Uploading a stub QSAR model

Motivation

In OCHEM you can use a wide variety of descriptors, machine learning methods and data processing techniques to build a predictive QSAR model. However, there are several cases when you have a model that was built using some specific tools or custom protocol, and you would like to introduce this model to OCHEM. For this purpose you can use the "Upload stub model" tool that OCHEM offers. It is a tool that allows you to create a stub OCHEM profile based on the real/predicted value pairs that you provide. As a result of model upload you get a public model profile similar to a normal OCHEM model. The difference is that it's impossible to apply this model to new compounds. Most common use cases for model upload include:

1) Publishing your datasets and model results - possibly as part of a collaboration or a journal publication

2) Model analysis using some of OCHEM tools (like Nearest neighbour analysis or MMP analysis).

Model upload preparation

Data preparation

First, you need to upload all the data records involved in the uploaded model. You can use the Batch data upload to ol to achieve this. After your upload is done, you need to organize your data in baskets according to what your model's training and validation sets are. On creating baskets please refer to the Working with datasets section of the documentation. Once your training and validation sets are ready, you can proceed to creating a model upload file.

Model upload file

A model upload file (XLS, CSV or SDF) generally contains two columns: one identifies one of the OCHEM records in your training or validation sets, and another holds the "predicted" value for this record. Therefore, the first column can be either **MOLECULE** or **EXTERNAL_ID**. The second column should be named **PREDICTION**.

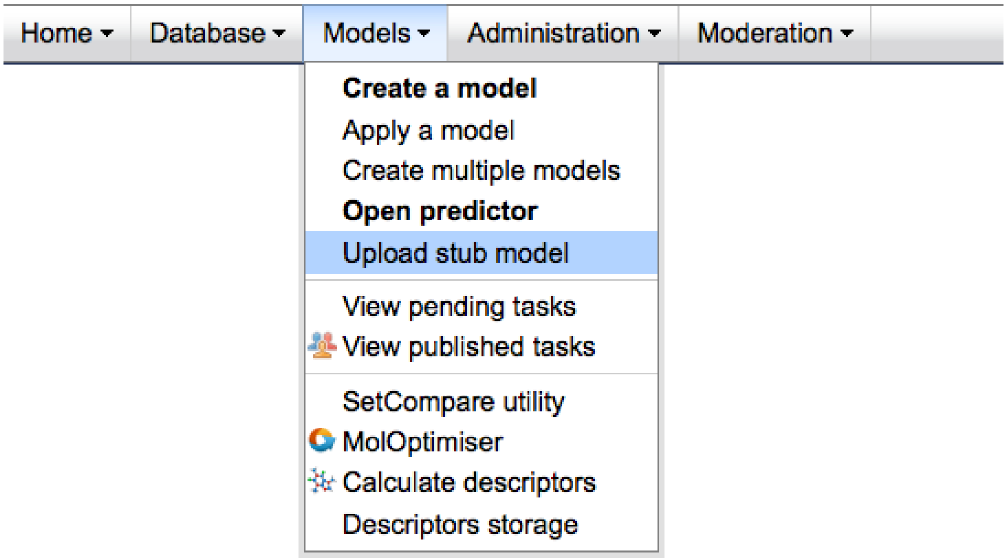
If you name the first column MOLECULE, you can fill the column with OCHEM - compatible molecule data. That is you can either use OCHEM molecule ids (e.g., *M1234*), smiles (e.g., *CCCN*), sdfs, etc. The tool will take this molecular data and try to find the record in your training and validation set baskets that corresponds to this molecule.

If you name the first column EXTERNAL_ID, you should fill the column with same external record identifiers that you used during your batch upload in the EXTERNAL_ID column of the batch upload sheet.

