Comprehensive modeling with multiple models

OCHEM allows creating multiple models with different configuration options. Hundreds of models can be created with a few clicks.
The utility for multiple models creation and management is referred to as "**comprehensive modeling**" utility.

Creation of multiple models

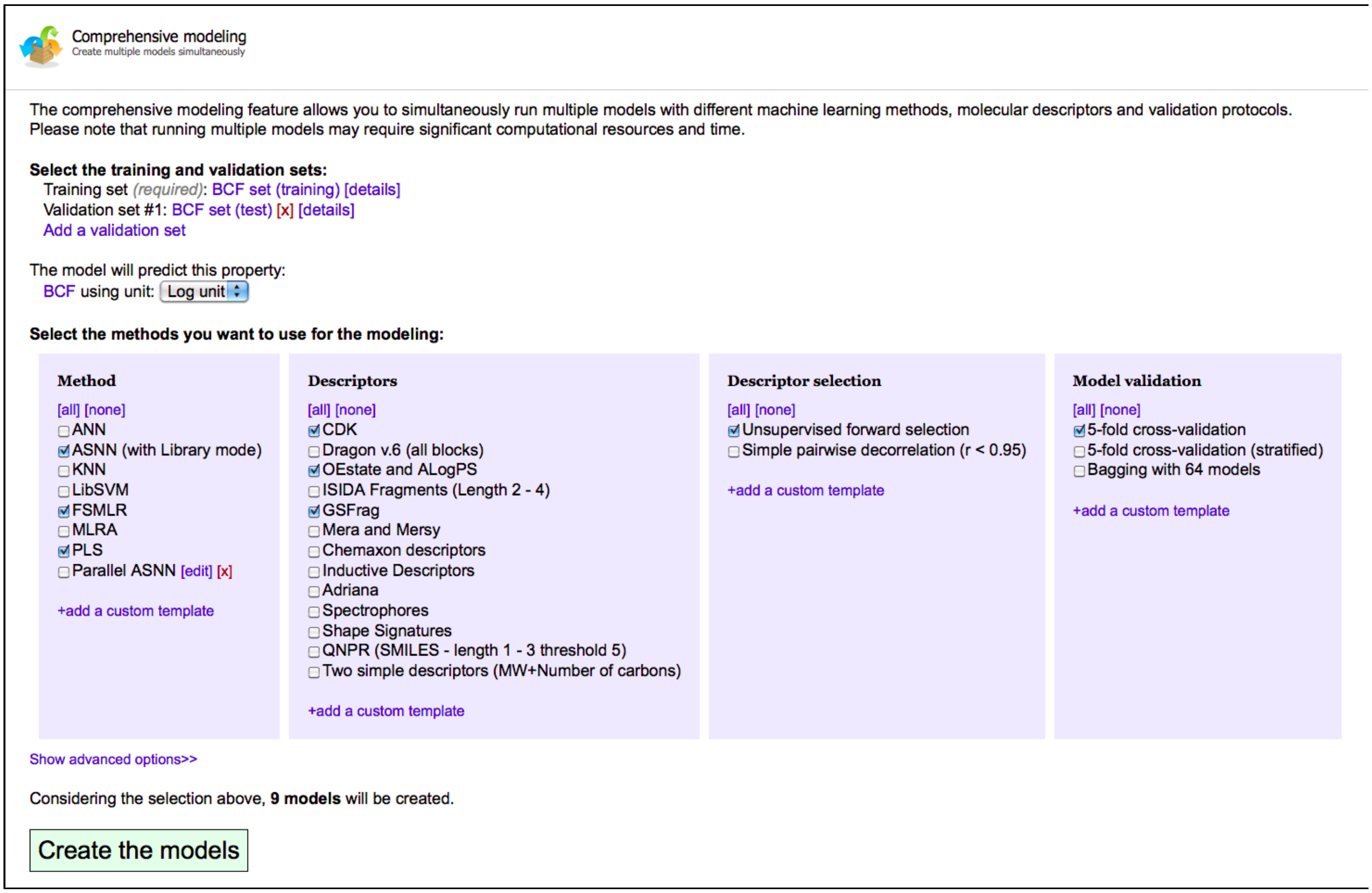
To access the comprehensive modelling dialogue, select the menu "Models > Create multiple models".

