Introduction to OCHEM

based on the OCHEM user's manual
from eADMET
http://docs.eadmet.com/
and
OCHEM tutorial handouts

Author: Vlad Kholodovych, Ph.D.

OIT/High Performance and Research Computing
and Robert Wood Johnson Medical School
Rutgers Biomedical and Health Sciences (RBHS)
185 South Orange Avenue, MSB, Room C-631
University Heights, Newark, NJ 07103
Phone: 973.972.0663/732.235.2113
Fax: 973.972.7412
kholodvl@rutgers.edu
http://tiny.cc/compchem

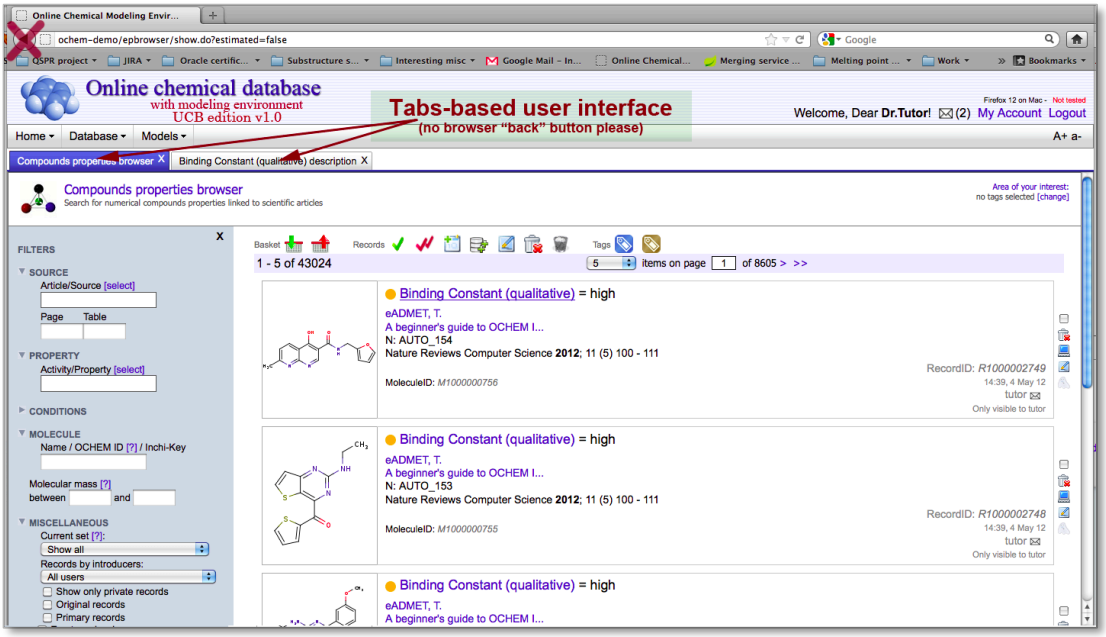
Before we start

OCHEM is a web-based platform. Users can access and operate it simply within a web browser, similarly to the way they access services like Gmail or Facebook.

The public version of OCHEM is available at **www.ochem.eu**.

To get the best experience it is recommended to use the latest version of Firefox browser. Chrome and Safari browsers are fully supported. At the moment OCHEM team does not support Internet Explorer and Konqueror Web browsers.

OCHEM user interface heavily utilizes “tabs” capability of web browsers. Often new dialogues and result windows get output in a new tab as shown below.



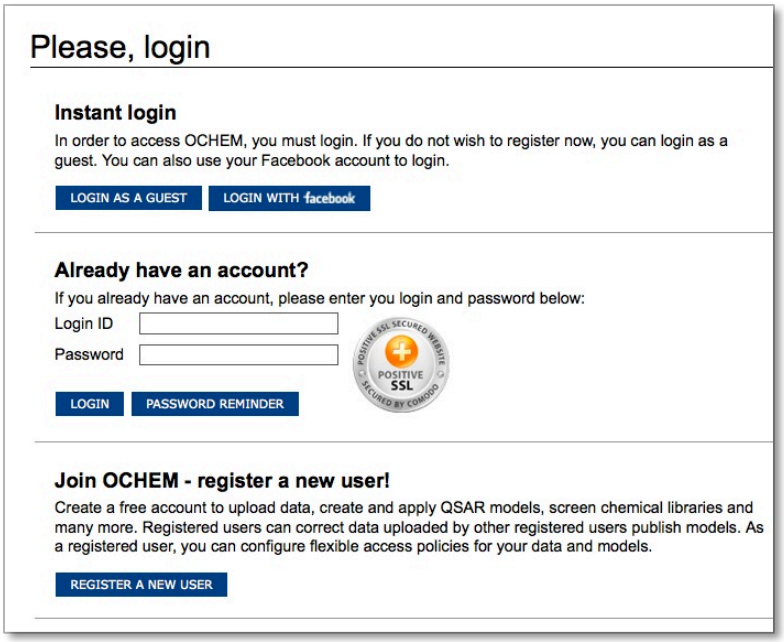
NOTE: *Please do not use the “back” button of a browser, as it is not fully compatible with the “tab” interface of OCHEM. Instead follow the navigation buttons at the bottom of OCHEM windows or open the links from the previous tab or OCHEM main menu.*

User account registration and login

A public version of OCHEM platform is free however a user needs to register an account and to be logged into his/her account to fully operate the OCHEM.

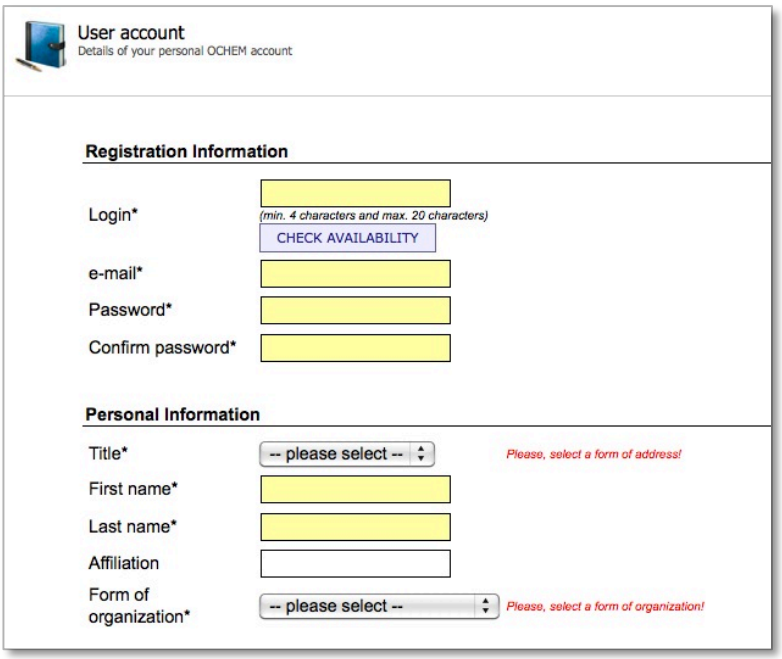
NOTE: *It is possible to log into OCHEM server without registration as a guest. Guest accounts have limited access to OCHEM resources and reserved mainly for users to get familiar with the interface*.

For this tutorial you will need to have a registered user account.



If you have already OCHEM account please log in using your username and password.

If you do not have an account yet please create a new one by clicking on the button "**REGISTER A NEW USER**" at the bottom of the page.



Select a login name (it should be between 4 and 20 characters) and check its availability by clicking on a "**Check availability**" button. If the selected login name is available continue by filling all mandatory fields (marked with an asterisk), read the Terms of Service agreement and accept it electronically by clicking on "**I accept. Create my account**" button at the bottom of the page.

You should receive a confirmation email with a link to activate your account within 10 min. Please follow the instructions in this email to complete your registration.
NOTE: *You may be able to log into your account immediately after accepting Service agreement without a confirmation email.*

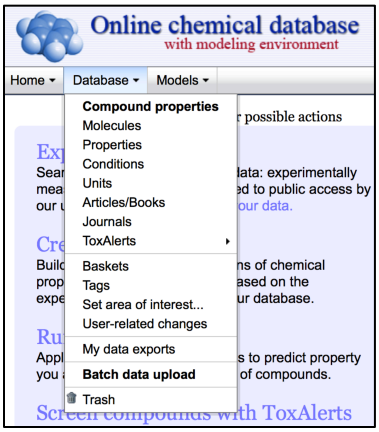
OCHEM SYSTEM OVERVIEW

OCHEM is an open source system that is moderated and managed by users in a way similar to Wikipedia. As a registered user you have a full access to all data on OCHEM server. You may modify, add, delete any record in OCHEM database; you may create your own QSAR models and make them available for all other registered users.

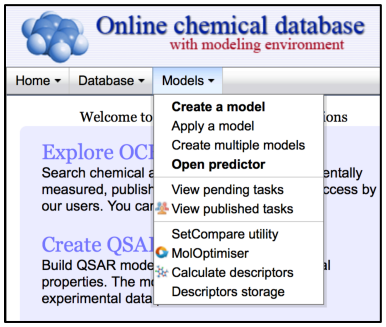
Please be EXTRA CAREFUL when you making any modifications to existing data or adding new properties to OCHEM database.

OCHEM platform consists of two major components: the database of experimental data and the modeling framework.

