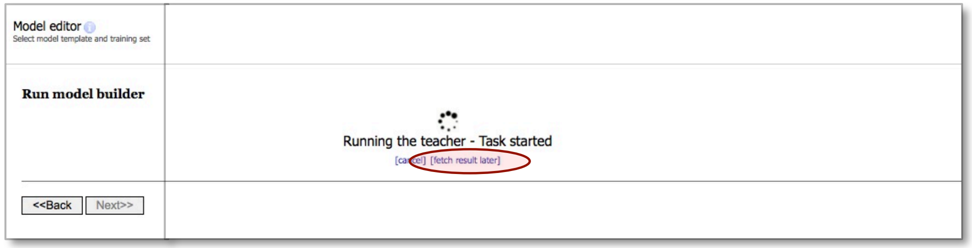
Preferred calculation server: (is available for developers only)

9. Finally, we are ready to start calculations. Before starting, please provide the name for your future model.

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

11. Please click “start calculations” to start the model training process.

## Distributed model calculation



Model editor
Select model template and training set

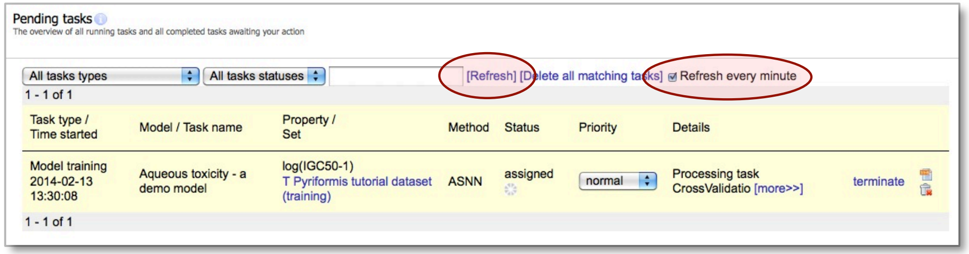
**Run model builder**

Running the teacher - Task started
[cancel] [fetch result later]

<<Back Next>>

A waiting-screen that shows you the status of the calculations.
The training process is automatically distributed to several internal calculation units, but still for large datasets it can take a while to complete (from minutes to weeks).

**12.** Although we could have waited, we will opt to click “fetch result later” to get an overview of currently submitted tasks to the system.



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

All tasks types | All tasks statuses | [Refresh] [Delete all matching tasks] ☑ Refresh every minute

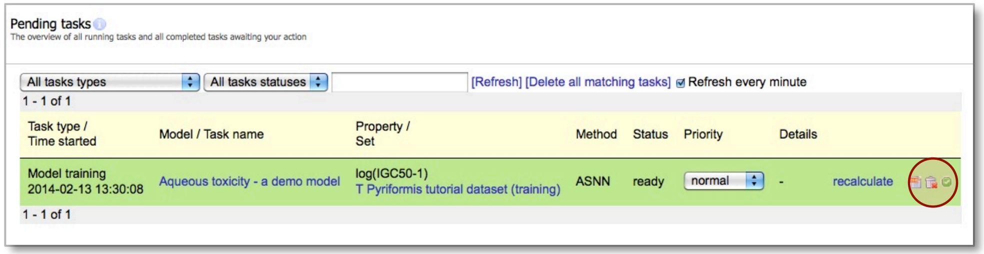
1 - 1 of 1

| Task type /<br>Time started | Model / Task name | Property /<br>Set | Method | Status | Priority | Details | |
|---|---|---|---|---|---|---|---|
| Model training<br>2014-02-13<br>13:30:08 | Aqueous toxicity - a demo model | log(IGC50-1)<br>T Pyriformis tutorial dataset (training) | ASNN | assigned | normal | Processing task CrossValidatio [more>>] | 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 tasks that are currently running on the system or have been finished, but not yet fetched by the user.

**13.** Here you can observe the status, terminate running tasks or fetch ready tasks. Please, click “refresh” or check the box for Refresh every minute” to actualize the page.



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

All tasks types | All tasks statuses | [Refresh] [Delete all matching tasks] ☑ Refresh every minute

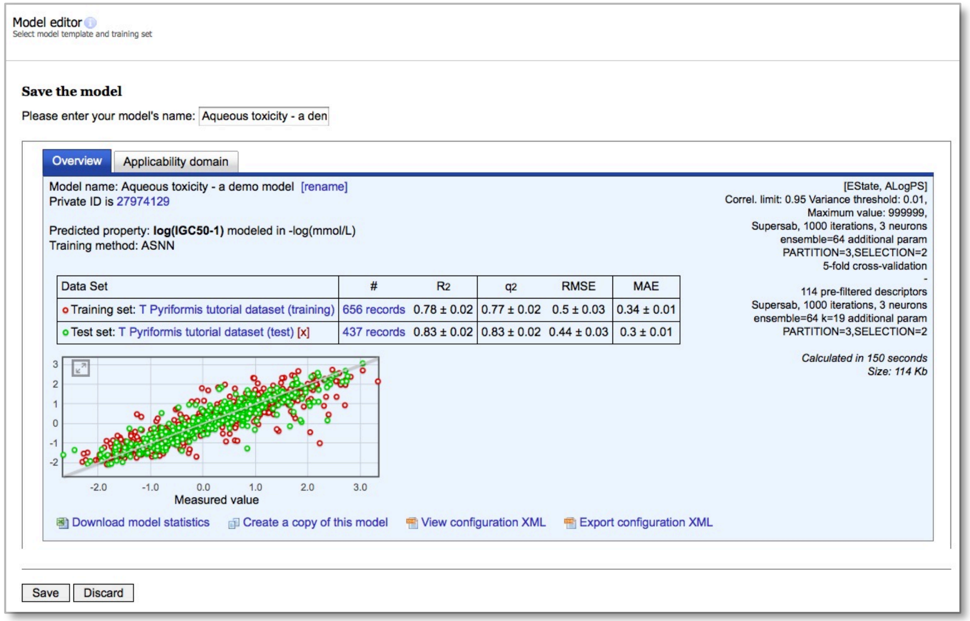
1 - 1 of 1

| Task type /<br>Time started | Model / Task name | Property /<br>Set | Method | Status | Priority | Details | |
|---|---|---|---|---|---|---|---|
| Model training<br>2014-02-13 13:30:08 | Aqueous toxicity - a demo model | log(IGC50-1)<br>T Pyriformis tutorial dataset (training) | ASNN | ready | normal | - | recalculate |

1 - 1 of 1

**14.** When the task has finished, please click the green check button or the model name link to fetch the model and investigate the statistics of the model.

## Save your model



Model editor

Select model template and training set

**Save the model**

Please enter your model's name: Aqueous toxicity - a den

Model name: Aqueous toxicity - a demo model [rename]
Private ID is 27974129

Predicted property: **log(IGC50-1)** modeled in -log(mmol/L)
Training method: ASNN

| Data Set | # | R2 | q2 | RMSE | MAE |
|---|---|---|---|---|---|
| ○ Training set: T Pyriformis tutorial dataset (training) | 656 records | 0.78 ± 0.02 | 0.77 ± 0.02 | 0.5 ± 0.03 | 0.34 ± 0.01 |
| ○ Test set: T Pyriformis tutorial dataset (test) [x] | 437 records | 0.83 ± 0.02 | 0.83 ± 0.02 | 0.44 ± 0.03 | 0.3 ± 0.01 |

[EState, ALogPS]
Correl. limit: 0.95 Variance threshold: 0.01,
Maximum value: 999999,
Supersab, 1000 iterations, 3 neurons
ensemble=64 additional param
PARTITION=3,SELECTION=2
5-fold cross-validation
-
114 pre-filtered descriptors
Supersab, 1000 iterations, 3 neurons
ensemble=64 k=19 additional param
PARTITION=3,SELECTION=2

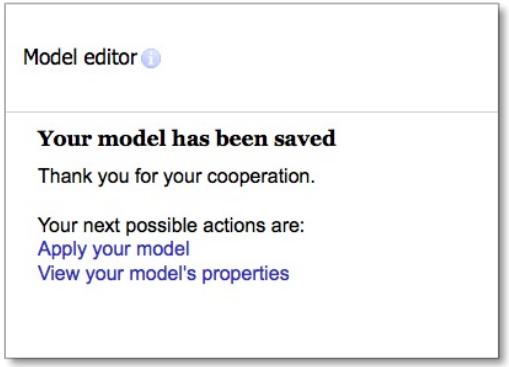
*Calculated in 150 seconds*
*Size: 114 Kb*

If the calculation was successful, you can see the profile of the ready model. Before saving the model the profile can be investigated further:

**15.** The model profile shows information about the training configuration (used descriptors, machine-learning method and predicted property. It shows the training process statistics like the training set size and correlation coefficients (R2, Q2) and deviation measures of the predicted values to the observed values (RMSE, MAE).
On the interactive plot training results of single structures can be inspected. E.g. the calculated descriptor values and the predicted value.

This important dialog is explained in more detail in a chapter on its own. For now please save your model with a meaningful name.

