Supplementary Information to article:

The development of models to predict melting and pyrolysis point data associated with several hundreds thousands compounds mined from patents

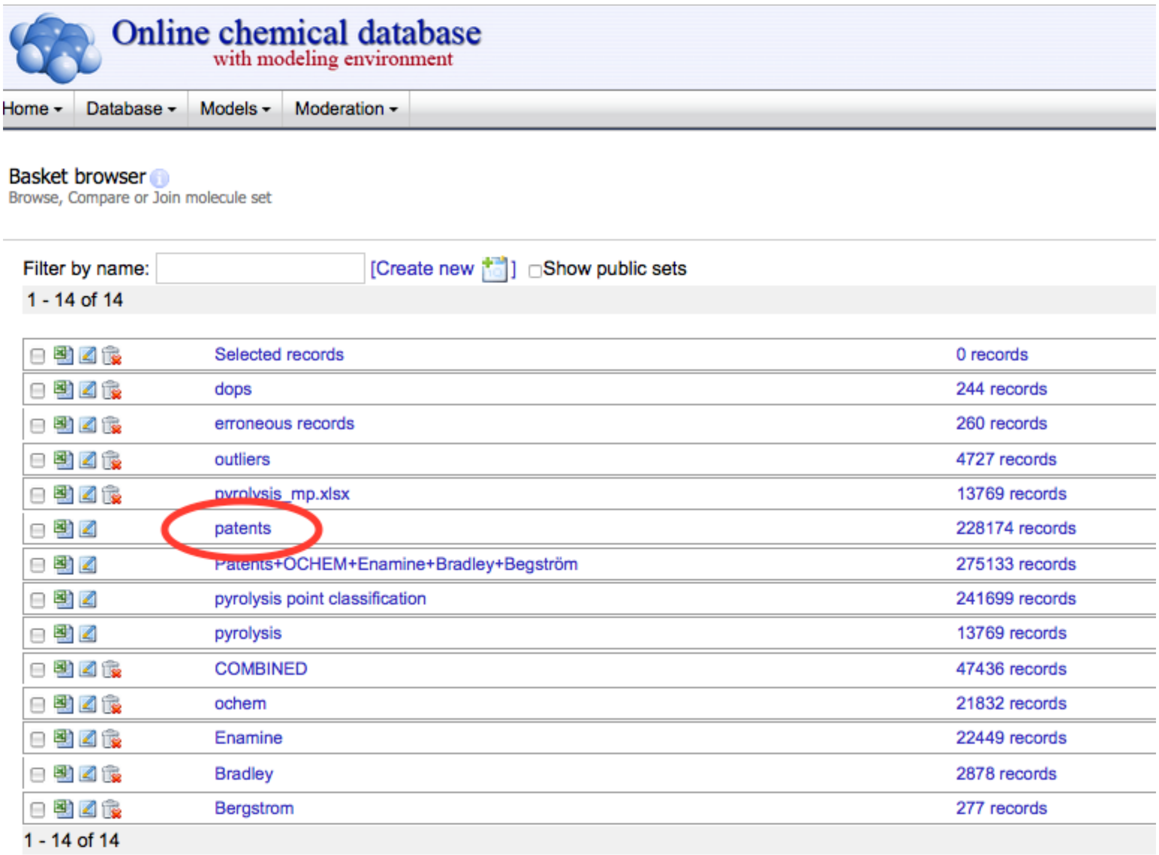
Igor V. Tetko,[1,*] Daniel Lowe[2] and Antony J. Williams[3]

1) Helmholtz Zentrum München für Gesundheit und Umwelt (HMGU), Institute of Structural Biology, Ingolstaedter Landstrasse 1, b. 60w, D-85764 Neuherberg, Germany and BigChem GmbH, D-85764 Neuherberg, Germany i.tetko@helmholtz-muenchen.de
2) NextMove Software Limited, Innovation Centre (Unit 23), Cambridge Science Park, Cambridge CB4 0EY, UK, daniel@nextmovesoftware.com
3) ChemConnector Inc., 904 Tamaras Circle, Wake Forest, NC-27587, USA, Williams.Antony@epa.gov

Description of the protocols used to develop the Melting Point consensus model.