The **database** contains records of experimentally measured biological and physicochemical properties of small molecules together with information about conditions under which the experiments have been conducted and references to the resources from which the data have been collected. The most common sources of information on OCHEM are published articles in scientific journals, books, and commercial databases.
OCHEM records can be accessed through Database menu. Any operation related to data browsing and manipulation (adding new records, uploading datasets of molecules, modifying the existing datasets, etc) can be found here.



The **modeling framework** module provides variety of tools for development of predictive computational models for compounds from the OCHEM database.
Through the Model menu user can select and calculate molecular descriptors (OCHEM supports more that 20 types of state-of-the-art molecular descriptors from different vendors), specify and run various machine-learning algorithms (ANN, PLS, SVM, etc), analyze statistics of new and existing QSAR and classification models, perform validation and apply the best models to predict properties of new compounds.



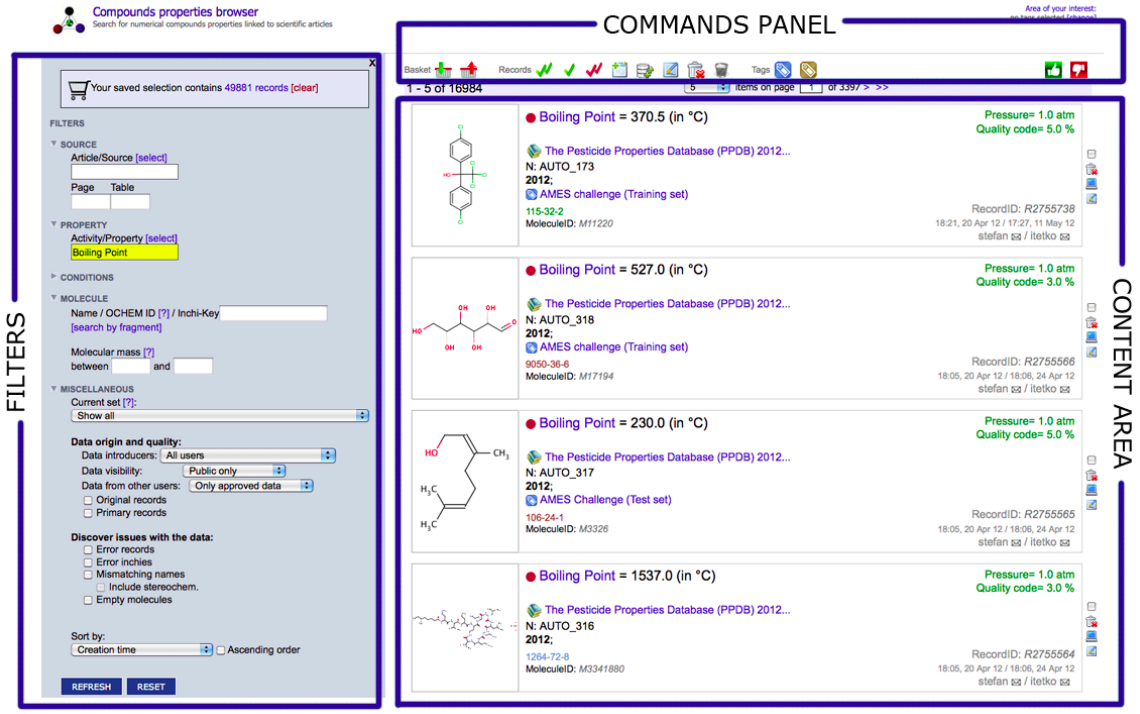
User browsers

OCHEM user interface is based on individual windows known as user browsers. OCHEM operates through browsers with all database entities (molecular properties, molecules, journal, conditions, properties, units, baskets, models, etc).

In this tutorial we will use different browsers however the overall organization of all browser windows in OCHEM is similar. Briefly, OCHEM browser layout includes three major components

Filters
Commands panel
Content area

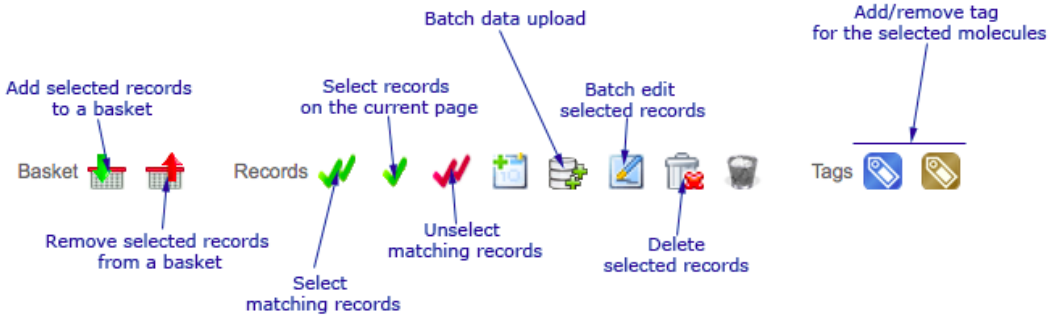
Below is a screenshot of one of the most important “Compounds properties” browser.



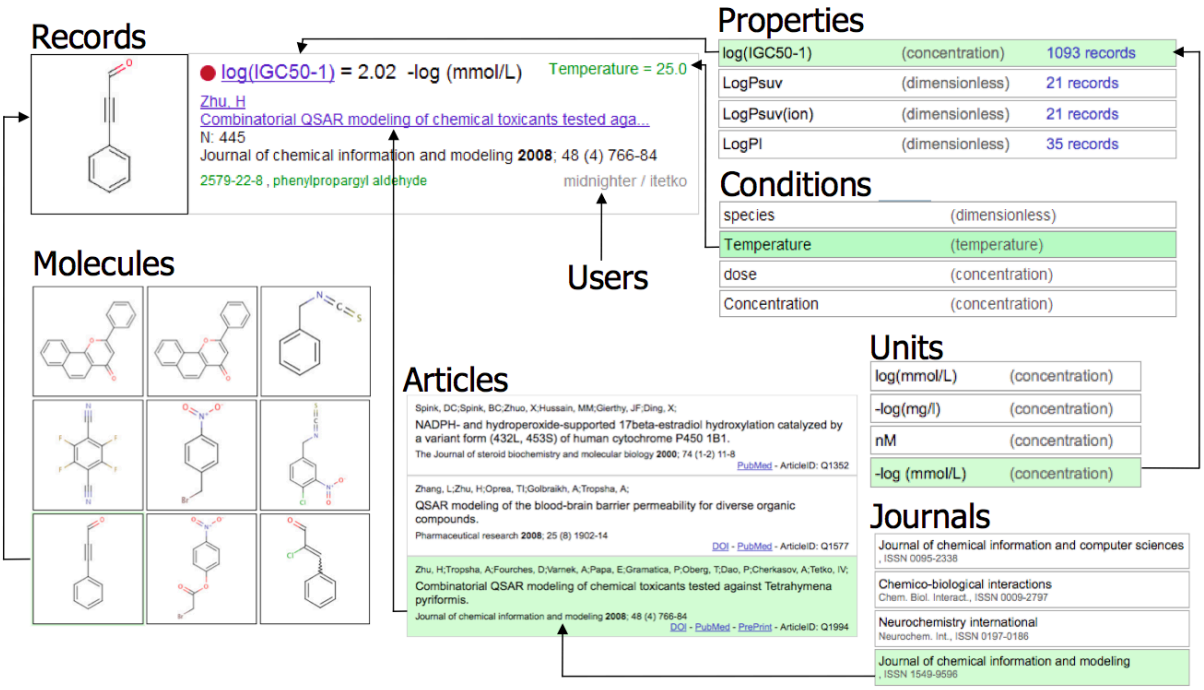
The **filters** panel allows user to specify content by applying various restrictions. In “Compounds properties” browser they are:

original source of data (scientific publication, journal); property (physicochemical parameter and biological activity); conditions of experiments (for example, pressure for boiling point; S9 activation for Ames test); molecular structure (substructure search, filters by molecular mass range), etc.

The **commands** panel is used to select/modify/delete/organize the content of the data records. Commands may be applied to an inidividual record (selected with a checkbox at the right side of each record) or for multiple selections.

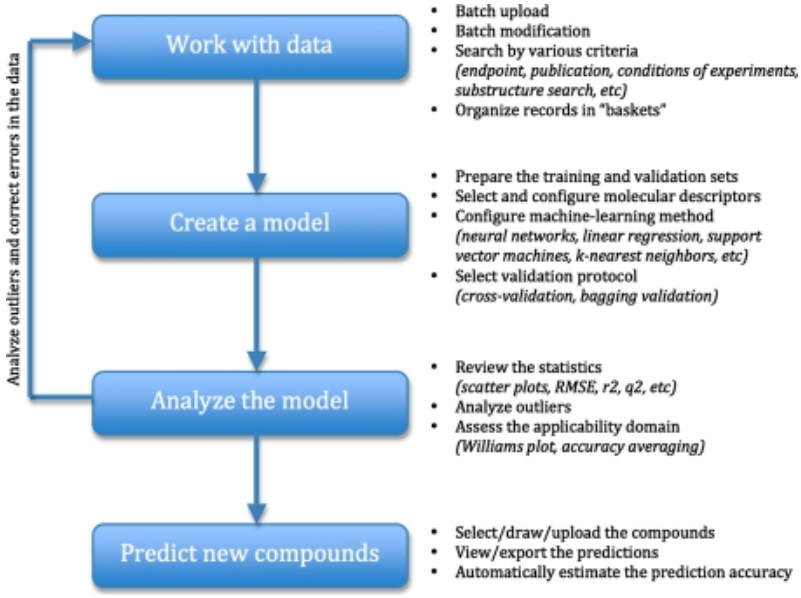


In the **content area** the actual data matching the specified filters are shown. If the content does not fit in one window it could be displayed page by page by selecting a page number at the top of the content area panel.
Detailed explanation of the content area of “Compounds properties” browser is shown below.