You have successfully built a prediction model on the OCHEM platform



Model editor

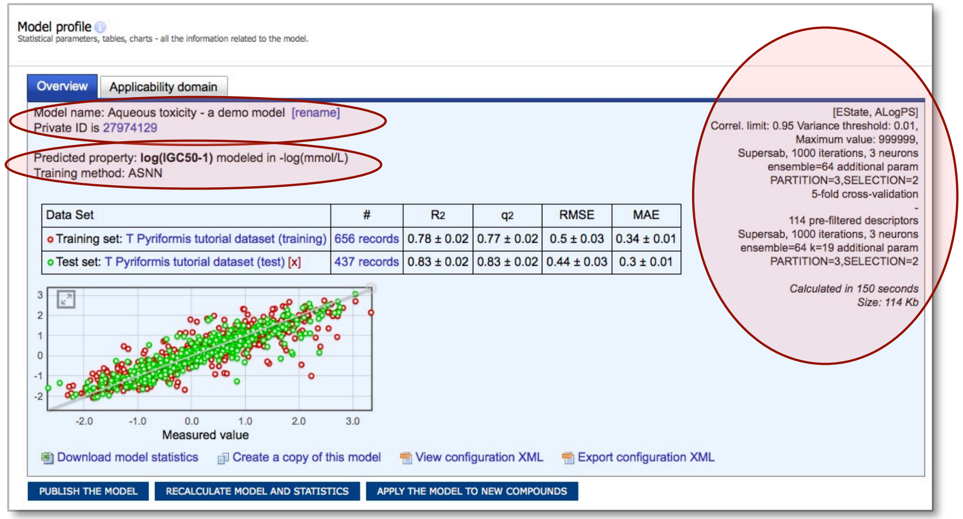
**Your model has been saved**

Thank you for your cooperation.

Your next possible actions are:
Apply your model
View your model's properties

**16.** After saving the model, you can directly apply it to predict new compounds. Before running predictions, we will investigate the model profile page in more detail.

## 3.3 Model profile



Model profile
Statistical parameters, tables, charts - all the information related to the model.

Model name: Aqueous toxicity - a demo model [rename]
Private ID is 27974129

Predicted property: **log(IGC50-1)** modeled in -log(mmol/L)
Training method: ASNN

| Data Set | # | R2 | q2 | RMSE | MAE |
|---|---|---|---|---|---|
| ○ Training set: T Pyriformis tutorial dataset (training) | 656 records | 0.78 ± 0.02 | 0.77 ± 0.02 | 0.5 ± 0.03 | 0.34 ± 0.01 |
| ○ Test set: T Pyriformis tutorial dataset (test) [x] | 437 records | 0.83 ± 0.02 | 0.83 ± 0.02 | 0.44 ± 0.03 | 0.3 ± 0.01 |

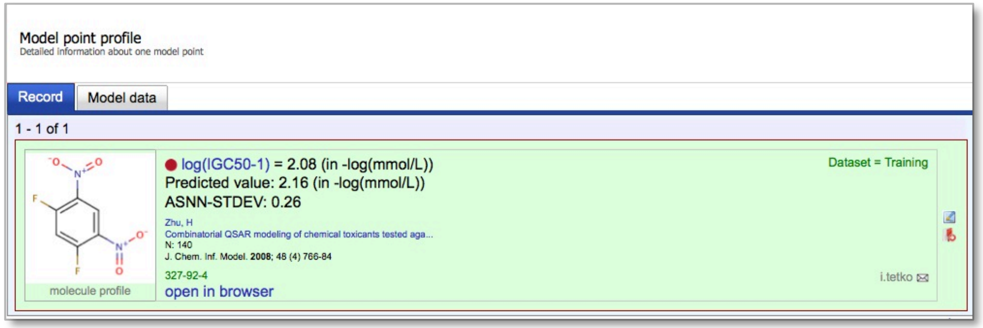
[EState, ALogPS]
Correl. limit: 0.95 Variance threshold: 0.01,
Maximum value: 999999,
Supersab, 1000 iterations, 3 neurons
ensemble=64 additional param
PARTITION=3,SELECTION=2
5-fold cross-validation
-
114 pre-filtered descriptors
Supersab, 1000 iterations, 3 neurons
ensemble=64 k=19 additional param
PARTITION=3,SELECTION=2

*Calculated in 150 seconds*
*Size: 114 Kb*

The model profile contains all the information related to the performance of the model: **statistical parameters**, **interactive scatter plot** for the single structures, links to the **data sets** used for the model and various operations, like export of the model, application of the model to new compounds, etc.

There is information about the **model name** and **public visibility**, the **predicted property** and the **training method**, used **descriptors** and **validation method**, **training duration** and **model size.**

Typical model statistics are shown in a summary table **(5)**. There are the data sets (training set and given external validation sets) with direct links to their profiles or the records respectively showing the set size. For each set the coefficient of determination „**$R^2$**“, cross-validated $R^2$ called „**$Q^2$**“, root mean squared error „**RMSE**“ and mean absolute error „**MAE**“.



Model point profile
Detailed information about one model point

1 - 1 of 1

● log(IGC50-1) = 2.08 (in -log(mmol/L))
Predicted value: 2.16 (in -log(mmol/L))
ASNN-STDEV: 0.26
Zhu, H
Combinatorial QSAR modeling of chemical toxicants tested aga...
N: 140
J. Chem. Inf. Model. **2008**; 48 (4) 766-84
327-92-4
open in browser

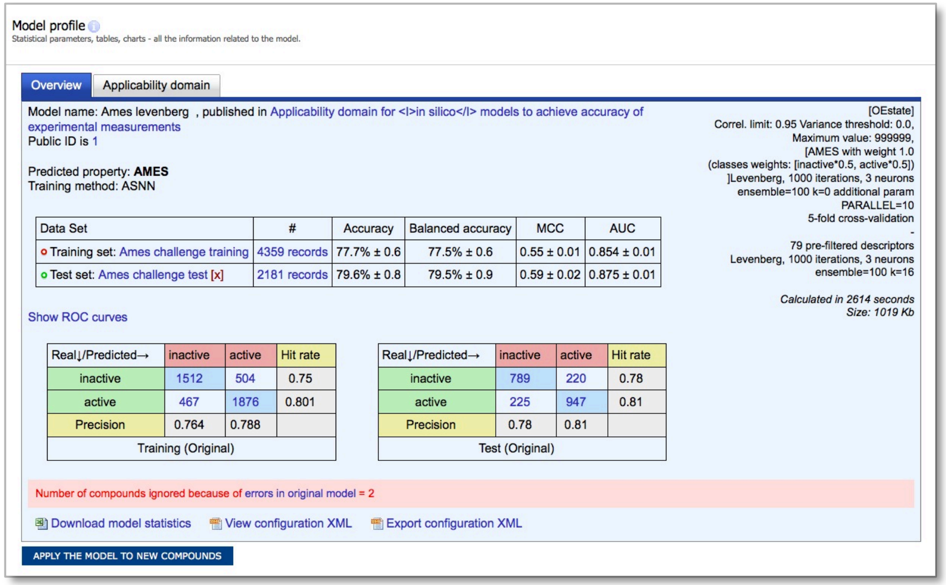
Dataset = Training

Furthermore with the interactive scatterplot, showing predicted versus measured values, single records can be further investigated, e.g. comparison of predicted and measured value or inspection of calculated descriptor values for this structure.

Each point on the scatter plot is clickable and will open the “model point profile” 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.

# OCHEM tutorial handout



Model profile

Statistical parameters, tables, charts - all the information related to the model.

Model name: Ames levenberg , published in Applicability domain for <I>in silico</I> models to achieve accuracy of experimental measurements
Public ID is 1

Predicted property: **AMES**
Training method: ASNN

[OEstate]
Correl. limit: 0.95 Variance threshold: 0.0,
Maximum value: 999999,
[AMES with weight 1.0
(classes weights: [inactive*0.5, active*0.5])
]Levenberg, 1000 iterations, 3 neurons
ensemble=100 k=0 additional param
PARALLEL=10
5-fold cross-validation
-
79 pre-filtered descriptors
Levenberg, 1000 iterations, 3 neurons
ensemble=100 k=16

*Calculated in 2614 seconds*
*Size: 1019 Kb*

| Data Set | # | Accuracy | Balanced accuracy | MCC | AUC |
|---|---|---|---|---|---|
| ○ Training set: Ames challenge training | 4359 records | 77.7% ± 0.6 | 77.5% ± 0.6 | 0.55 ± 0.01 | 0.854 ± 0.01 |
| ○ Test set: Ames challenge test [x] | 2181 records | 79.6% ± 0.8 | 79.5% ± 0.9 | 0.59 ± 0.02 | 0.875 ± 0.01 |

Show ROC curves

| Real↓/Predicted→ | inactive | active | Hit rate |
|---|---|---|---|
| inactive | 1512 | 504 | 0.75 |
| active | 467 | 1876 | 0.801 |
| Precision | 0.764 | 0.788 | |

Training (Original)

| Real↓/Predicted→ | inactive | active | Hit rate |
|---|---|---|---|
| inactive | 789 | 220 | 0.78 |
| active | 225 | 947 | 0.81 |
| Precision | 0.78 | 0.81 | |