Overview

Below, we summarize steps used for the development of the consensus model. The illustrative Figures are followed by a brief description of the respective step.



1 - 14 of 14

Selected records	0 records
dops	244 records
erroneous records	260 records
outliers	4727 records
pyrolysis_mp.xlsx	13769 records
patents	228174 records
Patents+OCHEM+Enamine+Bradley+Begström	275133 records
pyrolysis point classification	241699 records
pyrolysis	13769 records
COMBINED	47436 records
ochem	21832 records
Enamine	22449 records
Bradley	2878 records
Bergstrom	277 records

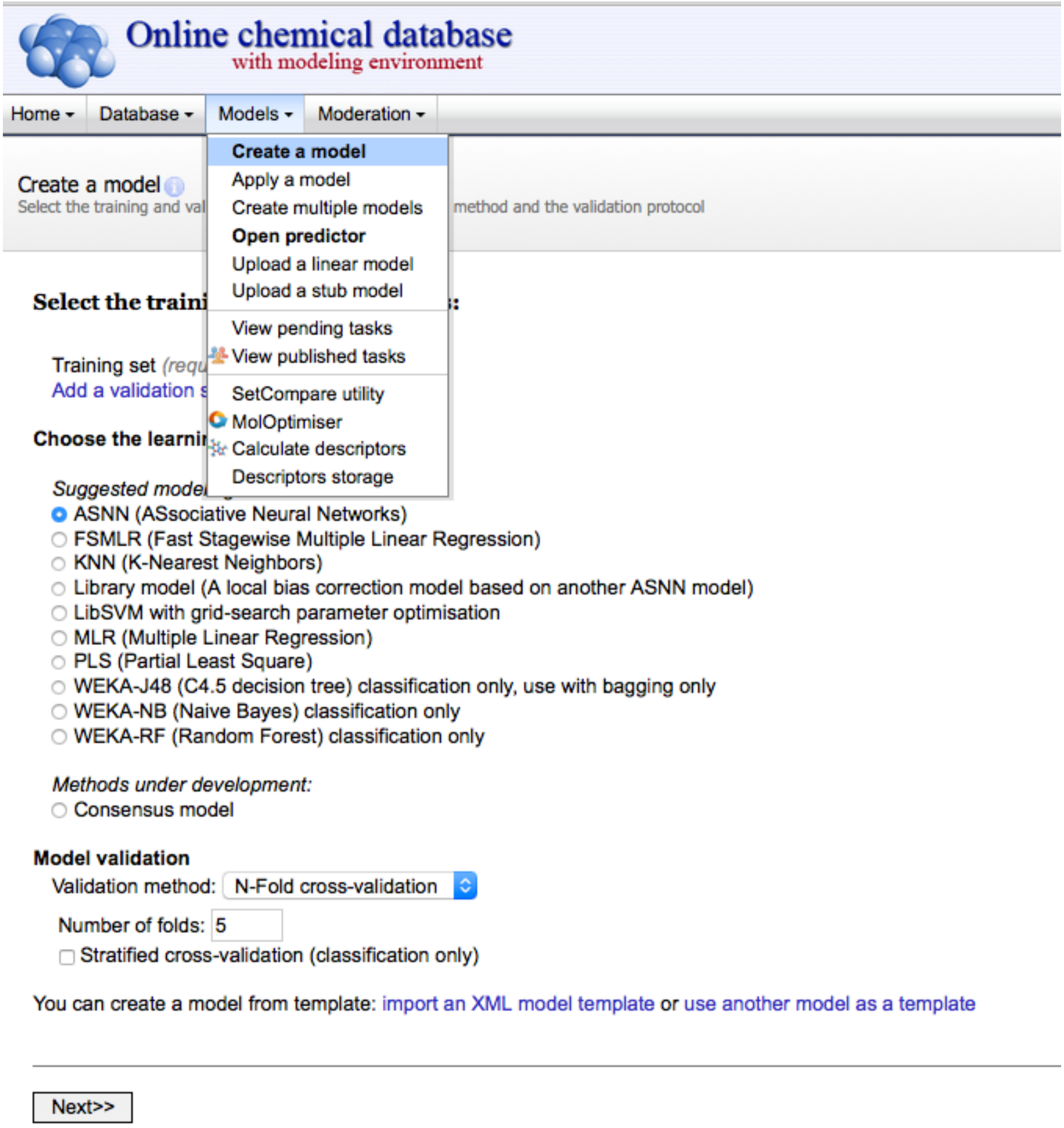
1 - 14 of 14

Step 1. Data upload.

The data for the Melting Point (MP) model are uploaded in the OCHEM database (for these steps see tutorial at https://www.youtube.com/watch?v=vdJWiS4wSaQ) and formed the basket “patents”.

Step 2. Identification of LibSVM parameters.

A preliminary analysis using a grid search (see Methods) was run to select the optimal parameters for the SVM model. It included the following several substeps.



Create a model

Select the training and val… method and the validation protocol

Select the traini…:

Training set *(requ…*

Add a validation s…

Choose the learni…

Suggested model…

● ASNN (ASsociative Neural Networks)
○ FSMLR (Fast Stagewise Multiple Linear Regression)
○ KNN (K-Nearest Neighbors)
○ Library model (A local bias correction model based on another ASNN model)
○ LibSVM with grid-search parameter optimisation
○ MLR (Multiple Linear Regression)
○ PLS (Partial Least Square)
○ WEKA-J48 (C4.5 decision tree) classification only, use with bagging only
○ WEKA-NB (Naive Bayes) classification only
○ WEKA-RF (Random Forest) classification only

Methods under development:

○ Consensus model

Model validation

Validation method: N-Fold cross-validation

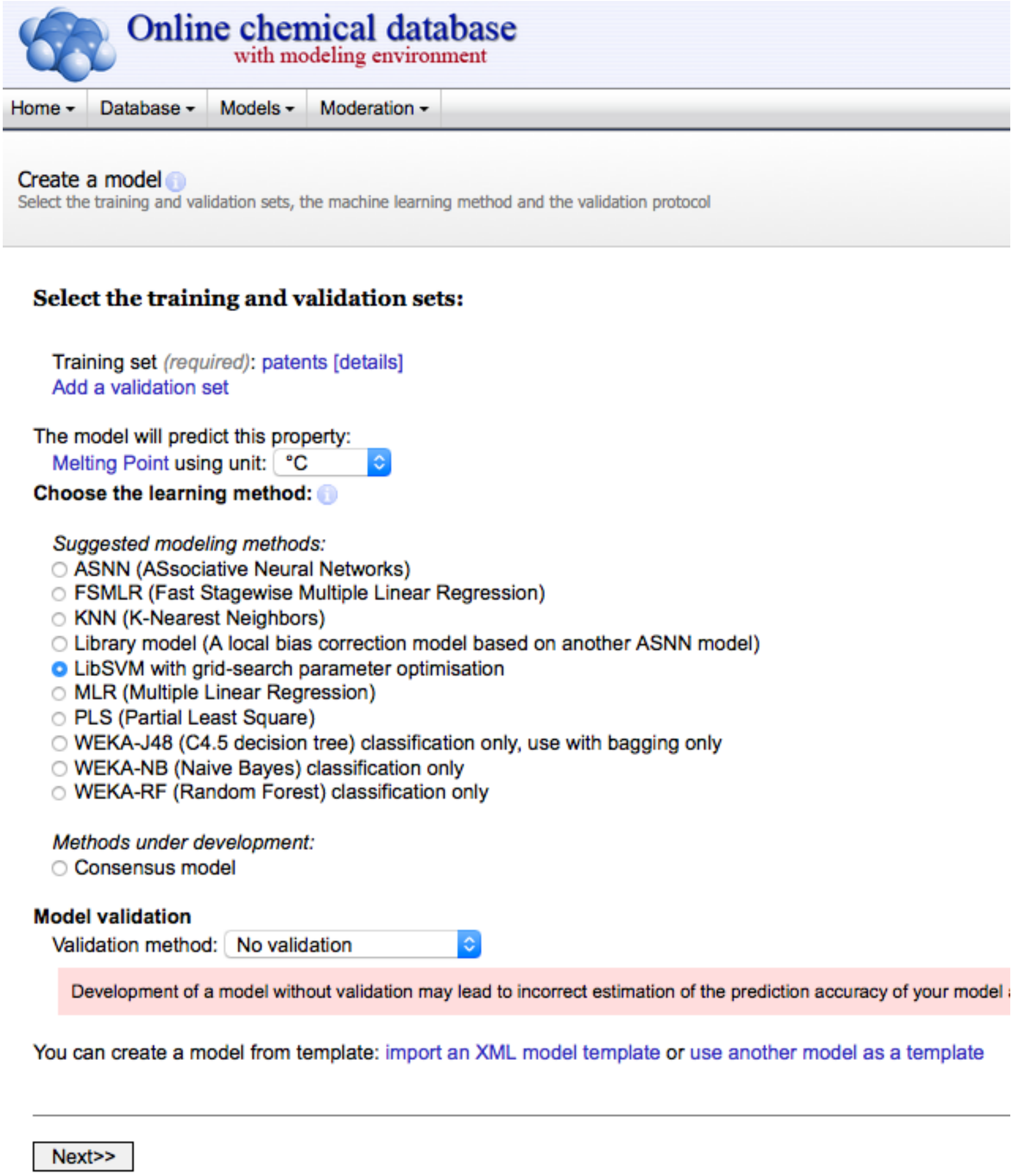
Number of folds: 5

□ Stratified cross-validation (classification only)

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

Next>>

Step 2.1 Starting model development.



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)*: patents [details]

Add a validation set

The model will predict this property:

Melting Point using unit: °C

Choose the learning method:

Suggested modeling methods:

○ ASNN (ASsociative Neural Networks)

○ FSMLR (Fast Stagewise Multiple Linear Regression)

○ KNN (K-Nearest Neighbors)

○ Library model (A local bias correction model based on another ASNN model)

● LibSVM with grid-search parameter optimisation

○ MLR (Multiple Linear Regression)

○ PLS (Partial Least Square)

○ WEKA-J48 (C4.5 decision tree) classification only, use with bagging only

○ WEKA-NB (Naive Bayes) classification only

○ WEKA-RF (Random Forest) classification only

Methods under development:

○ Consensus model

Model validation

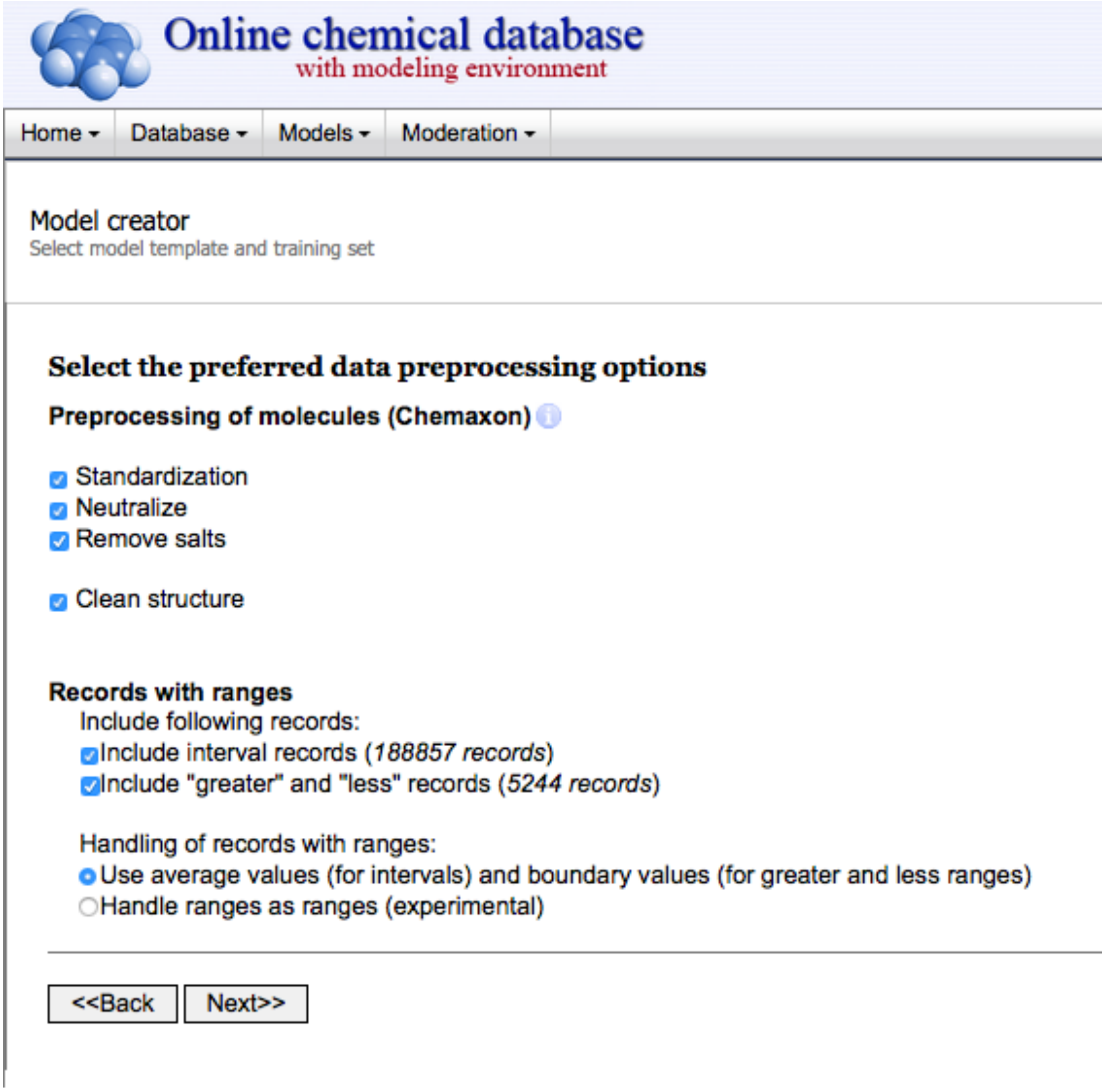
Validation method: No validation

Development of a model without validation may lead to incorrect estimation of the prediction accuracy of your model

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

2.2 Selection of the dataset, method and validation protocol to identify the best set of LibSVM parameters.

Since the calculations in this step required very large CPU resources and LibSVM parameters were optimized for the whole set, “no validation” method was used. This method is normally not recommended since it can result in data overfitting. In this case only three LibSVM parameters were optimized for N = 228k compounds and thus there was no problem with overfitting. Notice that these calculations required more than 600 CPU-days of calculations.