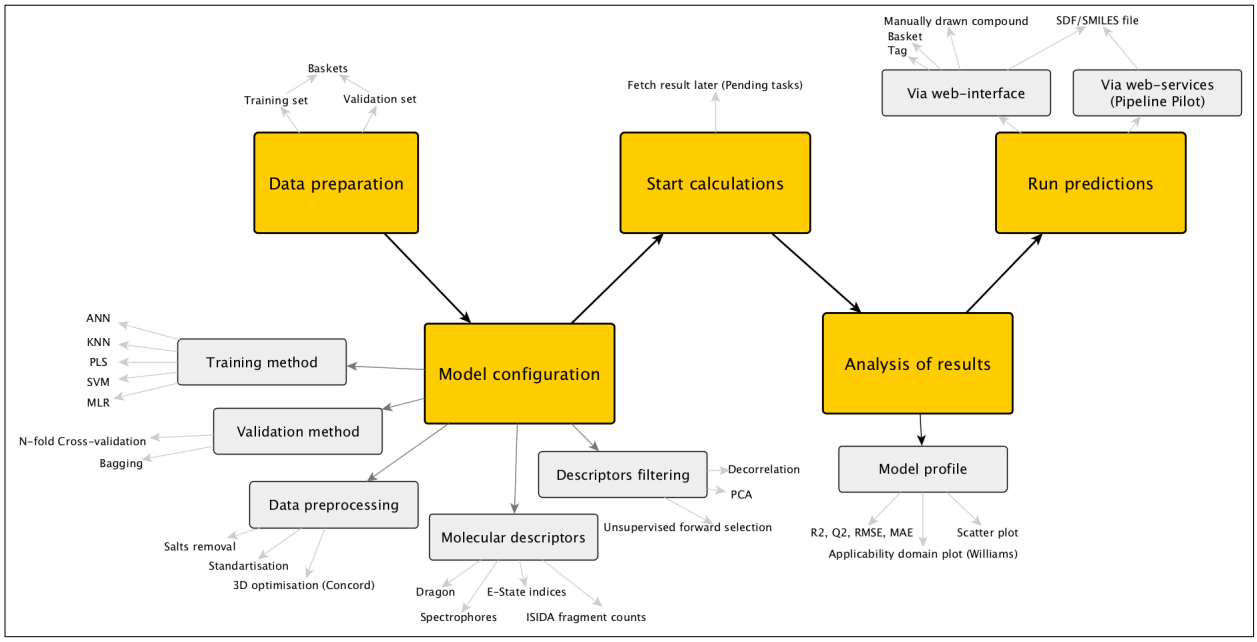
In this tutorial you will learn how to create a predictive QSAR model, validate it and apply to a set of new compounds.

A typical QSAR modeling framework is shown below.



http://docs.eadmet.com/display/MAN/Modeling+framework.

Tutorial workflow



1. DATA PREPARATION

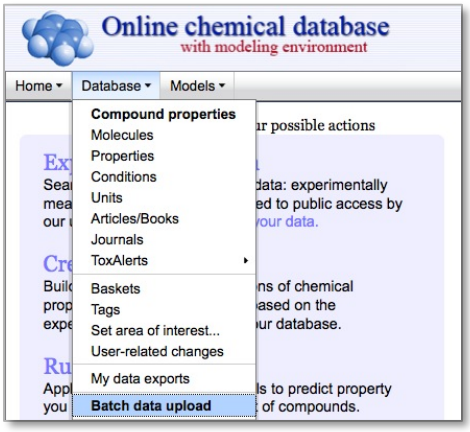
In this tutorial, we are going to work with about a thousand of molecules with known aqueous toxicity (measured as a growth inhibition concentration for T. pyriformis). Fortunately we do not need to introduce each record individually into OCHEM database. The current version of OCHEM supports a batch upload of data (molecular structures and properties) in three different formats – excel file, sdf file and comma separated csv file.

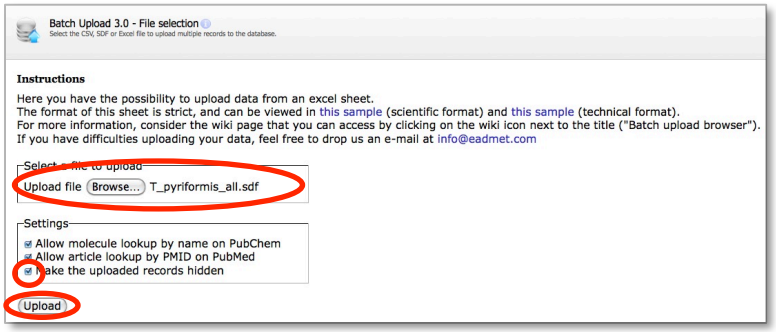
Download structural T_pyriformis files from http://tiny.cc/compchem
There are two files: excel and sdf. You may use either of them. Before we proceed with an upload of data to OCHEM webserver take a closer look on the content of those files.

Excel file: Open file in excel. For a data upload to OCHEM server at least three mandatory fields should be present: structural information as SMILES(or MOLECULE), NAME (or MOLECULEID if the structure is present in OCHEM database) and PROPERTY (registered property in OCHEM database). In our file we have SMILES, MOLECULEID and log(IGC50-1) field for the property. Close Excel.

SDF file: Open sdf file in a text editor (notepad, word, etc). After a structural information and $$$$ there are several fields <SMILES>, <N>, <log(IGC50-1)>, <CASRN>, however there is no NAME or MOLECULEID field present in the file. Fortunately another way to specify a name is by using CASRN number. If CASRN number is provided OCHEM batch loader automatically checks PubChem website for the name and makes a substitution in the database.
Close structural files and proceed with upload.

In the main OCHEM window from the "**Database**" menu select the "**Batch data upload**". The "Batch upload" wizard window will open.

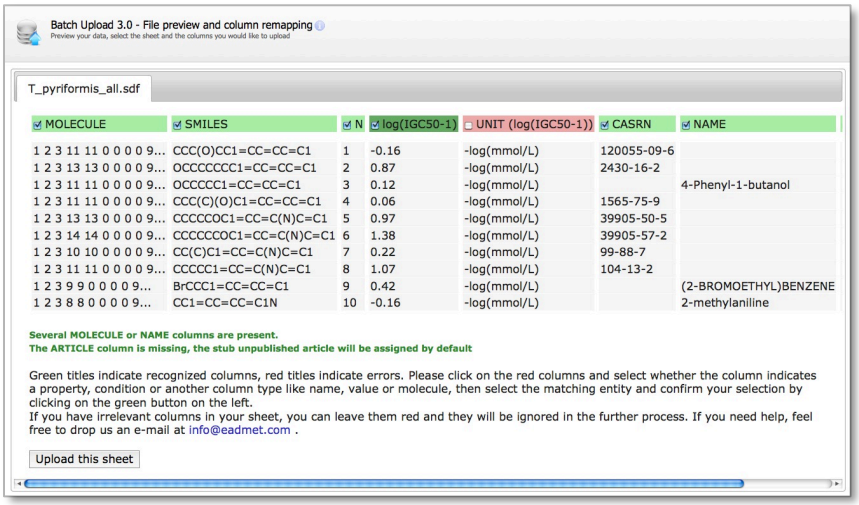


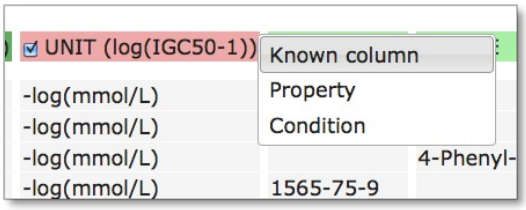


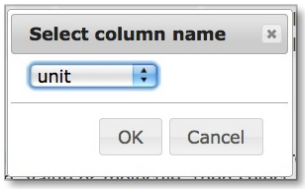
In the "**Upload file**" field click on Browse and select T_pyriformis_all.sdf file. This file contains about 1,000 measurements from the growth inhibition assay.

Make sure that "**Make the uploaded records hidden**" box is checked to avoid data conflicts with other OCHEM users. Click "**Upload**" to continue.

The file review window with "column remapping" tool will appear.