Select the necessary training and validation sets.

Select the predefined configuration templates for:

- molecular descriptors
- machine learning methods
- descriptors filtering
- validation protocols

Each combination of the selected templates will be used, therefore, if you select 2 types of descriptors, 3 machine learning methods, 1 descriptor filtering options and 2 validation options, then 2x3x1x2 = **12 models** will be created.



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)*: 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: Log unit

Select the methods you want to use for the modeling:

Method
[all] [none]
☐ ANN
☑ ASNN (with Library mode)
☐ KNN
☐ LibSVM
☑ FSMLR
☐ MLRA
☑ PLS
☐ Parallel ASNN [edit] [x]
+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
☐ Shape Signatures
☐ 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

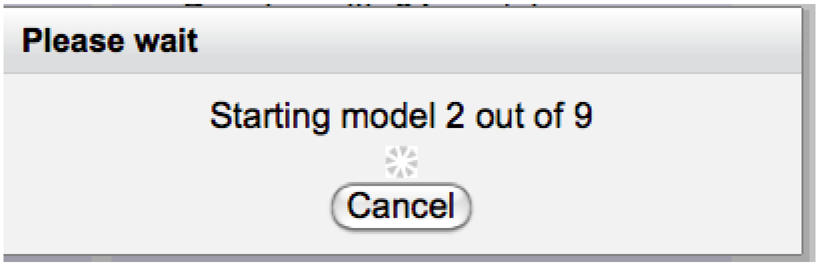
Model validation
[all] [none]
☑ 5-fold cross-validation
☐ 5-fold cross-validation (stratified)
☐ Bagging with 64 models
+add a custom template

Show advanced options>>

Considering the selection above, **9 models** will be created.

Create the models

Wait until the models have been started:

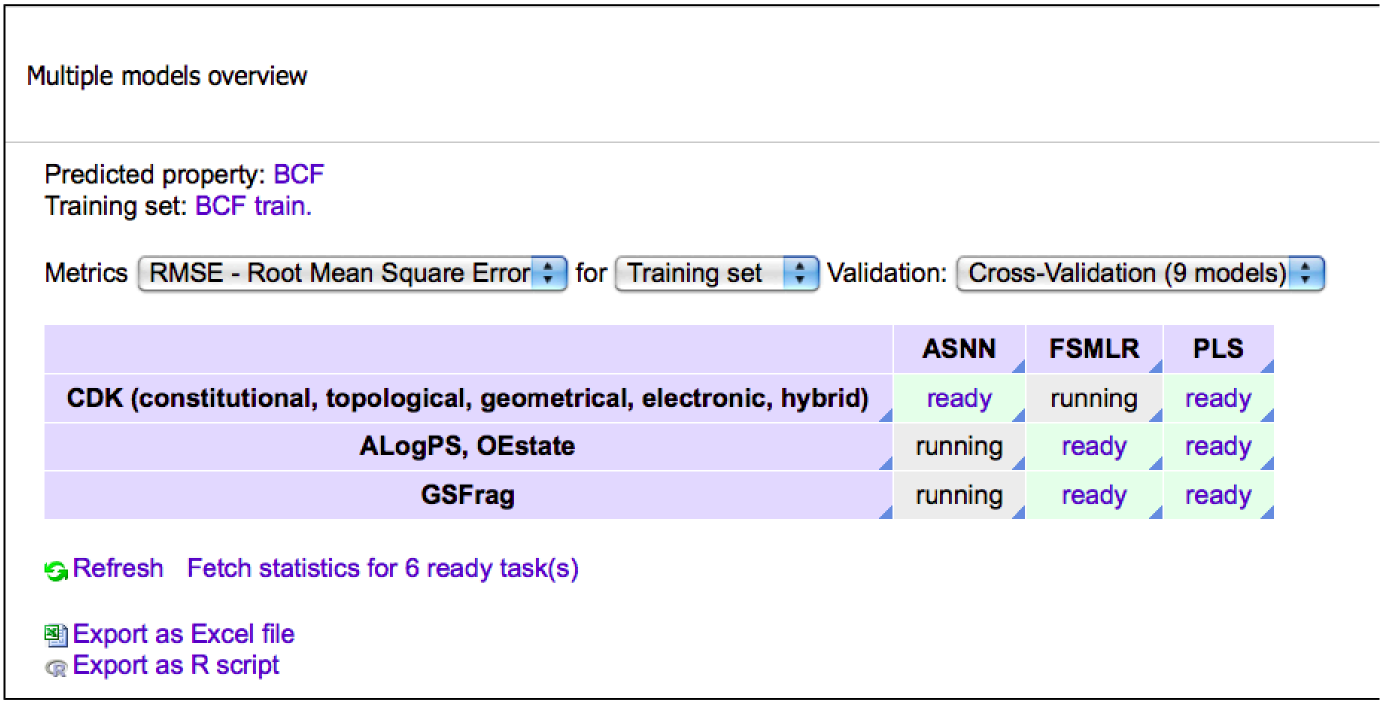


Please wait

Starting model 2 out of 9

Multiple models overview

The **multiple models overview** page will display the status of all the models, both running and completed:



Multiple models overview

Predicted property: BCF
Training set: BCF train.

Metrics: RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (9 models)

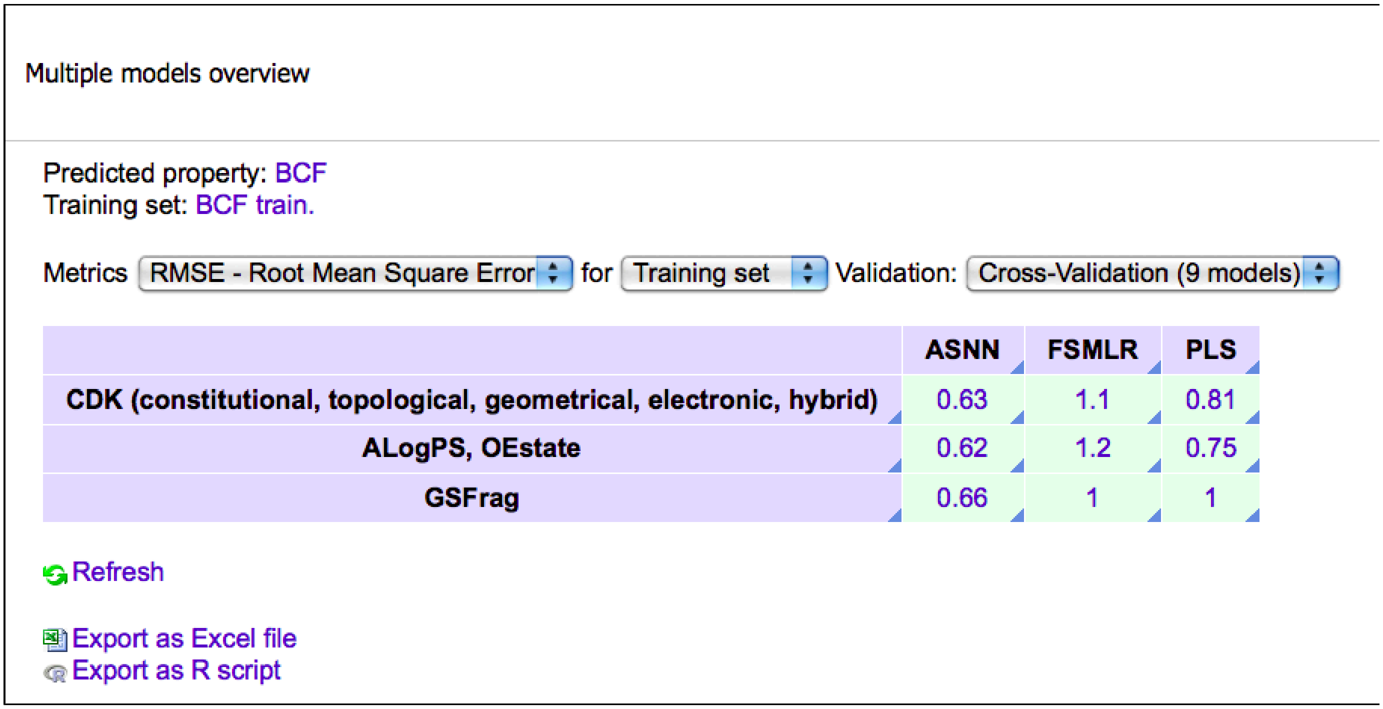
	ASNN	FSMLR	PLS
CDK (constitutional, topological, geometrical, electronic, hybrid)	ready	running	ready
ALogPS, OEstate	running	ready	ready
GSFrag	running	ready	ready