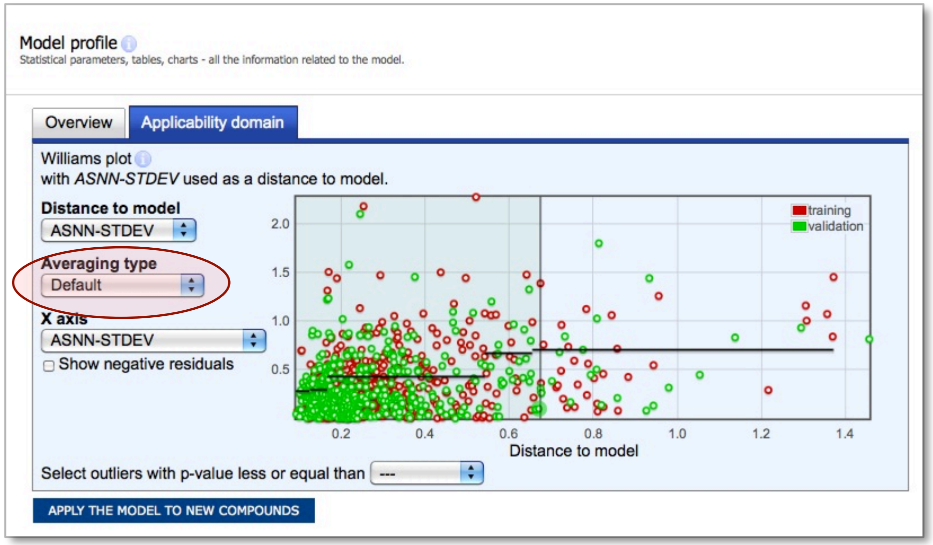
Test (Original)

Number of compounds ignored because of errors in original model = 2

For classification models, the model profile shows different statistical parameters. These are:

- accuracy, balanced accuracy, MCC, AUC,
- ROC curves
- Confusion matrices, where you can see the number of false positives, false negatives and so on.

## 3.4 Applicability domain



Model profile

Statistical parameters, tables, charts - all the information related to the model.

Overview | Applicability domain

Williams plot

with *ASNN-STDEV* used as a distance to model.

**Distance to model**: ASNN-STDEV

**Averaging type**: Default

**X axis**: ASNN-STDEV

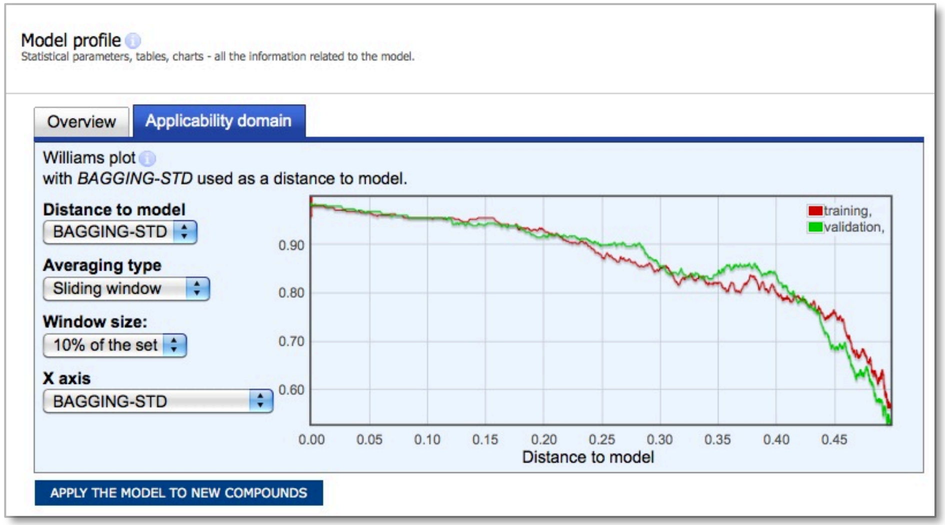
□ Show negative residuals

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

Each prediction in OCHEM is complemented with an accuracy estimate.

The key concept used for the accuracy estimation is so called **distance to model** (DM). DM is any measure of prediction uncertainty correlated with the prediction accuracy.

Usually, the prediction accuracy falls as DM grows, which is shown on so called **accuracy averaging plots** (shown on the left). The accuracy can be averaged in several manners: via bin-based averaging (used for regression models) and sliding window averaging (used for classification models).



Model profile

Statistical parameters, tables, charts - all the information related to the model.

Overview | Applicability domain

Williams plot

with *BAGGING-STD* used as a distance to model.

**Distance to model**: BAGGING-STD

**Averaging type**: Sliding window

**Window size:** 10% of the set

**X axis**: BAGGING-STD

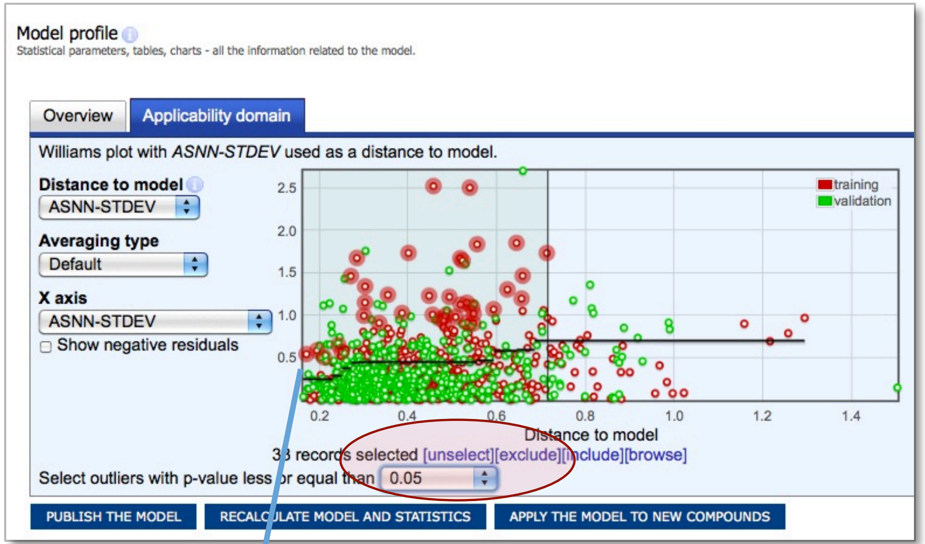
With the Applicability Domain (AD) user interface different display options of the AD can be selected. There is the “Distance to model” type, the “averaging type”, sliding “window size” and the label show at the x-axis.

## Smart outlier detection using AD plots

A practical feature of the applicability plots with their estimated distance to model is the automatic outlier selection according to a significance range of the p-value.

Since one measure of the distance to model follows a normal distribution, certain records can be selected and excluded from the training set.

Sometimes these “outliers” are not easily detectable as such just by inspecting the scatterplot.



Model profile

Statistical parameters, tables, charts - all the information related to the model.

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

Distance to model: ASNN-STDEV

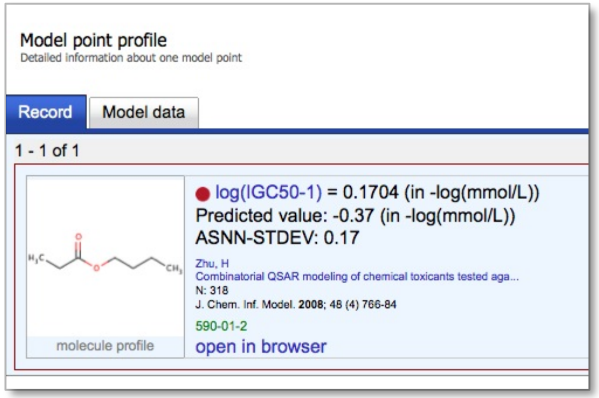
Averaging type: Default

X axis: ASNN-STDEV

□ Show negative residuals

38 records selected [unselect][exclude][include][browse]

Select outliers with p-value less or equal than 0.05



Model point profile

Detailed information about one model point

1 - 1 of 1

● log(IGC50-1) = 0.1704 (in -log(mmol/L))

Predicted value: -0.37 (in -log(mmol/L))

ASNN-STDEV: 0.17

Zhu, H

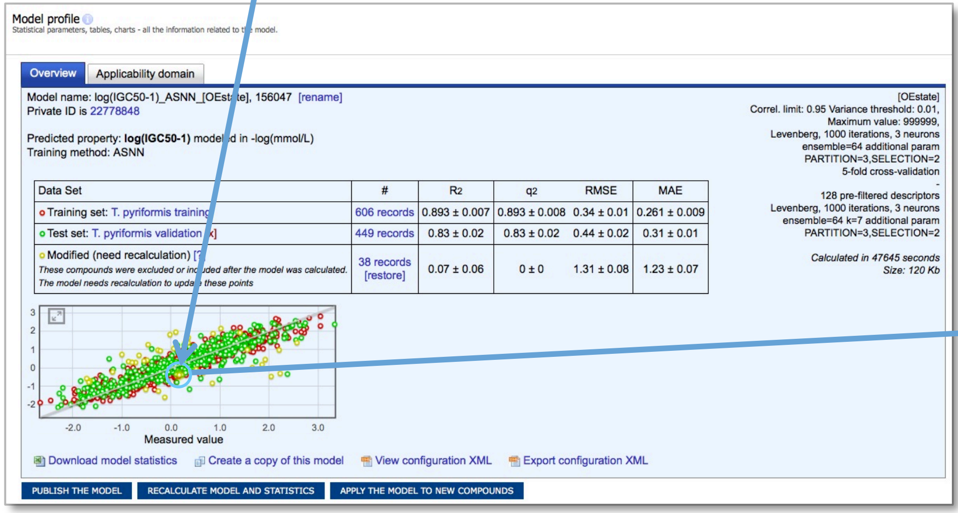
Combinatorial QSAR modeling of chemical toxicants tested aga...