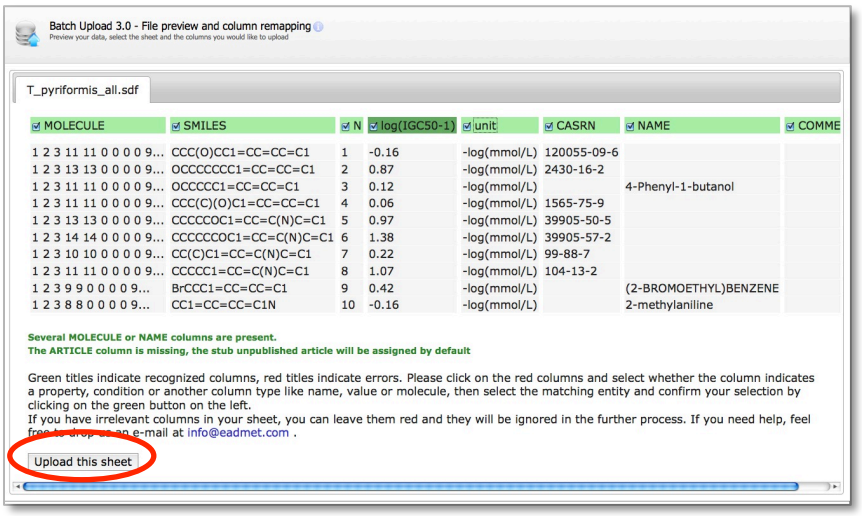






Here you can examine first few lines of the uploading file and make remapping of columns that are not recognized by the system.
NOTE: *Column headers are colour coded. Green means that the field is recognized by the system, red is not. The property column is dark green if this property is already registered in the OCHEM database.*

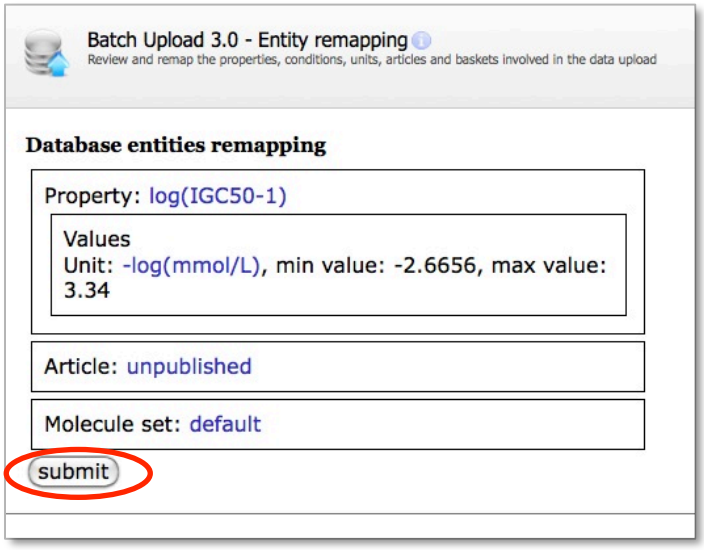
In our example the column with unit values was not recognized by the system (marked in red). We need to specify that these values represent units of "Aqueous toxicity". Click on the red column header. In a popped up menu window select "**Known column**" and then select "**unit**" from the list of known columns. Click **OK**.



Note that the color of column header has changed from red to green (a recognized unit), the name of the column is now "unit", and the checkbox in the column header is marked. Click "**Upload this sheet**" button to proceed.

You can review and make the final changes to uploading data on the next "Entity remapping" page.

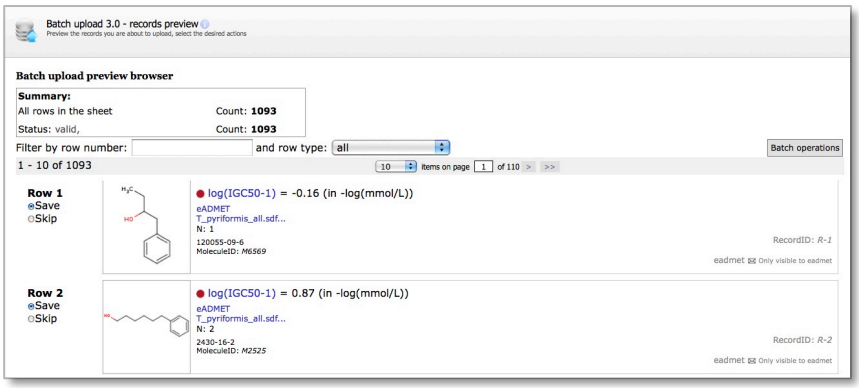
Since no source of data (a journal article or a book) has been specified for an uploading file this field was automatically marked as "unpublished" data by the system.



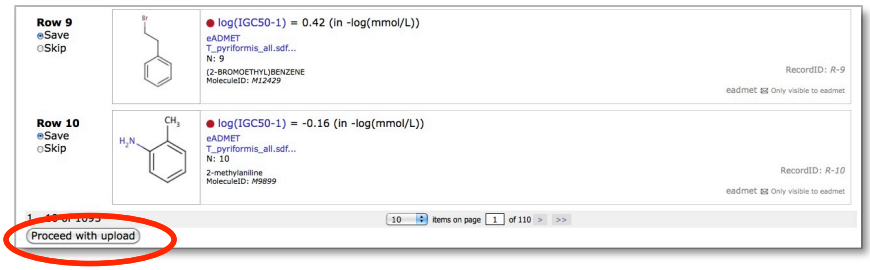
Click "**submit**" button to upload the data. A new dataset will be added as hidden data and it will be only visible to you (recommended option for the tutorial exercise).

NOTE: *If the size of the uploading set is more than 50000 it may take up to several hours to upload the data on a server. Please be patient.*

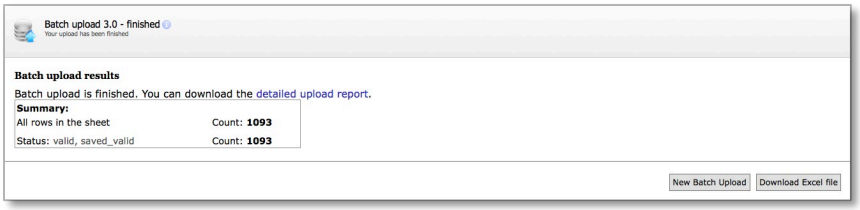
The last step in the batch upload is to examine the uploading data in a preview browser. Here you can check for any errors that may occur during the upload.
The statistical information about the dataset, the total number of processed records to be uploaded and the count of valid, erroneous and duplicated records among them are reported here. You may also exclude any individual records from upload on this page.



In our file all processed records are reported to be valid. Continue the upload by clicking on "**Proceed with upload**" button.



Finally the resulting page of the batch upload provides statistics about the uploaded data. You can review the uploaded data in the "Experimental property browser" on OCHEM or download a detailed report to your local computer.

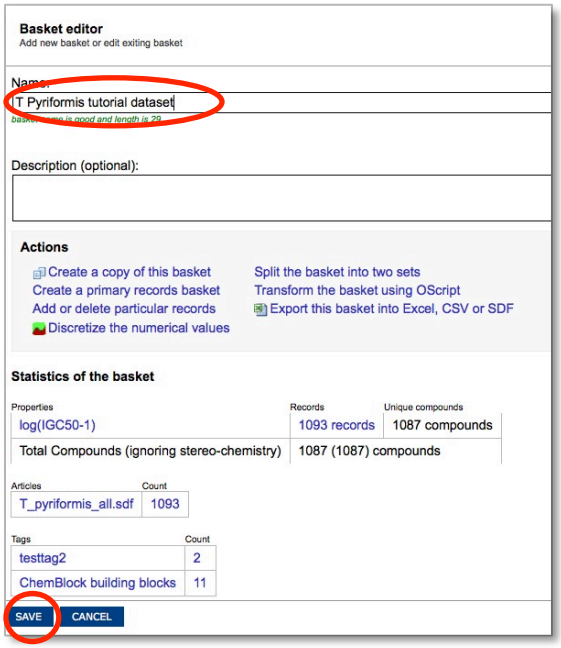


NOTE: *The uploaded data are automatically put into a new basket that can be accessed from* ***Data -> Basket menu***.

Experimental records on OCHEM server are organized into reusable sets and referred to as **baskets**. A user may create unlimited number of baskets, modify their content for specific tasks and rename them as needed. We will use baskets for the development of QSAR models.

Let's review your baskets in the basket browser accessible from ***Database > Baskets***

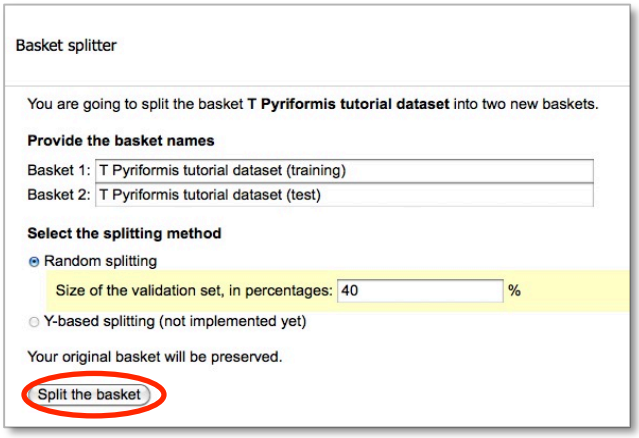
At this moment you should have at least one basket, the set of T_Pyriformis that we just uploaded.
Click on it.



In a new window, **Basket editor**, brief information about the basket is displayed. Here you can make changes to your baskets. Let's rename this basket. In the field "Name" put a new name "**T_Pyriformis_tutorial_dataset**" and click **Save**.

For typical QSAR modelling usually two sets of compounds are required. One is so called a Training set for creating (training) of a model and a Validation set (additional set of compounds that are not used for training) for verification of the new model robustness.