When the models are ready, we fetch the statistics by clicking the appropriate lick and overview the predictive performance of the models.
The multiple models overview allows to compare models by:

- RMSE
- R2
- Mean absolute error (MAE)

It is possible to compare the models by their performances for both training and validation sets.

The screenshot below displays RMSE values for 9 models developed for BCF (bio-concentration factor) endpoint based on the "BCF train" basket.



Multiple models overview

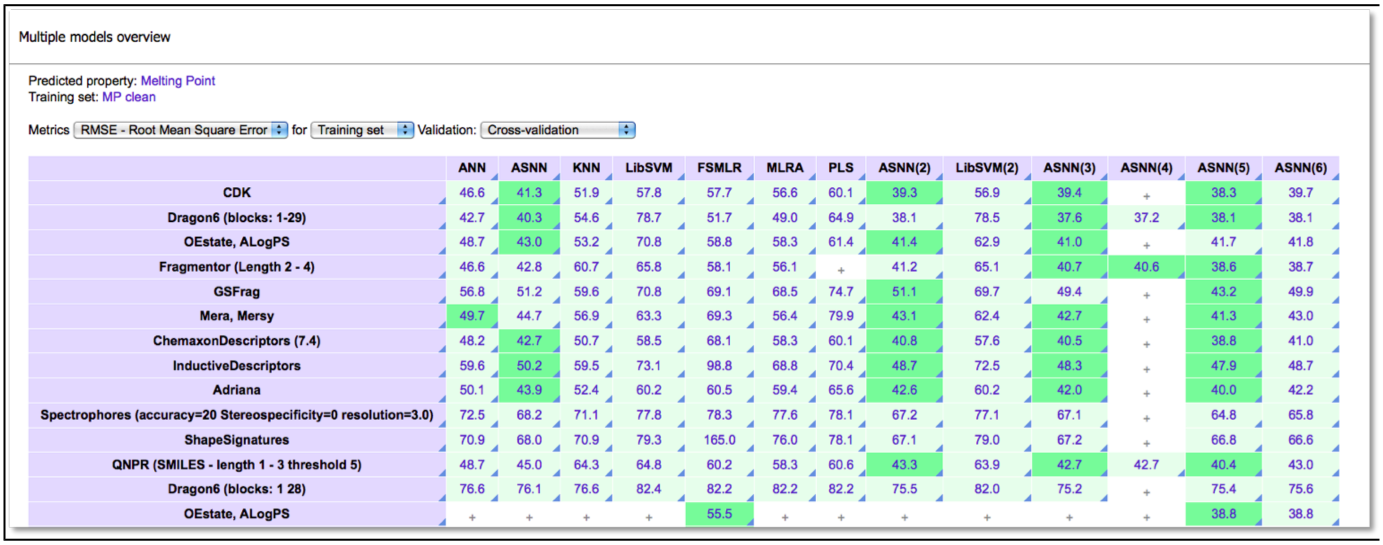
Predicted property: BCF
Training set: BCF train.

Metrics RMSE - Root Mean Square Error for Training set Validation: Cross-Validation (9 models)

	ASNN	FSMLR	PLS
CDK (constitutional, topological, geometrical, electronic, hybrid)	0.63	1.1	0.81
ALogPS, OEstate	0.62	1.2	0.75
GSFrag	0.66	1	1

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

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



Multiple models overview

Predicted property: Melting Point
Training set: MP clean

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

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