N: 318

J. Chem. Inf. Model. **2008**; 48 (4) 766-84

590-01-2

open in browser



Model profile

Statistical parameters, tables, charts - all the information related to the model.

Model name: log(IGC50-1)_ASNN_[OEstate], 156047 [rename]

Private ID is 22778848

Predicted property: **log(IGC50-1)** modeled in -log(mmol/L)

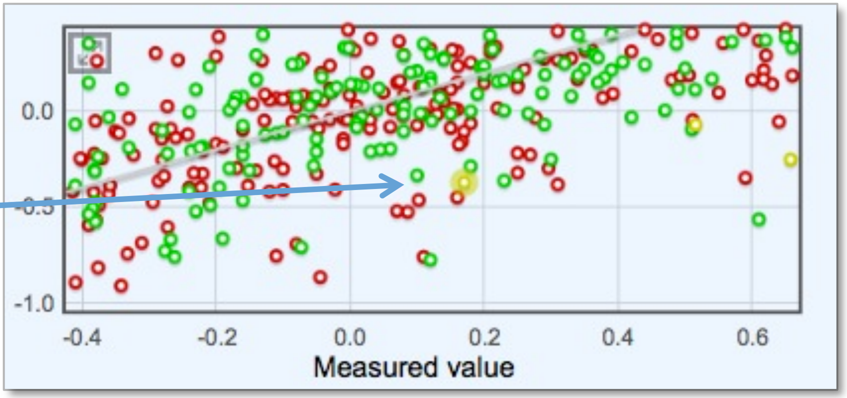
Training method: ASNN

| Data Set | # | R2 | q2 | RMSE | MAE |
|---|---|---|---|---|---|
| ○ Training set: T. pyriformis training | 606 records | 0.893 ± 0.007 | 0.893 ± 0.008 | 0.34 ± 0.01 | 0.261 ± 0.009 |
| ○ Test set: T. pyriformis validation [x] | 449 records | 0.83 ± 0.02 | 0.83 ± 0.02 | 0.44 ± 0.02 | 0.31 ± 0.01 |
| ○ Modified (need recalculation) [?]<br>*These compounds were excluded or included after the model was calculated. The model needs recalculation to update these points* | 38 records<br>[restore] | 0.07 ± 0.06 | 0 ± 0 | 1.31 ± 0.08 | 1.23 ± 0.07 |

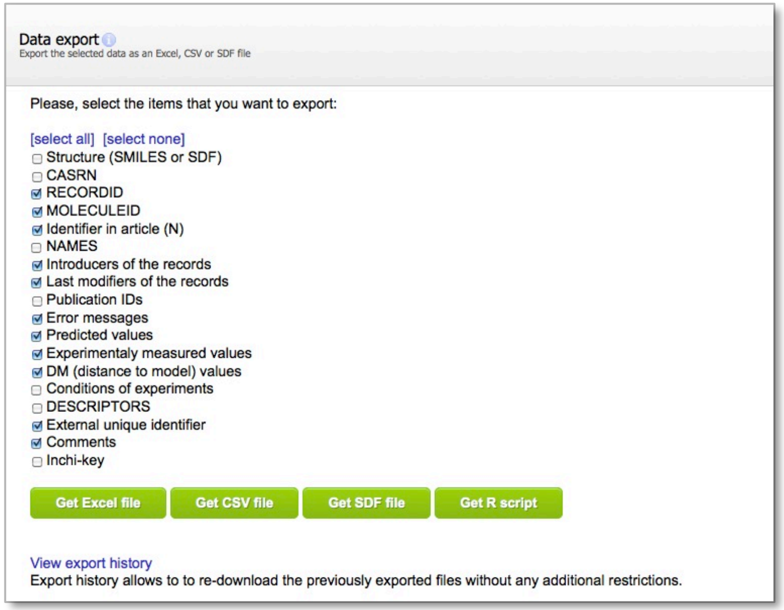
[OEstate]<br>
Correl. limit: 0.95 Variance threshold: 0.01,<br>
Maximum value: 999999,<br>
Levenberg, 1000 iterations, 3 neurons<br>
ensemble=64 additional param<br>
PARTITION=3,SELECTION=2<br>
5-fold cross-validation<br>
-<br>
128 pre-filtered descriptors<br>
Levenberg, 1000 iterations, 3 neurons<br>
ensemble=64 k=7 additional param<br>
PARTITION=3,SELECTION=2

*Calculated in 47645 seconds*

*Size: 120 Kb*



## Model export



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
☑ Experimentaly measured values
☑ DM (distance to model) values
□ Conditions of experiments
□ DESCRIPTORS
☑ External unique identifier
☑ Comments
□ Inchi-key

View export history
Export history allows to to re-download the previously exported files without any additional restrictions.

1. It is possible to export the data related to your model by clicking “Download model statistics in Excel format”. 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.

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

# 4 Working with data

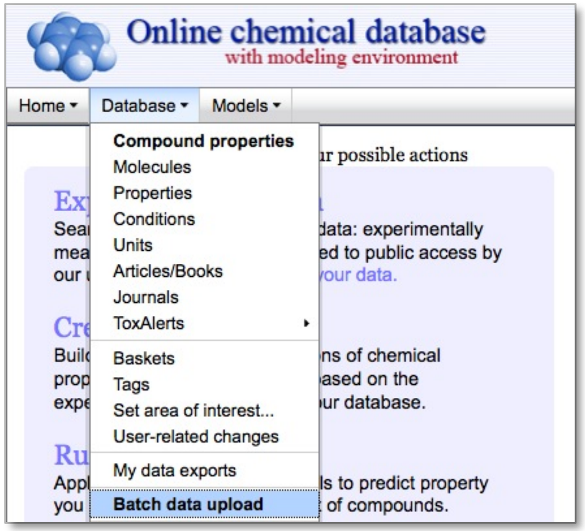
In this chapter, we learn how to upload data using

- Batch data upload
- Advanced basket management

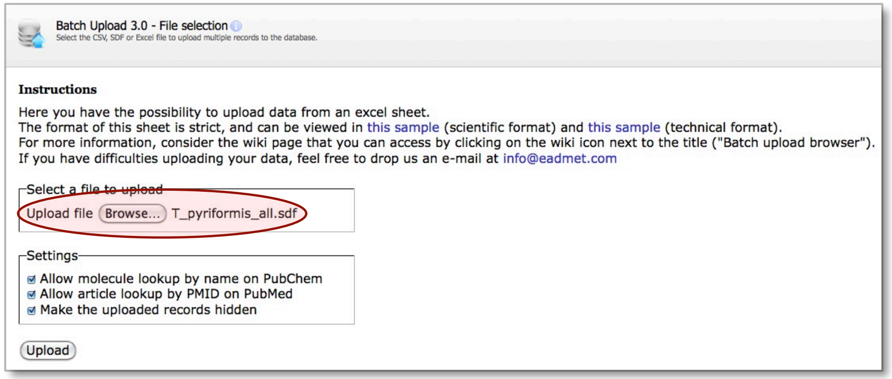
## 4.1 Batch data upload

Although you can introduce each record individually, this is usually not practically feasible. Instead, it is convenient to upload hundreds of thousands of records from external files, e.g. Excel or SD files. This can be done using the “Batch data upload” utility.

In this tutorial, we are going to upload about a thousand records for aquatic toxicity (namely, growth inhibition concentration for T. pyriformis).



1. Select the “Batch data upload” item in the “Database” submenu of the main OCHEM menu. You will open the first page of the “Batch upload wizard”.



Batch Upload 3.0 - File selection
Select the CSV, SDF or Excel file to upload multiple records to the database.

**Instructions**

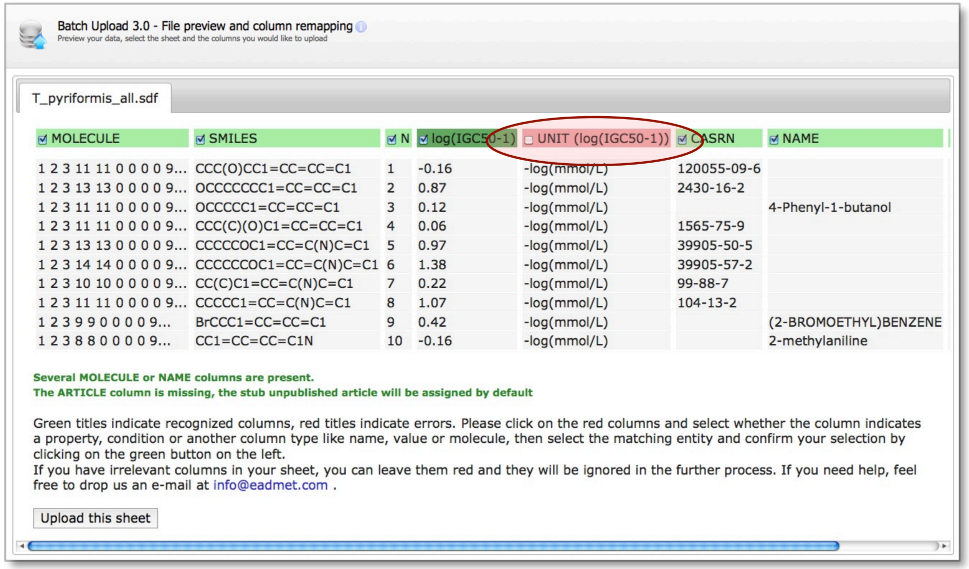
Here you have the possibility to upload data from an excel sheet.
The format of this sheet is strict, and can be viewed in this sample (scientific format) and this sample (technical format).
For more information, consider the wiki page that you can access by clicking on the wiki icon next to the title ("Batch upload browser").
If you have difficulties uploading your data, feel free to drop us an e-mail at info@eadmet.com

Select a file to upload
Upload file (Browse...) T_pyriformis_all.sdf

Settings
☑ Allow molecule lookup by name on PubChem
☑ Allow article lookup by PMID on PubMed
☑ Make the uploaded records hidden

(Upload)

2. Select your provided SD-file in the “Upload file” field. The tool supports SDF and XLS file formats. To remind you, this file contains about 1,000 measured values for the growth inhibition assay.
3. Make sure you select “make the uploaded records hidden” to avoid data conflicts with the other course participants.
4. Hit “Upload” to continue.