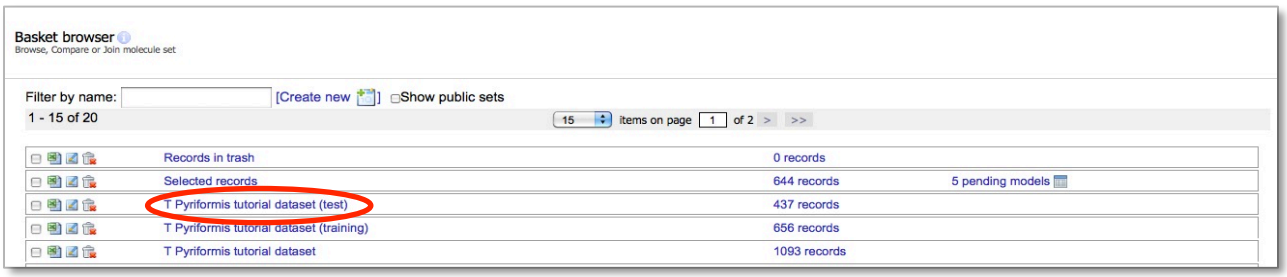
Open a Basket editor again.
Database-> Basket. Select **T_Pyriformis_tutorial_dataset.**
Now click "**Split your basket into two sets**".



Enter new names for baskets: Basket 1 will be used later as a training set – name it **T_Pyriformis_tutorial_dataset (training**)
Basket 2 will be our validation set – name it **T_Pyriformis_tutorial_dataset (test).**
For the splitting method select **Random splitting** and make a size of the validation set **40%**. It means that 40% of compounds will be in Basket 2 and 60% in Basket 1 after splitting. Click "**Split the basket**". The notification window about new baskets will pop up. Close it. Now you have two new baskets in the basket browser.

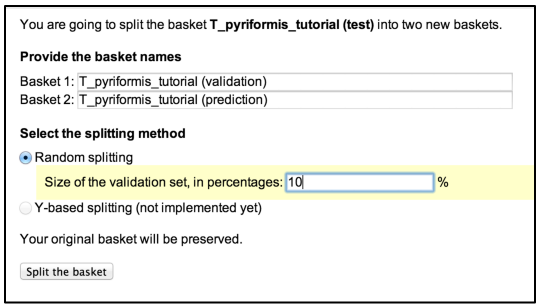
For this tutorial we need one more set of compounds. We will use it later to apply our QSAR model to for prediction of aqueous toxicity of new compounds.

Click on **T_Pyriformis_tutorial_dataset (test)** basket and split it into two more parts



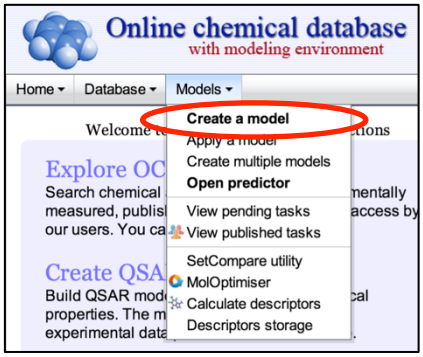
Enter new names for baskets:
Basket 1: **T_Pyriformis_tutorial_dataset (validation**)
Basket 2: **T_Pyriformis_tutorial_dataset (prediction).**
Size of the validation set: 10%. Click "**Split the basket**". Close the notification window.

We are ready to make a new QSAR model.



2. MODEL CONFIGURATION (see a workflow chart on page 7)

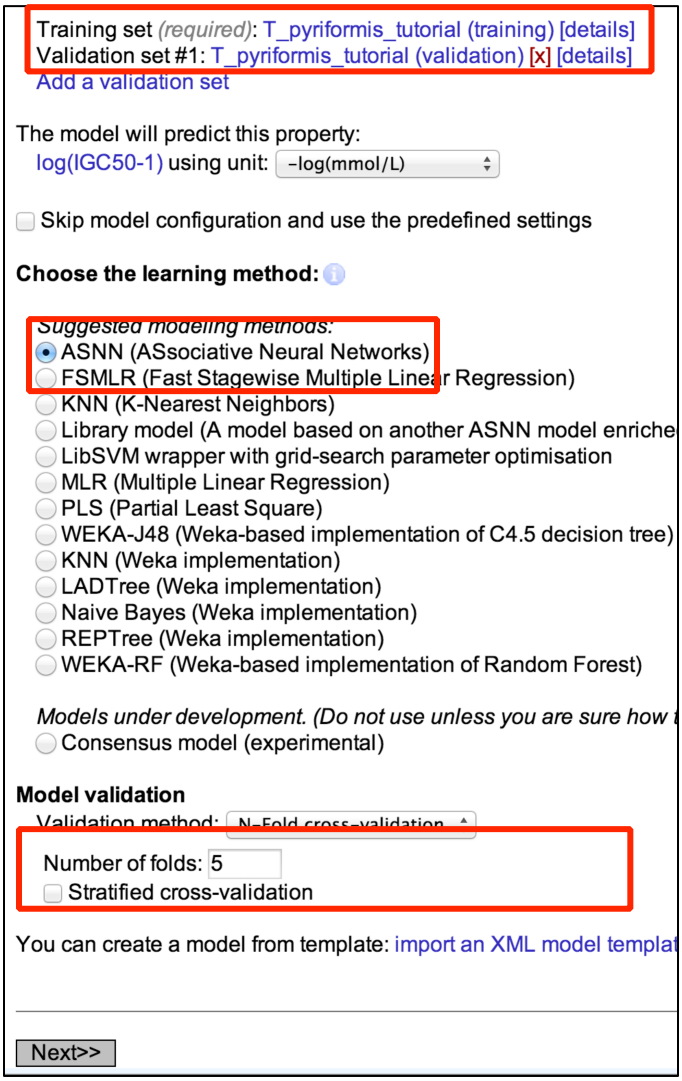
To start a new model development, open "**Models > Create a model**". In the next wizard window you need to select a training set, external validation sets (optional), a machine learning algorithm and a method for model internal validation.



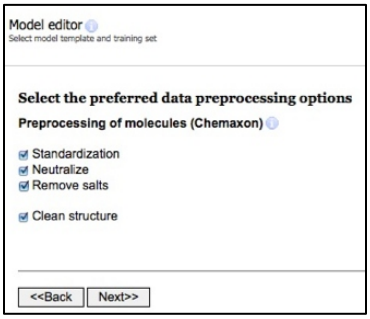
Click inside square brackets […] and select the training set ***T_pyriformis_tutorial (training)***.
Select a validation set ***T_pyriformis_tutorial (validation)***.
Make sure that the predicted property is set to –log(mmol/L).

OCHEM supports several dozen of machine learning methods for model training. For this tutorial, we will use the default one, associative neural networks (**ASNN**).
For internal model validation select "**N-fold cross-validation**" with **Number of folds: 5**. 5-fold cross-validation is the most common choice.

Click "**Next**".



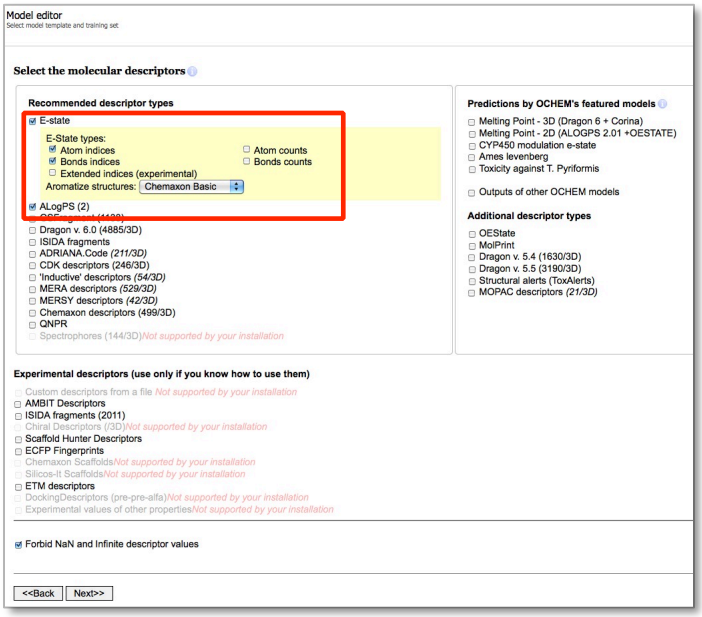
Next step is a sanity check for structural integrity of molecules in the dataset. Several options are available: standardization, neutralization, salt removal and cleaning the structural information stored in molecular files.



NOTE: *The last option, "**Clean structure**" is very important. If your uploaded dataset contains 3D information about molecules and you prefer to keep these coordinates intact this checkbox should be unchecked! It may be useful, for example, if your dataset is a set of ligands in preferred docking poses.*

We will use default values. Click **Next**

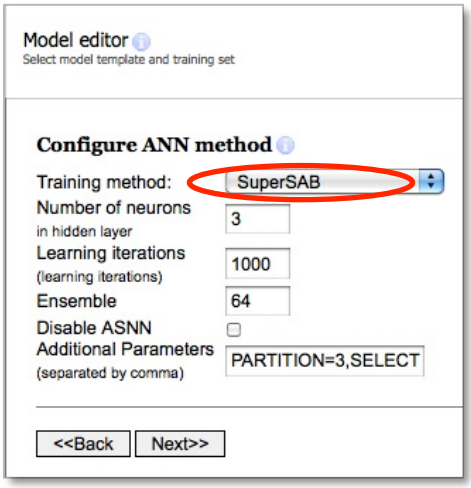
Next step is a selection of molecular descriptors.
OCHEM is a unique resource known for a vast collection of molecular descriptors.