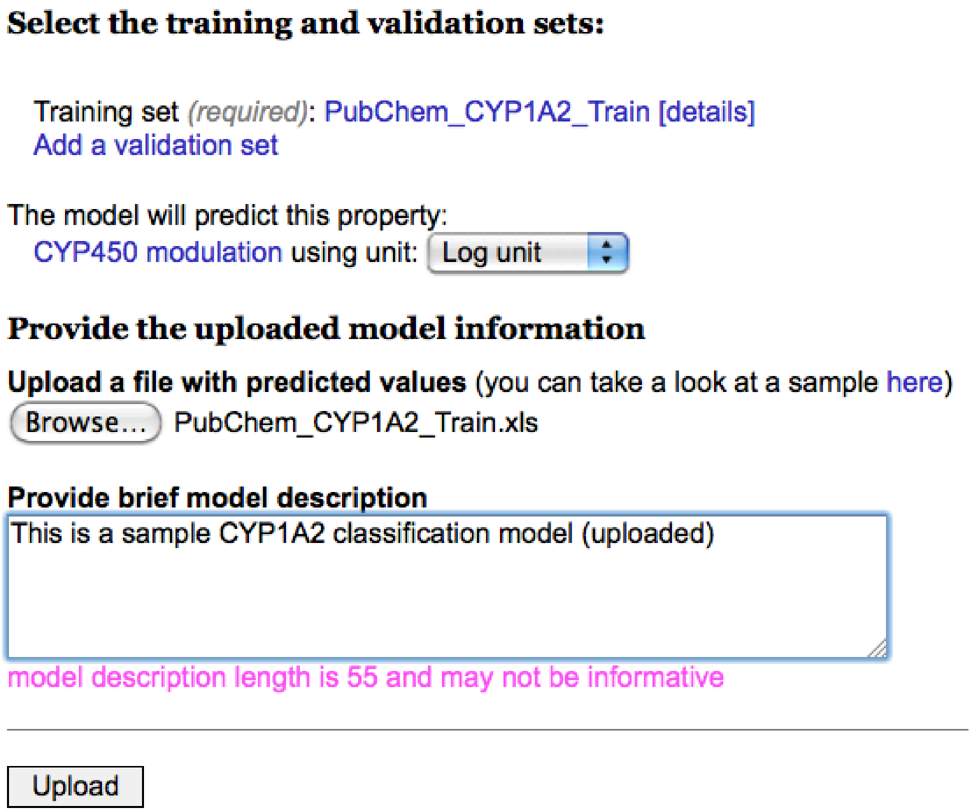
You can download an example of the sample file for uploading predicted classification value for CYP inhibition using OCHEM molecule ids here - model-upload.xls

Make sure, that your uploaded file has an exact one-to-one correspondence to the contents of your training and validation set baskets. Any extra data in the model upload file, or any records in the baskets that do not have a corresponding predicted value in the file are considered errors.

Model upload page



On the model upload page, select your training and validation sets, select your model upload file, and provide a brief description of the uploaded model.



Select the training and validation sets:

Training set *(required)*: PubChem_CYP1A2_Train [details]
Add a validation set

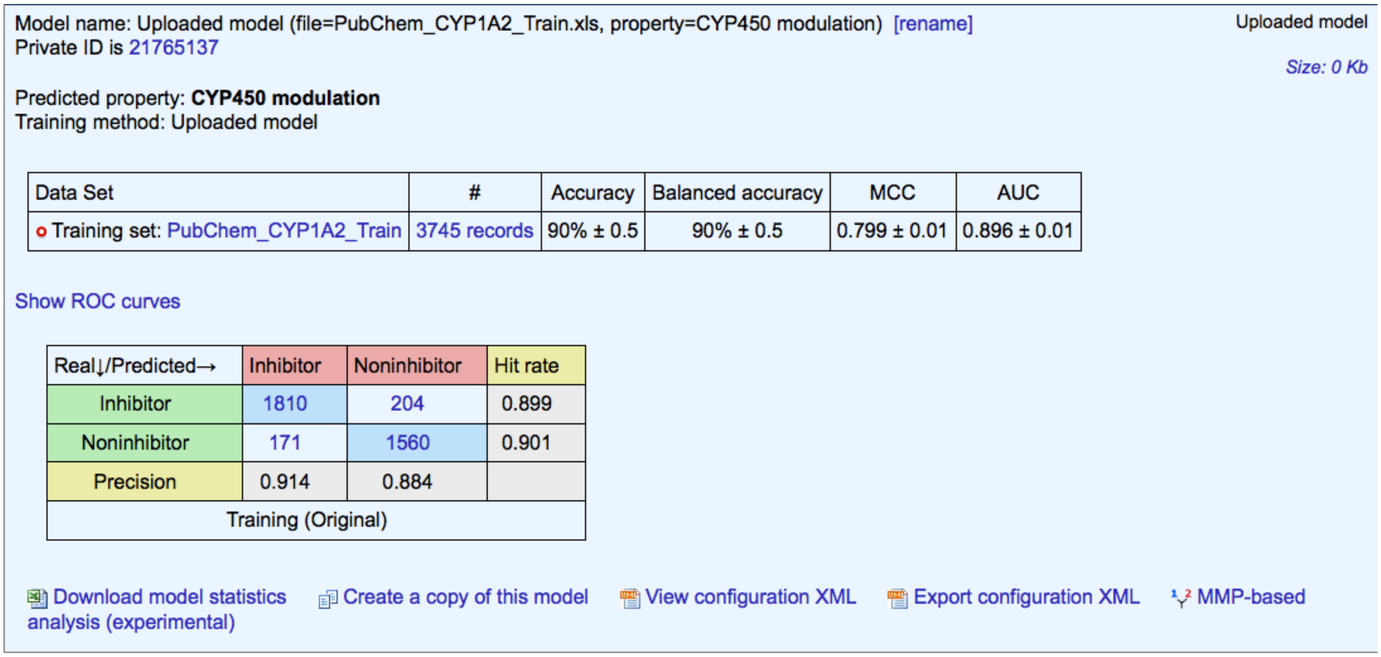
The model will predict this property:
CYP450 modulation using unit: Log unit