Batch Upload 3.0 - File preview and column remapping

Preview your data, select the sheet and the columns you would like to upload

T_pyriformis_all.sdf

| ☑ MOLECULE | ☑ SMILES | ☑ N | ☑ log(IGC50-1) | □ UNIT (log(IGC50-1)) | ☑ CASRN | ☑ NAME |
|---|---|---|---|---|---|---|
| 1 2 3 11 11 0 0 0 0 9... | CCC(O)CC1=CC=CC=C1 | 1 | -0.16 | -log(mmol/L) | 120055-09-6 | |
| 1 2 3 13 13 0 0 0 0 9... | OCCCCCCC1=CC=CC=C1 | 2 | 0.87 | -log(mmol/L) | 2430-16-2 | |
| 1 2 3 11 11 0 0 0 0 9... | OCCCCC1=CC=CC=C1 | 3 | 0.12 | -log(mmol/L) | | 4-Phenyl-1-butanol |
| 1 2 3 11 11 0 0 0 0 9... | CCC(C)(O)C1=CC=CC=C1 | 4 | 0.06 | -log(mmol/L) | 1565-75-9 | |
| 1 2 3 13 13 0 0 0 0 9... | CCCCCOC1=CC=C(N)C=C1 | 5 | 0.97 | -log(mmol/L) | 39905-50-5 | |
| 1 2 3 14 14 0 0 0 0 9... | CCCCCCOC1=CC=C(N)C=C1 | 6 | 1.38 | -log(mmol/L) | 39905-57-2 | |
| 1 2 3 10 10 0 0 0 0 9... | CC(C)C1=CC=C(N)C=C1 | 7 | 0.22 | -log(mmol/L) | 99-88-7 | |
| 1 2 3 11 11 0 0 0 0 9... | CCCCC1=CC=C(N)C=C1 | 8 | 1.07 | -log(mmol/L) | 104-13-2 | |
| 1 2 3 9 9 0 0 0 0 9... | BrCCC1=CC=CC=C1 | 9 | 0.42 | -log(mmol/L) | | (2-BROMOETHYL)BENZENE |
| 1 2 3 8 8 0 0 0 0 9... | CC1=CC=CC=C1N | 10 | -0.16 | -log(mmol/L) | | 2-methylaniline |

**Several MOLECULE or NAME columns are present.**

**The ARTICLE column is missing, the stub unpublished article will be assigned by default**

Green titles indicate recognized columns, red titles indicate errors. Please click on the red columns and select whether the column indicates a property, condition or another column type like name, value or molecule, then select the matching entity and confirm your selection by clicking on the green button on the left.

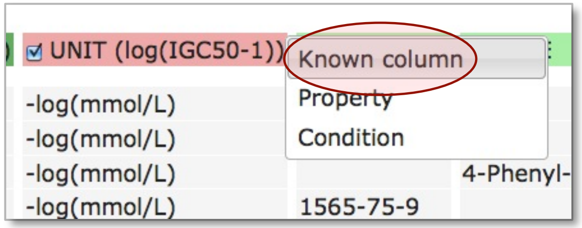
If you have irrelevant columns in your sheet, you can leave them red and they will be ignored in the further process. If you need help, feel free to drop us an e-mail at info@eadmet.com .

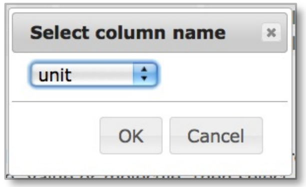
5. The second page of the wizard is the file review page with "column remapping" tool. Here you can preview the first few lines of your uploaded file and see which columns were recognized by the system. On this page, you also have the possibility to reassign column names and select/deselect columns for upload.

**Note:**

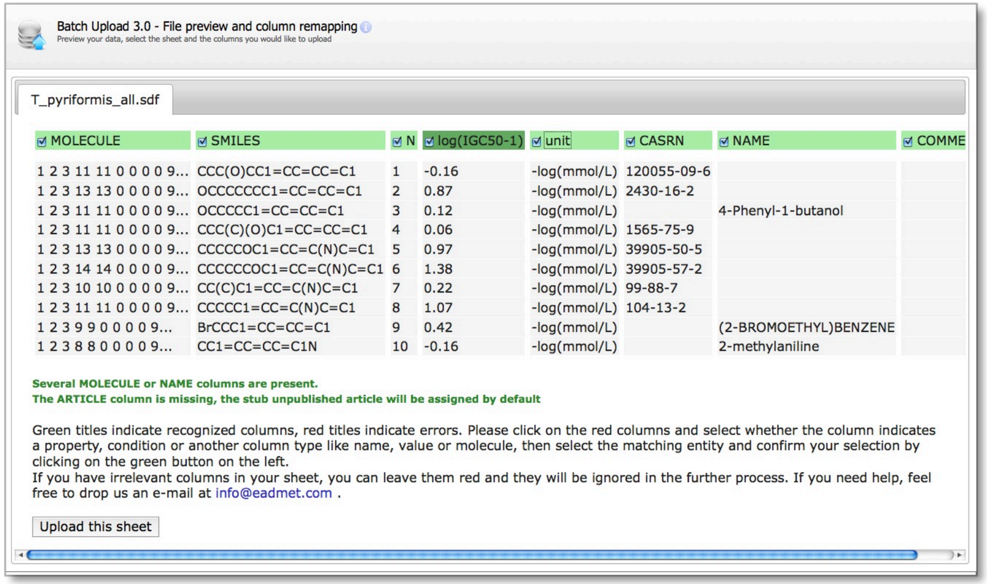
Column headers are colour coded. Green means recognized by the system, red means not recognized. The property is dark green if it is already in the system.

Columns can be remapped by clicking on the column header





6. For example, the column holding the data values is named "UNIT …" in the uploaded file. We need to specify that these values represent the unit for the "Aqueous toxicity" property. Click on the red unrecognized "Unit …" column header and select "Known column" from the popup menu. Then select unit from list of known columns.



Batch Upload 3.0 - File preview and column remapping

Preview your data, select the sheet and the columns you would like to upload

T_pyriformis_all.sdf

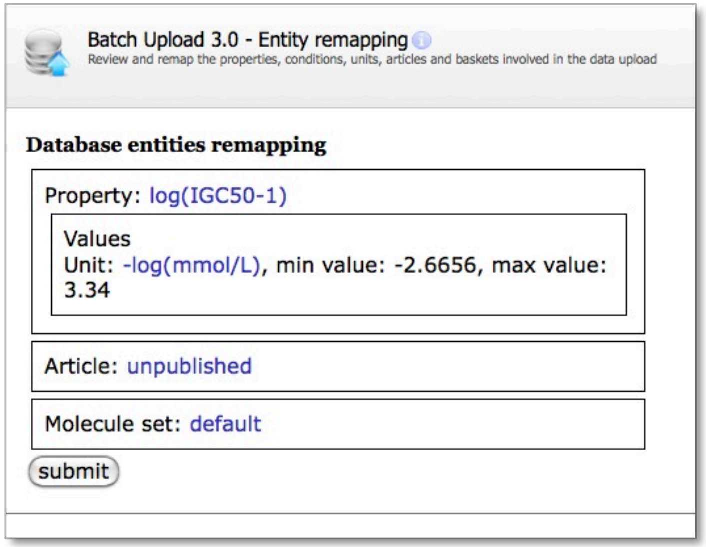
| ☑ MOLECULE | ☑ SMILES | ☑ N | ☑ log(IGC50-1) | ☑ unit | ☑ CASRN | ☑ NAME | ☑ COMME |
|---|---|---|---|---|---|---|---|
| 1 2 3 11 11 0 0 0 0 9... | CCC(O)CC1=CC=CC=C1 | 1 | -0.16 | -log(mmol/L) | 120055-09-6 | | |
| 1 2 3 13 13 0 0 0 0 9... | OCCCCCCC1=CC=CC=C1 | 2 | 0.87 | -log(mmol/L) | 2430-16-2 | | |
| 1 2 3 11 11 0 0 0 0 9... | OCCCCC1=CC=CC=C1 | 3 | 0.12 | -log(mmol/L) | | 4-Phenyl-1-butanol | |
| 1 2 3 11 11 0 0 0 0 9... | CCC(C)(O)C1=CC=CC=C1 | 4 | 0.06 | -log(mmol/L) | 1565-75-9 | | |
| 1 2 3 13 13 0 0 0 0 9... | CCCCCOC1=CC=C(N)C=C1 | 5 | 0.97 | -log(mmol/L) | 39905-50-5 | | |
| 1 2 3 14 14 0 0 0 0 9... | CCCCCCOC1=CC=C(N)C=C1 | 6 | 1.38 | -log(mmol/L) | 39905-57-2 | | |
| 1 2 3 10 10 0 0 0 0 9... | CC(C)C1=CC=C(N)C=C1 | 7 | 0.22 | -log(mmol/L) | 99-88-7 | | |
| 1 2 3 11 11 0 0 0 0 9... | CCCCC1=CC=C(N)C=C1 | 8 | 1.07 | -log(mmol/L) | 104-13-2 | | |
| 1 2 3 9 9 0 0 0 0 9... | BrCCC1=CC=CC=C1 | 9 | 0.42 | -log(mmol/L) | | (2-BROMOETHYL)BENZENE | |
| 1 2 3 8 8 0 0 0 0 9... | CC1=CC=CC=C1N | 10 | -0.16 | -log(mmol/L) | | 2-methylaniline | |