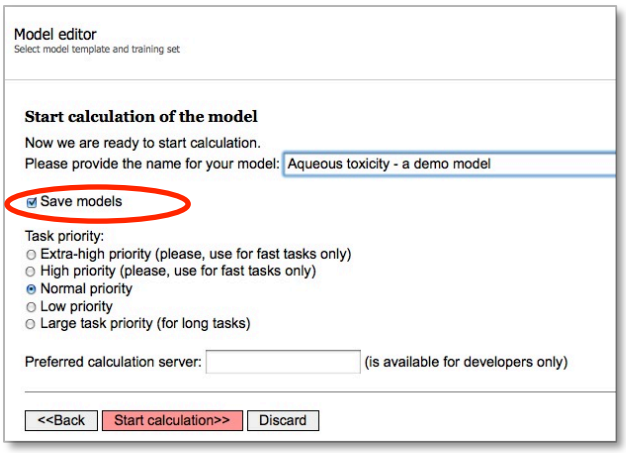
Several thousand descriptors from independent vendors and commercial software packages can be calculated here. The choice of molecular descriptors is one of the most important steps in developing of the successful QSAR model. It is common that the user tries different type of descriptors with different fitting techniques in the process of developing QSAR models. In this tutorial, we will use a default selection – simple topological **E-State** descriptors and **ALogPS**.

Click "**Next**"

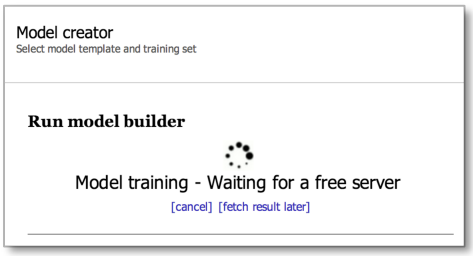
Another important aspect in making a good unbiased model is a quality of selected descriptors. They need to be orthogonal, or in other words to be independent from each other. In the next window you may select various filters to sift out highly correlated and empty descriptors. It allows significantly reduce number of descriptors and avoid unnecessary calculations for fitting algorithms.
We will use default values.
Click **Next**



Each fitting method requires a configuration of its parameters. OCHEM has a preset of most common default values associated with each particular method. In addition to default parameters users may select their own configuration.
For ANN we will use SuperSAB training.
Click **Next**

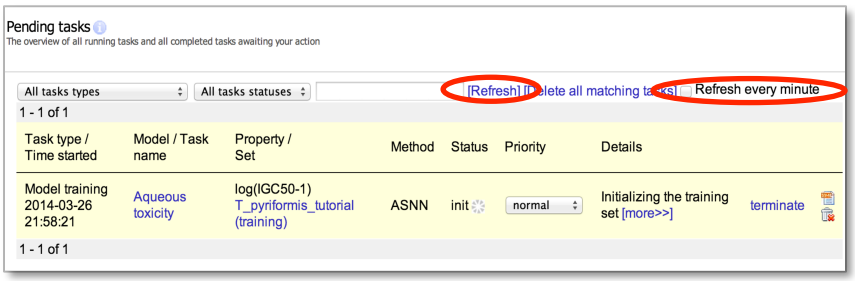


Finally, name you model "**Aqueous toxicity**" and make sure that the check box "Save models" is marked. We are ready to start calculations, click "**Start calculation**"

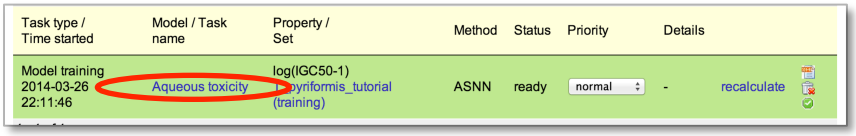


The model training will start and the current status of the model building process will be displayed. You may wait here for the modelling process completion, which may take from several minutes to several days depending on the size and complexity of the model setup or click on "**fetch result later**" and check the status of your job later.

To check the status of your job go to **Models - > View pending tasks**. The status of running jobs can be checked manually by clicking on **Refresh** link or get it updated every minute by selecting a corresponding checkbox.

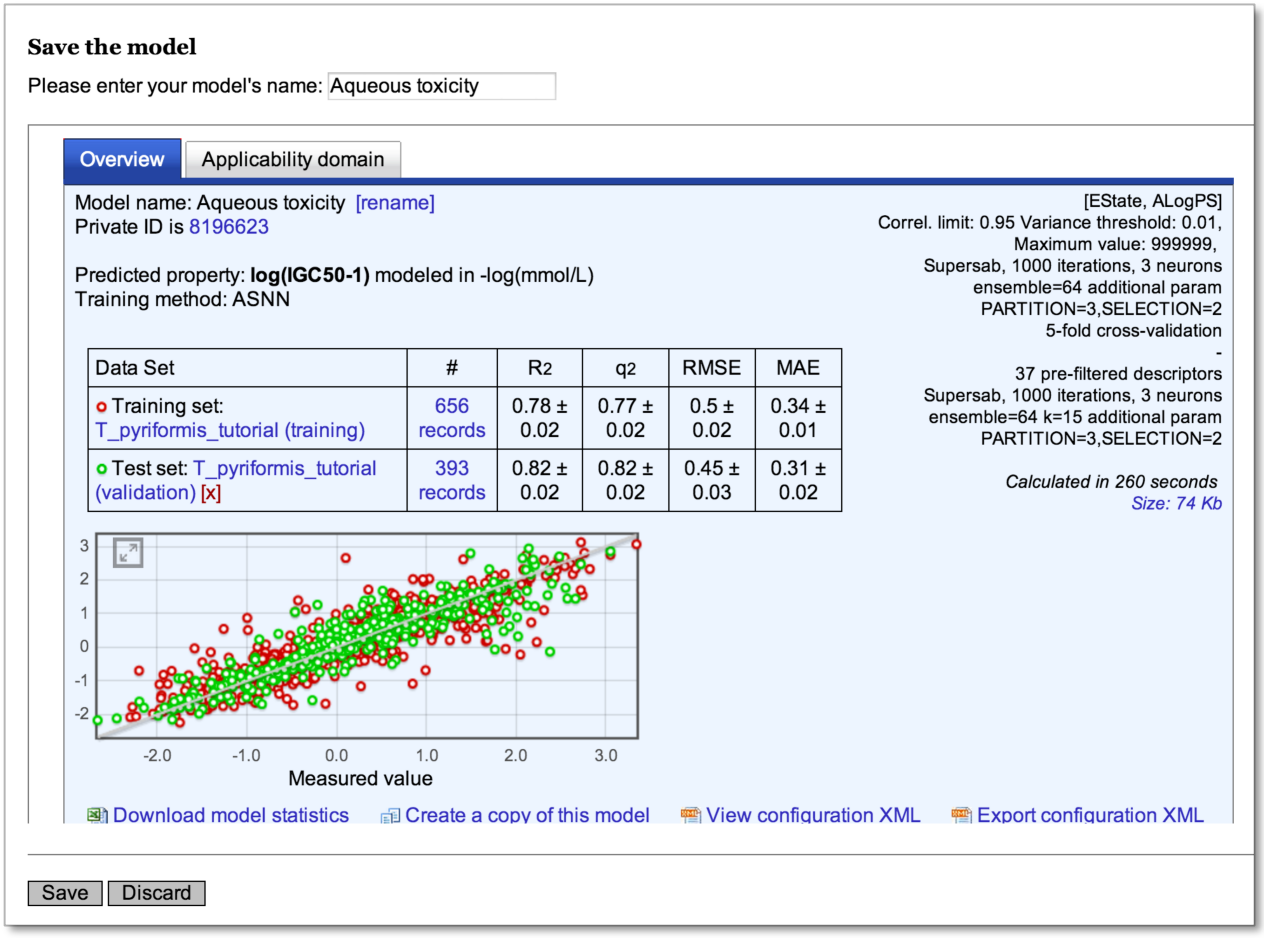


You may terminate the running job any time here or fetch the completed models for analysis. When the task is successfully finished the yellow background of the running jobs will change to green.



Click on a model name to see a model profile.

