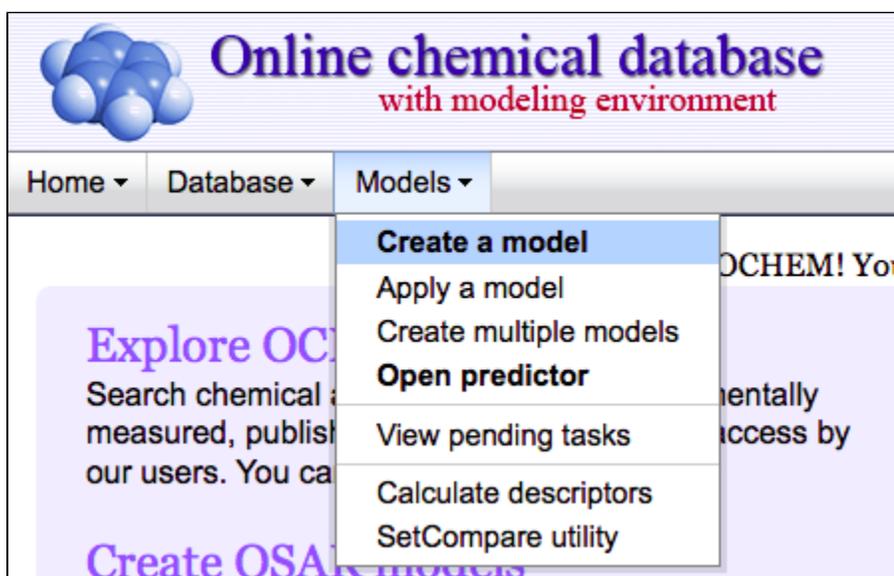


Creation of a single model

Model creation wizard

You can create a single model using web interface by accessing the "Models > Create a model" menu as shown below, which will open a wizard-like dialogue with a series of the necessary configuration steps.



First, you have to select a training and (optionally) a validation set. It is possible to have multiple validation sets. We assume that you have already [prepared the sets](#).

On the same screen, it is required to select:

- the unit of measurement for the model
- the desired machine learning method
- the desired validation protocol (cross validation, bagging or no validation)

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*): [BCF set \(training\) \[details\]](#)

Validation set #1: [BCF set \(test\) \[x\] \[details\]](#)

[Add a validation set](#)

The model will predict this property:

BCF using unit:

Choose the learning method:

Suggested modeling methods:

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

Models under development. (Do not use unless you are sure how to use):

- ANNC (Molecule-centric, experimental!) [W](#)
- BLASSO (Bayesian regression) [W](#)
- Consensus model (experimental) [W](#)
- KRR (Kernel Ridge Regression) [W](#)

Model validation

Validation method:

Number of folds:

Stratified cross-validation

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

The next screen allows to configure the preprocessing of molecular structures. The options include:

- Standardization (e.g., of nitro-groups)
- Neutralisation of ions
- Removal of salts (by keeping the counter-ion)
- Cleaning of structures (by converting it to SMILES and back to SD-file)

Model editor

Select model template and training set

Select the preferred data preprocessing options

Preprocessing of molecules (Chemaxon)

- Standardization 
- Neutralize 
- Remove salts 

- Clean structure 

<<Back

Next>>

Molecular descriptors

The next important step is the choice of molecular descriptors. OCHEM supports more than 20 descriptor packages provided by different vendors.

OCHEM policy is to integrate state-of-the art descriptors rather than to develop own solutions.

Select the molecular descriptors:

Suggested descriptors:

E-state [W](#)

E-State types:

Atom indices

Atom counts

Bonds indices

Bonds counts

Aromatize structures:

ALogPS (2) [W](#)

GSFragment (1138) [W](#)

Dragon v. 6.0 (4885/3D) [W](#)

ISIDA fragments [W](#)

ADRIANA.Code (211/3D) [W](#)

CDK descriptors (246/3D) [W](#)

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

MERA descriptors (529/3D) [W](#)

MERSY descriptors (42/3D) [W](#)

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

QNPR [W](#)

Spectrophores (144/3D) [W](#)

Additional or obsolete descriptors:

OEState [W](#)

MolPrint [W](#)

Dragon v. 5.4 (1630/3D) [W](#)

Dragon v. 5.5 (3190/3D) [W](#)

Structural alerts (ToxAlerts) [W](#)

MOPAC descriptors (21/3D) [W](#)

ShapeSignatures (3D) [W](#)

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

Custom descriptors from a file

AMBIT Descriptors [W](#)

ISIDA fragments (2011) [W](#)

Chiral Descriptors (/3D) [W](#)

Scaffold Hunter Descriptors [W](#)

Functional Groups [W](#)

ScrambledDragon (tmp) *Not supported by your installation*

ACD pKa (tmp) *Not supported by your installation*

Random descriptors (for testing) *Not supported by your installation*

ETM descriptors [W](#)

DockingDescriptors (pre-pre-alfa) [W](#) *Not supported by your installation*

Experimental values of other properties [W](#) *Not supported by your installation*

On the next step, we have to configure filtering of descriptors. The main filtering options include:

- elimination of constant or semi-constant descriptors
- unsupervised forward selection
- extraction of PCAs
- manual selection of the desired descriptors from a list

Model editor
Select model template and training set

Select filters of descriptors

Eliminate descriptors with less than unique values

Delete descriptors that have absolute values larger than

Delete descriptors that have variance smaller than

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

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

Perform principal component analysis

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

Normalisation parameters

Descriptors normalization

Values normalization

Machine learning method

Next, we have to configure the machine learning method. This step is method-specific. The screenshot below shows the configuration options for a neural network model.