**Several MOLECULE or NAME columns are present.**
**The ARTICLE column is missing, the stub unpublished article will be assigned by default**

Green titles indicate recognized columns, red titles indicate errors. Please click on the red columns and select whether the column indicates a property, condition or another column type like name, value or molecule, then select the matching entity and confirm your selection by clicking on the green button on the left.
If you have irrelevant columns in your sheet, you can leave them red and they will be ignored in the further process. If you need help, feel free to drop us an e-mail at info@eadmet.com .

7. Note that the column header has changed from red to green (recognized unit), the header name is now just unit, and the checkbox in the column header is checked, indicating that the column will be processed by the tool.

8. Click the “Upload this sheet” button to proceed to page three of the wizard.



Batch Upload 3.0 - Entity remapping

Review and remap the properties, conditions, units, articles and baskets involved in the data upload

**Database entities remapping**

Property: log(IGC50-1)

Values
Unit: -log(mmol/L), min value: -2.6656, max value: 3.34

Article: unpublished

Molecule set: default

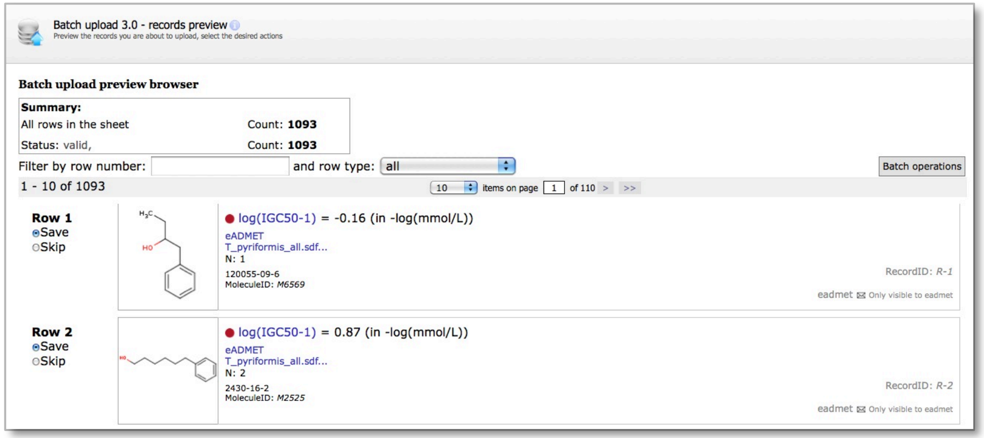
9. The third page of the wizard is the “entity remapping” page. You can review and change some aspects of the uploaded data (property, unit used for data upload, article, etc.)

10. Since no article has been specified in the data sheet, a stub “unpublished” was put instead of the article. Final corrections can be done here, e.g. correcting the unit, selection of a certain article or renaming of the basket.

11. With “submit” 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).

**Note**:

To upload data originally published in an article click on the “Unpublished” link in the “Article” section of the page.



Batch upload 3.0 - records preview
Preview the records you are about to upload, select the desired actions

**Batch upload preview browser**

**Summary:**
All rows in the sheet — Count: **1093**
Status: valid, — Count: **1093**

Filter by row number: and row type: all — Batch operations

1 - 10 of 1093 — 10 items on page 1 of 110 > >>

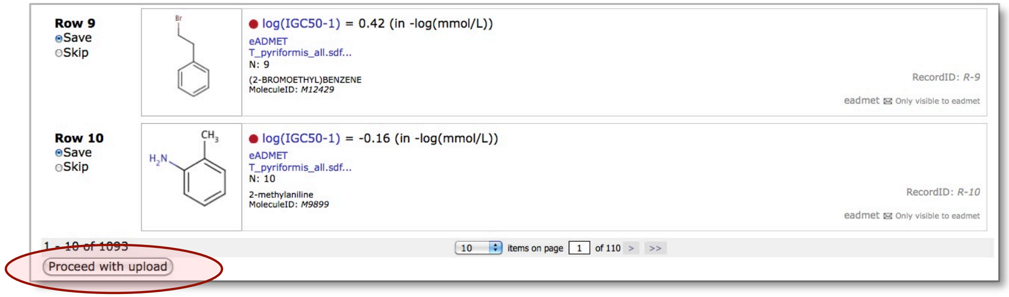
**Row 1** ● Save ○ Skip
● log(IGC50-1) = -0.16 (in -log(mmol/L))
eADMET
T_pyriformis_all.sdf...
N: 1
120055-09-6
MoleculeID: *M6569*
RecordID: *R-1*
eadmet Only visible to eadmet

**Row 2** ● Save ○ Skip
● log(IGC50-1) = 0.87 (in -log(mmol/L))
eADMET
T_pyriformis_all.sdf...
N: 2
2430-16-2
MoleculeID: *M2525*
RecordID: *R-2*
eadmet Only visible to eadmet

Depending on the size of the uploaded set, the process may take from seconds to hours for completion (e.g. more than 50000 data points).

12. 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 deselect individual records from the upload.



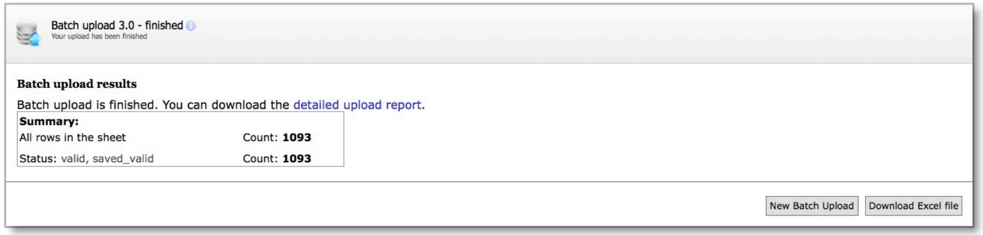
**Row 9** ● Save ○ Skip
● log(IGC50-1) = 0.42 (in -log(mmol/L))
eADMET
T_pyriformis_all.sdf...
N: 9
(2-BROMOETHYL)BENZENE
MoleculeID: *M12429*
RecordID: *R-9*
eadmet Only visible to eadmet

**Row 10** ● Save ○ Skip
● log(IGC50-1) = -0.16 (in -log(mmol/L))
eADMET
T_pyriformis_all.sdf...
N: 10
2-methylaniline
MoleculeID: *M9899*
RecordID: *R-10*
eadmet Only visible to eadmet

1 - 10 of 1093 — 10 items on page 1 of 110 > >>

13. Since all records being uploaded are valid, continue the upload by clicking the big “Proceed with upload” button.

    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).



Batch upload 3.0 - finished
Your upload has been finished

**Batch upload results**

Batch upload is finished. You can download the detailed upload report.

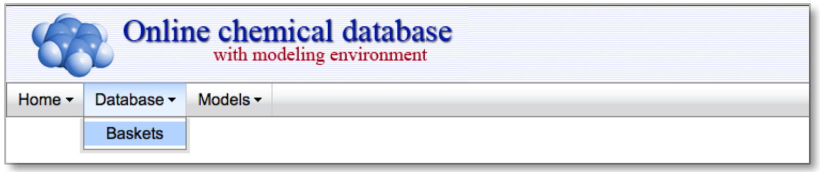
**Summary:**
All rows in the sheet — Count: **1093**
Status: valid, saved_valid — Count: **1093**

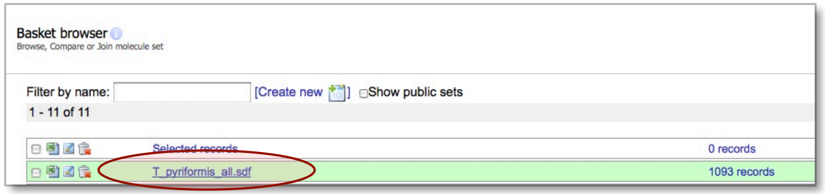
14. The final page of the batch data upload wizard gives some statistics about the uploaded data. You can review the uploaded data in the “Experimental property browser” or download a detailed report.

**Note:**

For your convenience, the uploaded data are automatically put into a newly created basket.