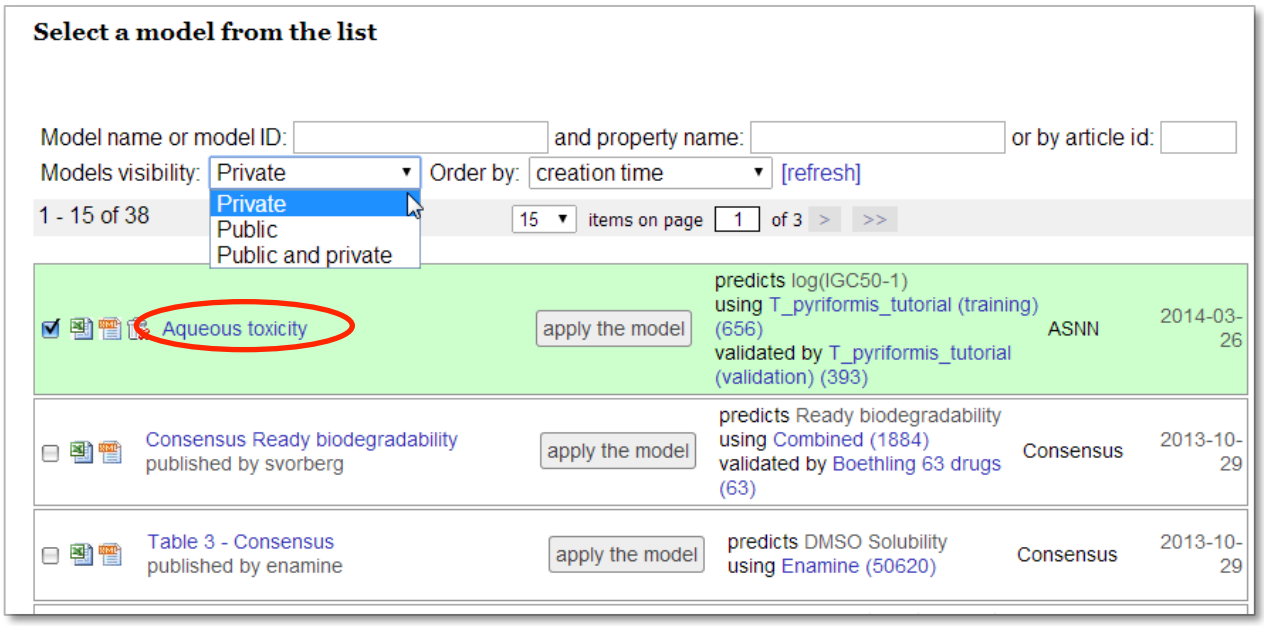
!!! CONGRATULATIONS !!!
You have successfully built a QSAR model on the OCHEM platform.
A model profile window contains information about a model initial configuration (used descriptors, machine-learning method and predicted property), basic statics for each set (correlation coefficient R and cross-validated correlation coefficient q, root mean squared error „RMSE“ and mean absolute error „MAE“).



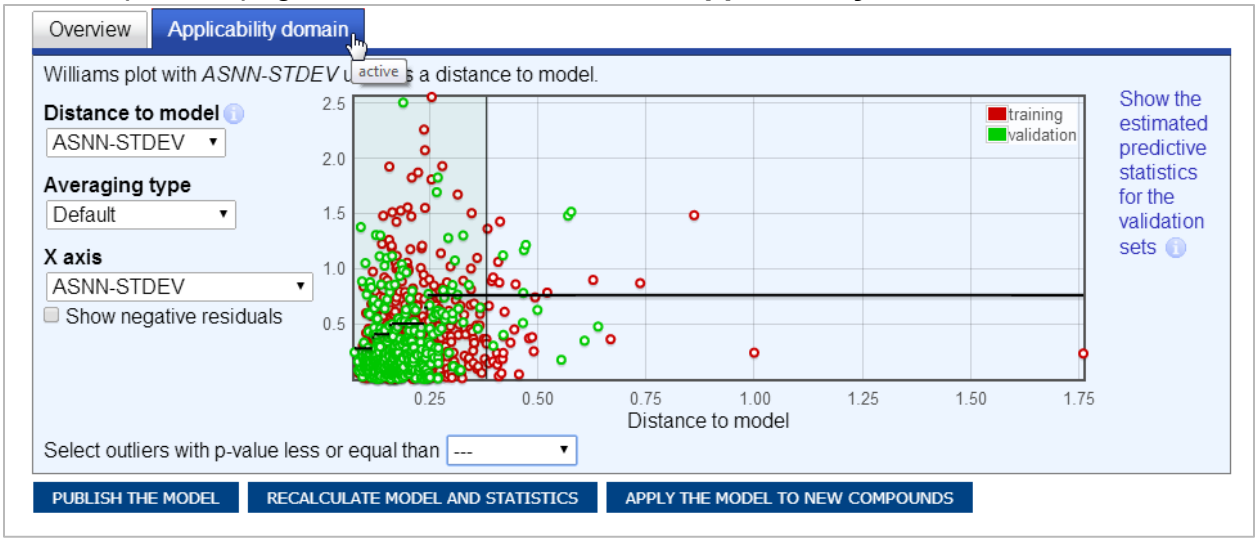
R^2 shows how well data are fit into the model while q^2 is a measure of a model stability. The higher those numbers the better model you have. RMSE and MAE represent the difference between the actual observations and the observation values predicted by the model. The smaller those numbers the better your model.
Important!! Write down numbers from the table on a piece of paper.

Below the table a scatter plot showing predicted versus observed (measured) values is shown. Each point in the plot is clickable and linked to a corresponding profile of this compound. Red dots on the plot represent a distribution of compounds from a training set, green points are compounds from a validation set.
As you can see some points are far away from the correlation fitting line. Most likely compounds represented by those points are outliers in our model. Let’s see if we can improve the model by removing outliers.
However before we start to play farther with our model let’s save it.
Click **Save** at the bottom of the page. After saving your model screen will disappear.

All saved models on OCHEM can be accessed through **Models -> Apply model**.

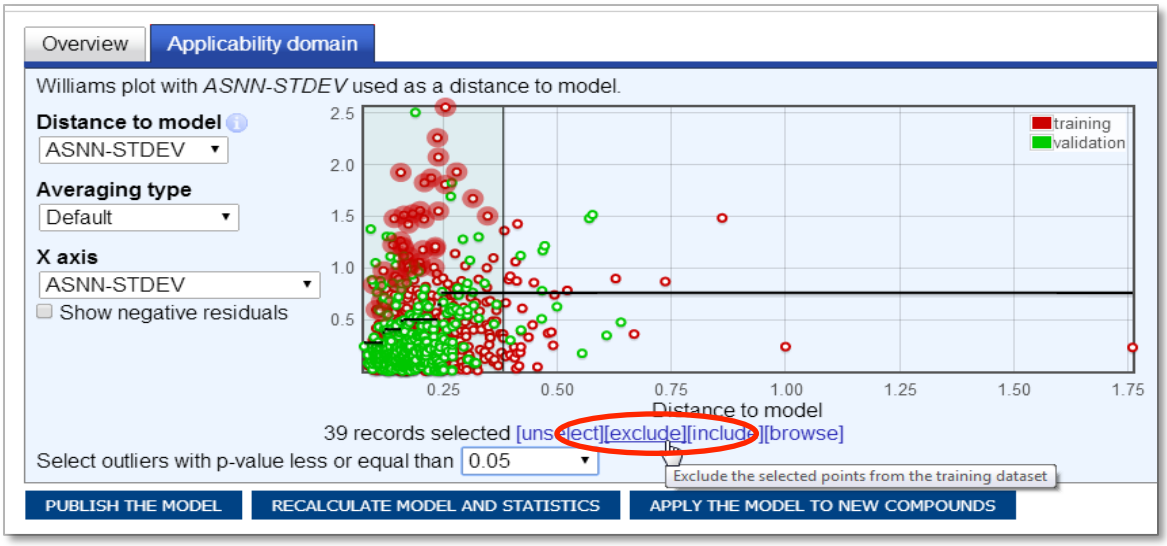


Both your private models and publicly available models published by other users are shown here. Let's find our model. In the header **Models visibility** select **Private**. Now only your model should be listed. You may apply your model to predict properties of new compounds by clicking on the button "Apply the model". However before doing this let's try to improve our model by excluding some outliers.
Click on the model name "**Aqueous toxicity**" instead to bring the model profile window. You may select outliers manually on a scatter plot however OCHEM has a very useful feature when outliers can be selected and excluded from a training set based on the p-value.
At the top of the page click on the second tab – **Applicability domain**.

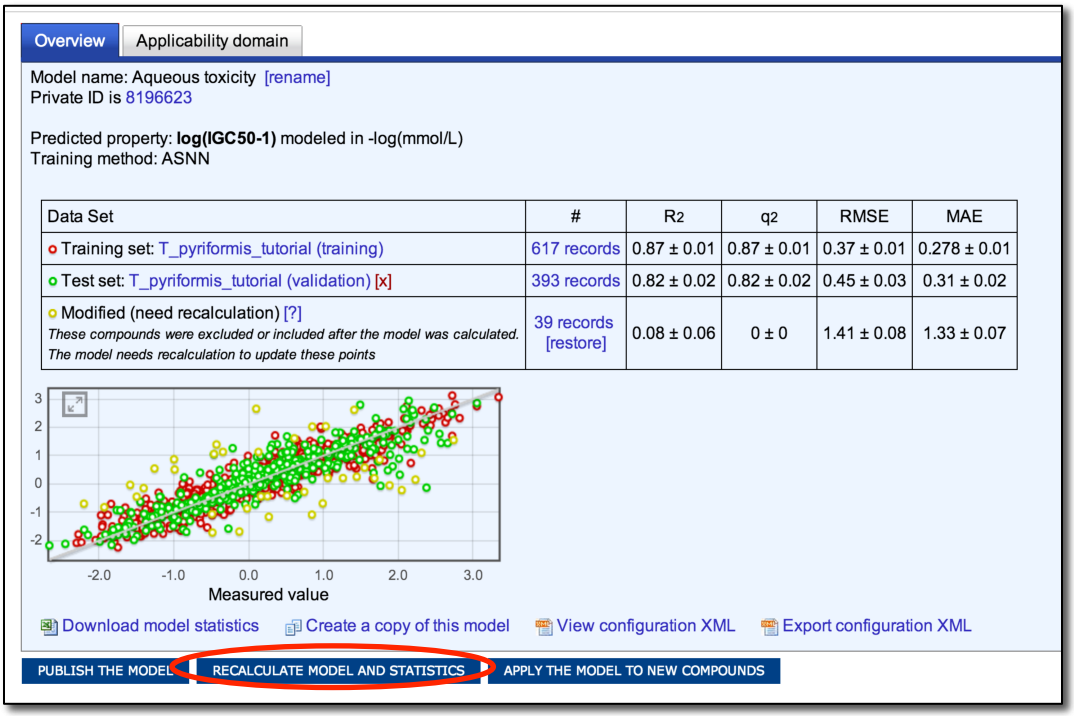


The Williams plot for standard deviation is shown here. Immediately below the plot you may select outliers based on the various level of p-value.