Provide the uploaded model information

Upload a file with predicted values (you can take a look at a sample here)
Browse… PubChem_CYP1A2_Train.xls

Provide brief model description
This is a sample CYP1A2 classification model (uploaded)
model description length is 55 and may not be informative

If everything goes well and the uploaded model file does not contain any errors, you will be presented by an Uploaded model OCHEM profile



Model name: Uploaded model (file=PubChem_CYP1A2_Train.xls, property=CYP450 modulation) [rename]
Private ID is 21765137

Uploaded model

Size: 0 Kb

Predicted property: **CYP450 modulation**
Training method: Uploaded model

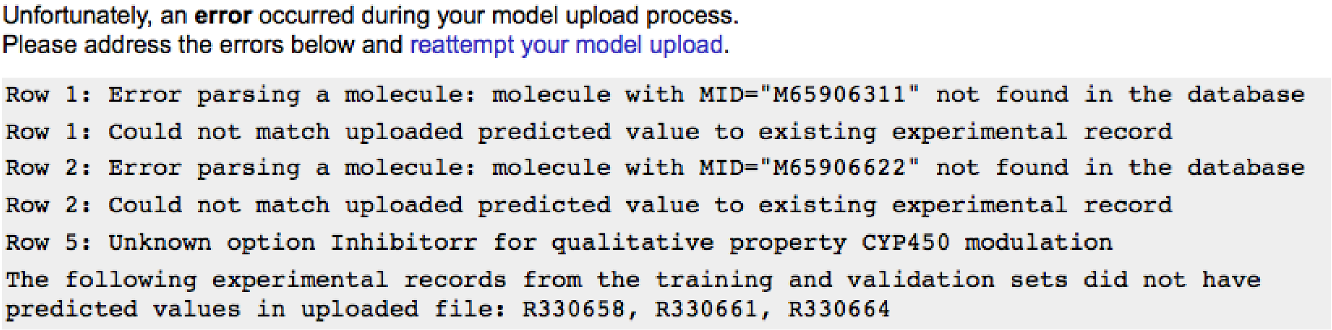
Data Set	#	Accuracy	Balanced accuracy	MCC	AUC
○ Training set: PubChem_CYP1A2_Train	3745 records	90% ± 0.5	90% ± 0.5	0.799 ± 0.01	0.896 ± 0.01

Show ROC curves

Real↓/Predicted→	Inhibitor	Noninhibitor	Hit rate
Inhibitor	1810	204	0.899
Noninhibitor	171	1560	0.901
Precision	0.914	0.884	
Training (Original)			

Download model statistics
Create a copy of this model
View configuration XML
Export configuration XML
MMP-based analysis (experimental)

Otherwise, a list of errors will be presented.



Unfortunately, an **error** occurred during your model upload process.
Please address the errors below and reattempt your model upload.

```
Row 1: Error parsing a molecule: molecule with MID="M65906311" not found in the database
Row 1: Could not match uploaded predicted value to existing experimental record
Row 2: Error parsing a molecule: molecule with MID="M65906622" not found in the database
Row 2: Could not match uploaded predicted value to existing experimental record
Row 5: Unknown option Inhibitorr for qualitative property CYP450 modulation
The following experimental records from the training and validation sets did not have
predicted values in uploaded file: R330658, R330661, R330664
```