Naturally, the other machine learning methods (e.g., linear regression, PLS, random forest, etc) will have different options.

Model editor

Select model template and training set

Configure ANN method

Training method:	<input type="text" value="SuperSAB"/>
Number of neurons in hidden layer	<input type="text" value="3"/>
Learning iterations (learning iterations)	<input type="text" value="1000"/>
Ensemble	<input type="text" value="64"/>
Disable ASNN	<input type="checkbox"/>
Additional Parameters (separated by comma)	<input type="text"/>

<<Back

Next>>

Initiating the calculations

Now, we are ready to start calculations. In the dialog below, we have to give our model a name, define the priority of this task and, finally, initiate the calculations.

We recommend using high and extra-high priorities only for fast tasks (e.g., models with less than 300-500 molecules in the training set).

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:

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)

<<Back

Start calculation>>

Discard

The next screen shows the progress of calculations. It is possible to fetch results later (in one hour, one day or more) from the registry of pending tasks.

Run model builder



Running the teacher - Task started

[cancel] [fetch result later]

<<Back
Next>>

The registry of pending tasks displays the status of all currently running calculation tasks. Once the task is ready, it will become "green", and it will be possible to fetch it by clicking a green checkbox icon on the right.

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

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

1 - 15 of 146 15 items on page 1 of 10 >>

Task type / Time started	Model	Property / Set	Method	Status	Priority	Details
Model training 2013-01-23 13:56:45	BCF_ASNN_[EState, ALogPS], 162292	BCF BCF set (training)	ANN	assigned	high	Task started terminate

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

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

1 - 15 of 146 15 items on page 1 of 10 >>

Task type / Time started	Model	Property / Set	Method	Status	Priority	Details
Model training 2013-01-23 13:56:45	BCF_ASNN_[EState, ALogPS], 162292	BCF BCF set (training)	ANN	ready	high	- recalculate

After you fetch the calculation task you will be presented with a model profile.

The model profile displays basic model statistics (RMSE, R2, etc), scatter plot, applicability domain plots, etc. On this step, you can either save your model or discard it in case if the model performance is unsatisfactory.

Model editor

Select model template and training set

Save the model

Please enter your model's name:

Overview

Applicability domain

Model name: BCF_ASNN_[EState, ALogPS], 162292 [rename]

Public ID is 37902450

Predicted property: BCF

Training method: ANN

[EState, ALogPS]

Correl. limit: 0.95 Variance threshold: 0.01,

Maximum value: 999999,

Supersab, 1000 iterations, 3 neurons

ensemble=64

5-fold cross-validation

-

61 pre-filtered descriptors

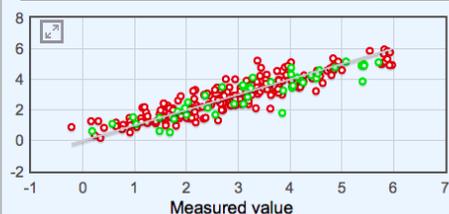
Supersab, 1000 iterations, 3 neurons

ensemble=64

Calculated in 37 seconds

Size: 54 Kb

Data Set	#	R2	q2	RMSE	MAE
• Training set: BCF set (training)	192 records	0.83 ± 0.04	0.82 ± 0.05	0.57 ± 0.05	0.45 ± 0.05
• Test set: BCF set (test) [x]	46 records	0.8 ± 0.1	0.8 ± 0.1	0.6 ± 0.2	0.5 ± 0.1



[Download model statistics in Excel format](#) [Create a copy of this model](#) [View configuration XML](#) [Export configuration XML](#)

Save

Discard

Accessing your model

After you save your model, it will become available from the browser of models:

Online chemical database
with modeling environment

Home ▾ Database ▾ **Models ▾** Moderation ▾

Create a model
Apply a model
Create multiple models
Open predictor
View pending tasks
Calculate descriptors
SetCompare utility

Pending tasks
The overview of all running tasks

All tasks types
1 - 15 of 145
Task type / Time started

The browser of models displays all the publicly available models developed by other OCHEM users as well as your own private models. Here, it is possible to search for models by various criteria.

The displayed information includes:

- model name
- training and validation sets

- machine learning method

Each available model has a model profile described above, which contains the detailed statistics, the scatter plots and confusion matrices for the classification models.

Models applicator 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: Order by: [refresh]

1 - 15 of 75 items on page of 5 > >>

    BCF_ASNN_[Estate, ALogPS], 162292	predicts BCF using BCF set (training) (192) validated by BCF set (test) (46)	ANN	2013-01-23
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------	-----	------------