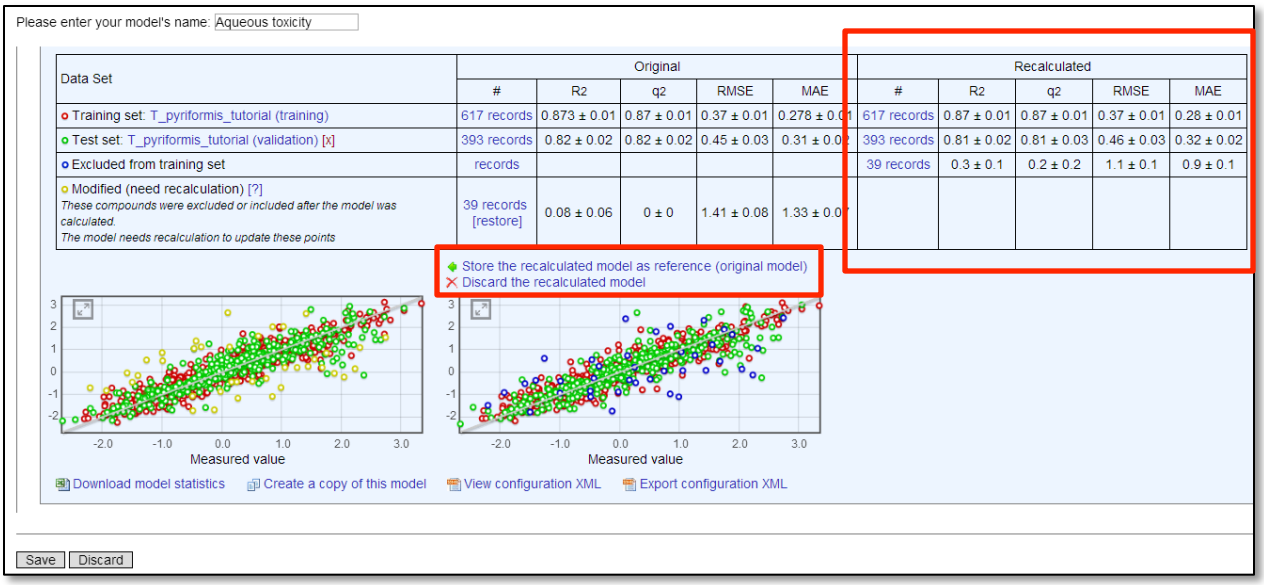
Select p < 0.05. Molecules from a training set with significance below selected p-value will become highlighted.



Click on **"exclude"** link to exclude this molecule from a training set.
Switch back to model overview (tab **Overview**). There is a new row in the table showing that 39 records have been excluded from the training set (this number can be different in your case because the original dataset was split randomly and you may have slightly different compounds in your training set). All excluded compounds are highlighted in yellow on the scatter plot. Statistical numbers for the training set also changed to new estimated values (may be higher or lower than originals). However it is just estimation, our model has been made with all compounds and it is possible that some of excluded outliers have significant influence on the model statistics. We need to recalculate the model and statistics. Click on the corresponding button at the bottom.



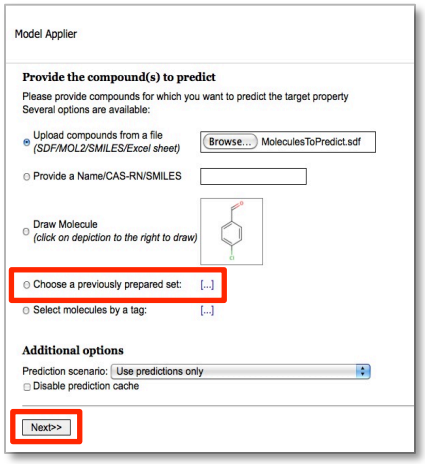
A tiny notification at the bottom of the page shows that the calculation started. Check the status of your job in **Models -> View pending tasks**. Do not forget to select "**Refresh every minute"** checkbox or refresh the page manually.
When calculations complete click on the model name to bring updated model statistics for analysis.
A new block of statistical values for recalculated model is added to the table.



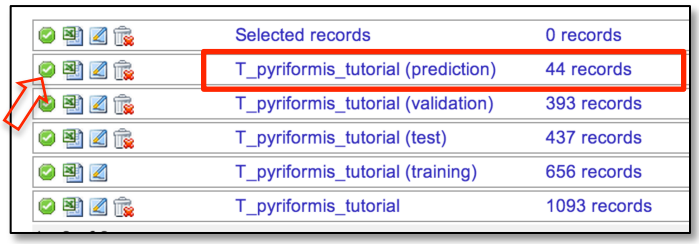
Unfortunately OCHEM has a small glitch: the statistics reported in the table for Original and Recalculated models are the same. We can't compare them here.
Remember I asked you to write down statistical values from the table on a piece of paper. Find this paper now. Compare the numbers that you have before with updated numbers. If new numbers are better (R^2 and q^2 higher) we will keep a new model. If the new numbers are worse than we revert the model to a previous stage.
Choose an appropriate link below the table for accepting or rejecting a new model. Then click **Save**.

Okay, now we have a good QSAR model that we may use for prediction for compounds that have not been yet tested.
Remember we created the third dataset at the beginning of this tutorial. Let's predict aqueous toxicity for these compounds.
Go to **Models -> Apply a model**
Find your model on the list – use option Private if there are many other models – and click on "**Apply the model**" button.

In a new window "**Model applier**" you can select a set of molecules for prediction, upload a new structural file or even draw a structure using embedded molecular editor. We will use our earlier prepared set.

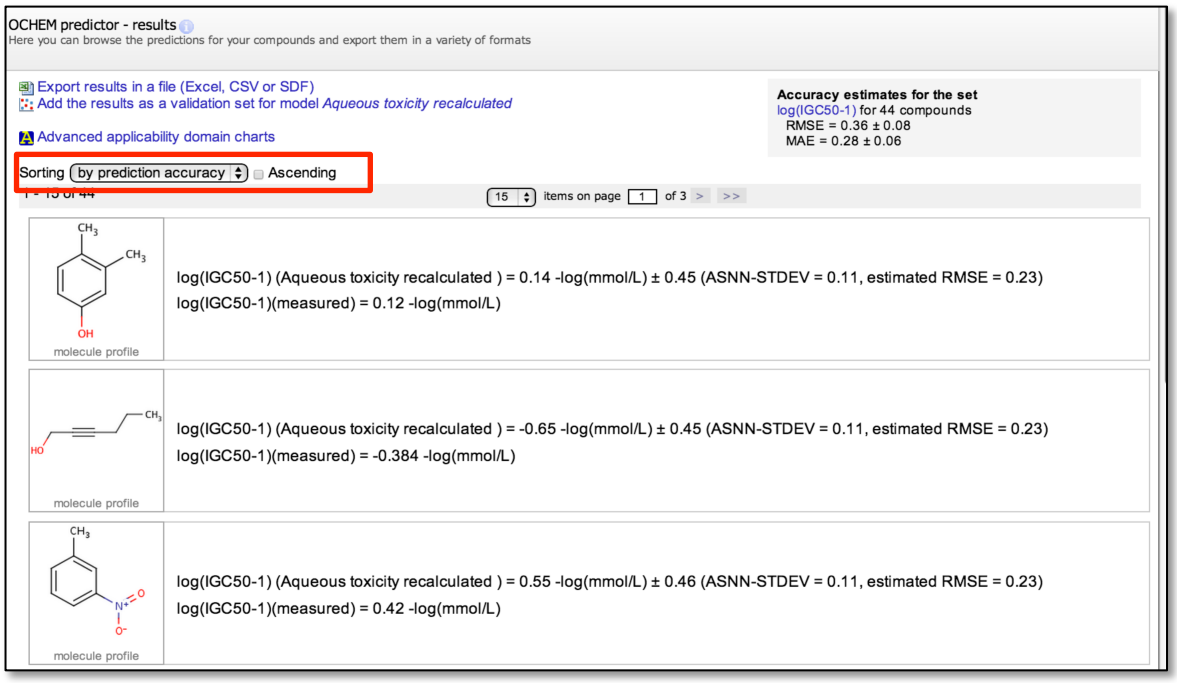


Select "**Choose a previously prepared set**" and click on **[…]**. In the basket browser select "**T_pyriformis_tutorial (prediction)**" set and click on green checkmark.



We are ready to run prediction. Click **Next** in the Model Applier.

Shortly a new window with predicted results will appear. Sort results "**by prediction accuracy**".
In the main panel you may examine predictions of aqueous toxicity for each individual compound. Because our set for prediction was extracted from a bigger set of compounds with known (experimentally measured) toxicity we can see how accurate our QSAR model is in prediction. However if you make a prediction for new compounds with unknown experimental values you entirely rely on your model.



You may save prediction results in a separate file (excel, csv or sdf) or add compounds from this set as a new validation set and recalculate your QSAR model.