## 4.2 Advanced data management

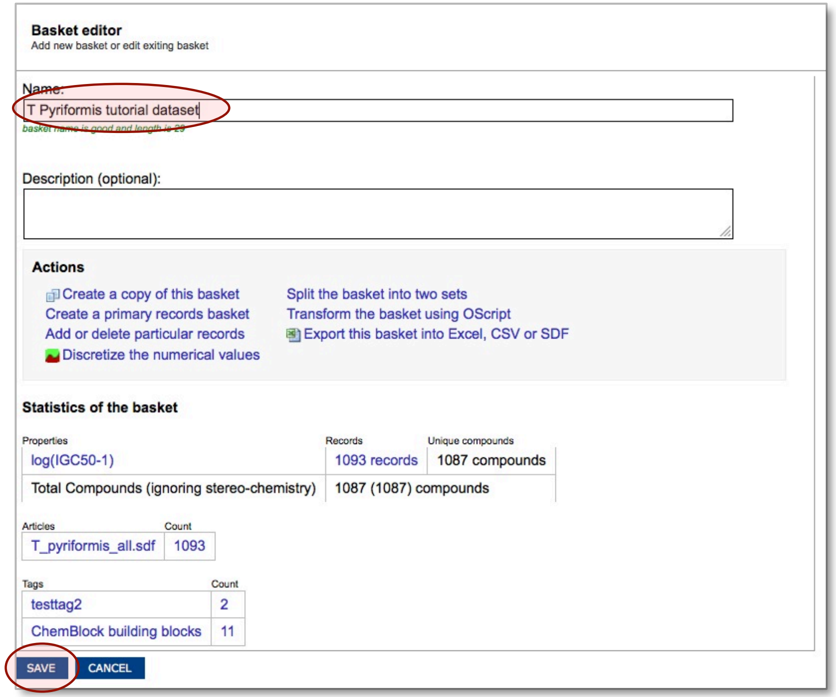



Basket browser
Browse, Compare or Join molecule set

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

1 - 11 of 11

| | | |
|---|---|---|
| □ | Selected records | 0 records |
| □ | T_pyriformis_all.sdf | 1093 records |


**Basket editor**
Add new basket or edit exiting basket

Name:
T Pyriformis tutorial dataset
basket name is good and length is 29

Description (optional):

**Actions**

- Create a copy of this basket
- Create a primary records basket
- Add or delete particular records
- Discretize the numerical values
- Split the basket into two sets
- Transform the basket using OScript
- Export this basket into Excel, CSV or SDF

**Statistics of the basket**

| Properties | Records | Unique compounds |
|---|---|---|
| log(IGC50-1) | 1093 records | 1087 compounds |
| Total Compounds (ignoring stereo-chemistry) | 1087 (1087) compounds | |

| Articles | Count |
|---|---|
| T_pyriformis_all.sdf | 1093 |

| Tags | Count |
|---|---|
| testtag2 | 2 |
| ChemBlock building blocks | 11 |

OCHEM allows combining experimental records into reusable sets. Such sets of records are referred to as **baskets**. Baskets have names and each user can virtually have as many baskets as required. Baskets are typically used as training or validation sets for the development of QSAR models.

### Basket browser and basket profile

1. Review a list of all your baskets in the basket browser accessible from *Database > Baskets* menu

The list should contain the basket created during the batch upload process from the previous tutorial.

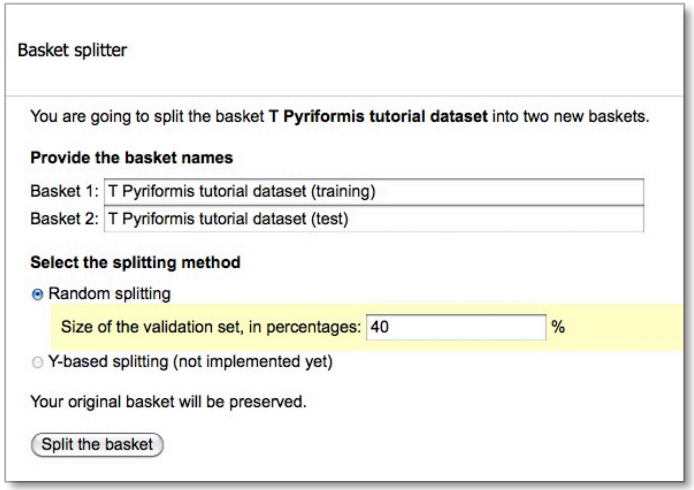
2. To open the profile of a particular basket, click on its name.

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.

Please give your basket a name of your choice (e.g., “T. Pyriformis tutorial dataset”) and click save.

### Splitting the basket

4. Click the link to randomly split your basket into two subsets.



Basket splitter

You are going to split the basket **T Pyriformis tutorial dataset** into two new baskets.

**Provide the basket names**

Basket 1: T Pyriformis tutorial dataset (training)

Basket 2: T Pyriformis tutorial dataset (test)

**Select the splitting method**

● Random splitting

Size of the validation set, in percentages: 40 %

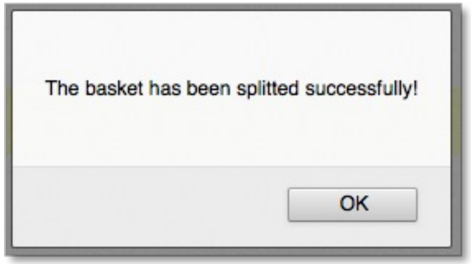
○ Y-based splitting (not implemented yet)

Your original basket will be preserved.

5. Enter names for the two new baskets to be created

6. Select the percentage by which the molecules are divided randomly in the training and validation set.

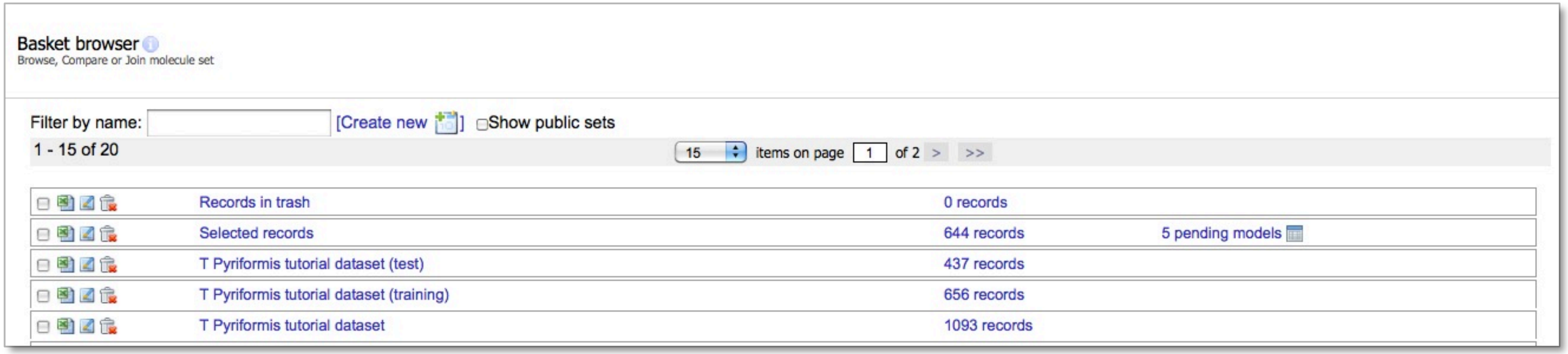
7. Click "split the basket" and wait for a couple of seconds.



The basket has been splitted successfully!

8. After notification that the new baskets have been created, you are forwarded to the basket browser.
9. The process is complete and you can see two new baskets in the basket browser.

We will use these baskets further in the model development tutorial.



Basket browser

Browse, Compare or Join molecule set

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

1 - 15 of 20 — 15 items on page 1 of 2 > >>

| | | | |
|---|---|---|---|
| □ | Records in trash | 0 records | |
| □ | Selected records | 644 records | 5 pending models |
| □ | T Pyriformis tutorial dataset (test) | 437 records | |
| □ | T Pyriformis tutorial dataset (training) | 656 records | |
| □ | T Pyriformis tutorial dataset | 1093 records | |

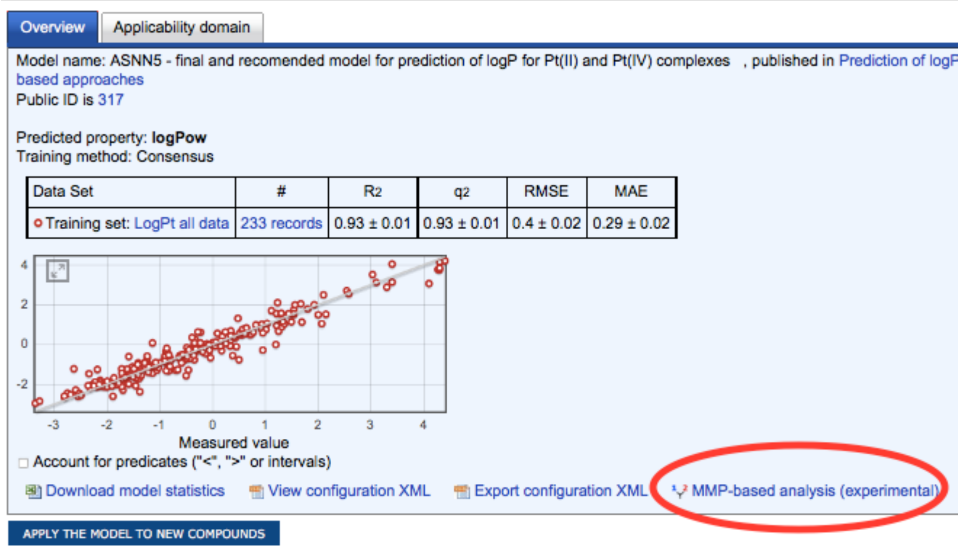
# 5. MMP Analysis

In the previous chapters, you have learned to perform the most basic steps of a QSAR modelling. In this chapter, we will review advanced features:

Interpretation of models using Molecular Matched Pairs

Matched Molecular Pair analysis to interpret models.

1. Click on the highlighted region to open the MMP plot



Model name: ASNN5 - final and recomended model for prediction of logP for Pt(II) and Pt(IV) complexes , published in Prediction of logP based approaches
Public ID is 317

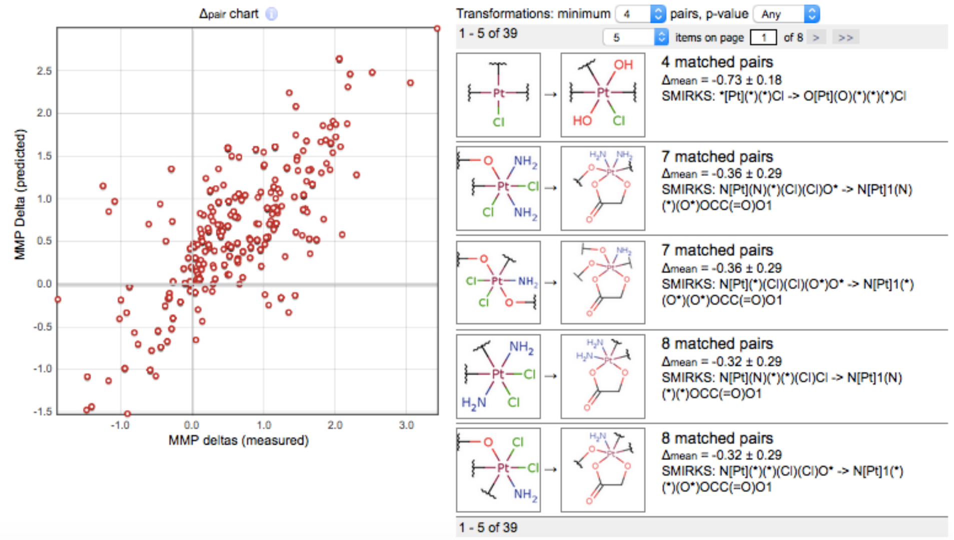
Predicted property: **logPow**
Training method: Consensus

| Data Set | # | R2 | q2 | RMSE | MAE |
|---|---|---|---|---|---|
| ○ Training set: LogPt all data | 233 records | 0.93 ± 0.01 | 0.93 ± 0.01 | 0.4 ± 0.02 | 0.29 ± 0.02 |

□ Account for predicates ("<", ">" or intervals)

Download model statistics
View configuration XML
Export configuration XML
MMP-based analysis (experimental)
APPLY THE MODEL TO NEW COMPOUNDS

2. Points near to the diagonal corresponds to the MMPs which were learnt. Points in 2nd and 4th quadrants correspond MMPs for which even the sign was incorrectly predicted.



Transformations: minimum 4 pairs, p-value Any
1 - 5 of 39 5 items on page 1 of 8 > >>

4 matched pairs
$\Delta_{mean}$ = -0.73 ± 0.18
SMIRKS: *[Pt](*)(*)Cl -> O[Pt](O)(*)(*)(*)Cl

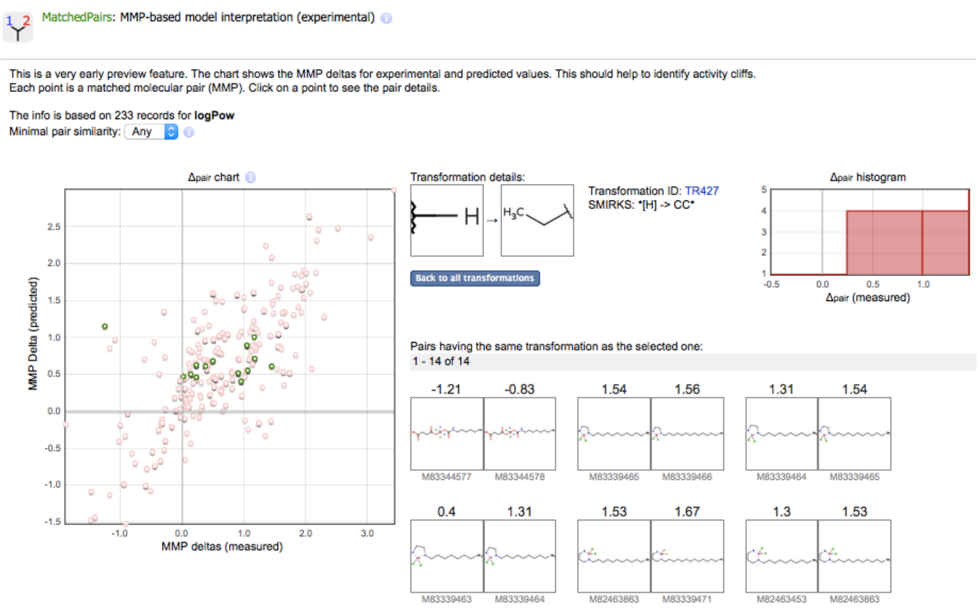
7 matched pairs
$\Delta_{mean}$ = -0.36 ± 0.29
SMIRKS: N[Pt](N)(*)(Cl)(Cl)O* -> N[Pt]1(N)(*)(O*)OCC(=O)O1

7 matched pairs
$\Delta_{mean}$ = -0.36 ± 0.29
SMIRKS: N[Pt](*)(Cl)(Cl)(O*)O* -> N[Pt]1(*)(O*)(O*)OCC(=O)O1

8 matched pairs
$\Delta_{mean}$ = -0.32 ± 0.29
SMIRKS: N[Pt](N)(*)(*)(Cl)Cl -> N[Pt]1(N)(*)(*)OCC(=O)O1

8 matched pairs
$\Delta_{mean}$ = -0.32 ± 0.29
SMIRKS: N[Pt](*)(*)(Cl)(Cl)O* -> N[Pt]1(*)(*)(O*)OCC(=O)O1

1 - 5 of 39



MatchedPairs: MMP-based model interpretation (experimental)

This is a very early preview feature. The chart shows the MMP deltas for experimental and predicted values. This should help to identify activity cliffs. Each point is a matched molecular pair (MMP). Click on a point to see the pair details.

The info is based on 233 records for **logPow**

Minimal pair similarity: Any

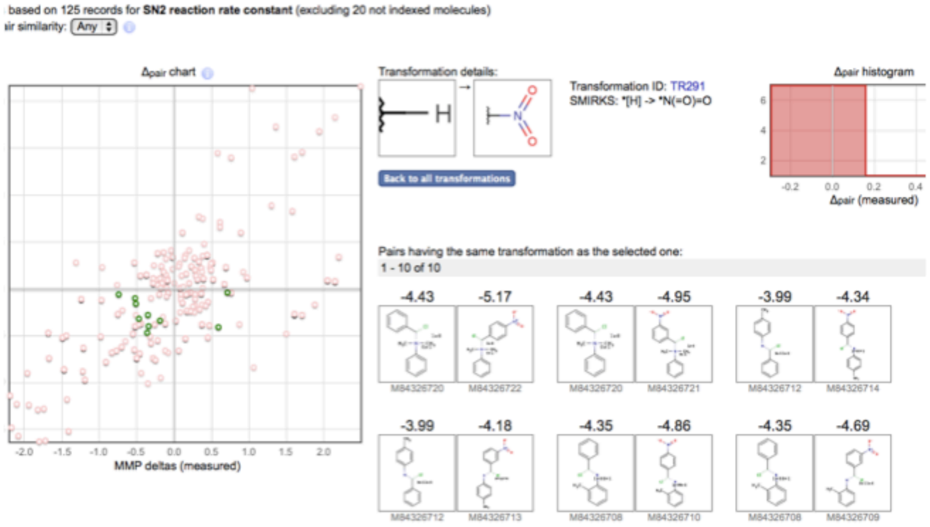
Transformation details:

Transformation ID: TR427
SMIRKS: *[H] -> CC*

Pairs having the same transformation as the selected one:

1 - 14 of 14

3. Clicking on each point will show the transformation responsible for each MMP. In the shown case the point in the 4th quadrant does not agree with qualitative changed of property (logP) for other transformations. It is likely to be an error, indeed addition of two carbon atoms CC increases logP – this is not an expected behavior.



based on 125 records for **SN2 reaction rate constant** (excluding 20 not indexed molecules)

ir similarity: Any

Transformation details:

Transformation ID: TR291
SMIRKS: *[H] -> *N(=O)=O

Pairs having the same transformation as the selected one:

1 - 10 of 10

4. The same approach can be used to interpret chemical reactions. Here the two groups of pairs of reactions correspond to two types of reactios.

## Acknowledgement

I thank Dr. Sushko, Dr. Novotarskyi, Mr. Körner and other members of my team for their work on the development of the OCHEM platform and preparation of the earlier version of this tutorial.