Model creator
Select model template and training set

Select the preferred data preprocessing options

Preprocessing of molecules (Chemaxon)

☑ Standardization
☑ Neutralize
☑ Remove salts

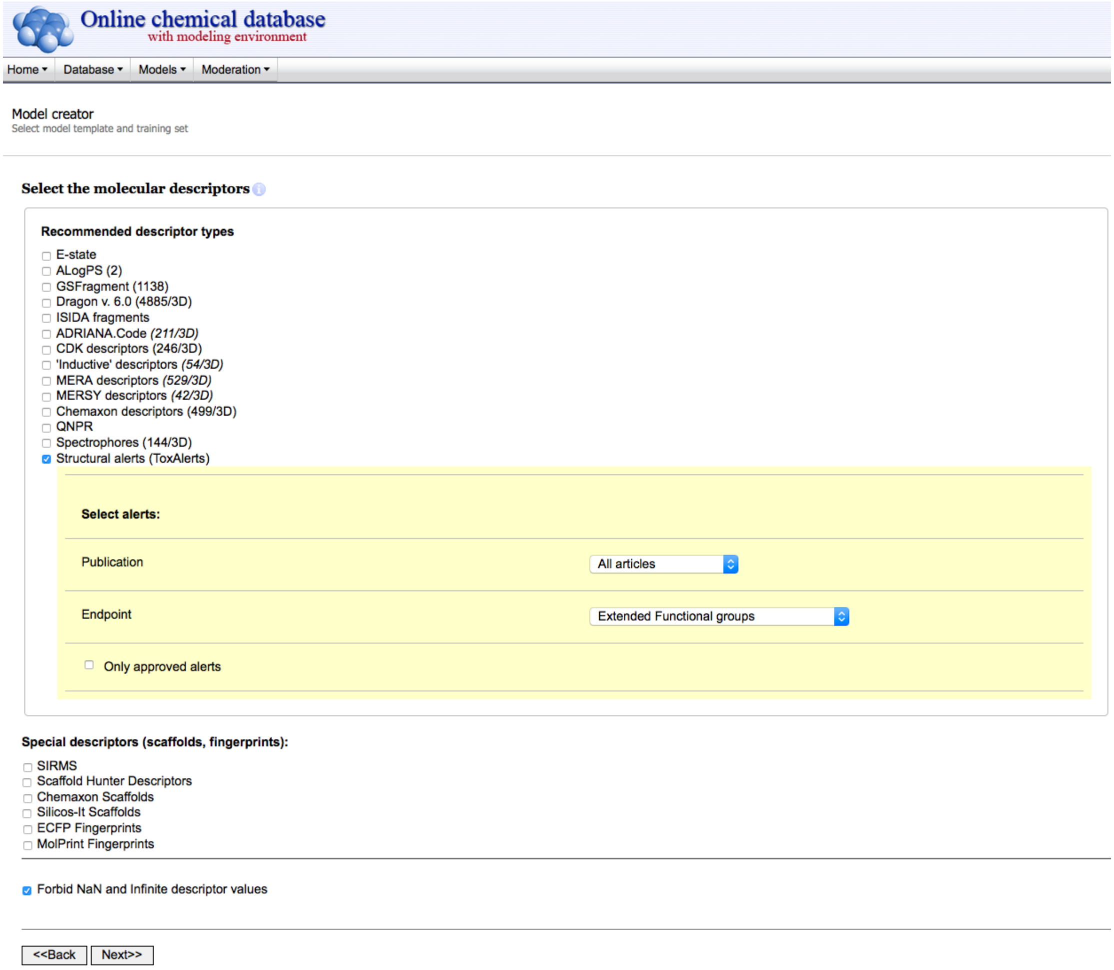
☑ Clean structure

Records with ranges
Include following records:
☑Include interval records (*188857 records*)
☑Include "greater" and "less" records (*5244 records*)

Handling of records with ranges:
●Use average values (for intervals) and boundary values (for greater and less ranges)
○Handle ranges as ranges (experimental)

2.3 Selection of preprocessing options.

The options included use of records with intervals and ranges for model development.



Model creator
Select model template and training set

Select the molecular descriptors

Recommended descriptor types

- □ E-state
- □ 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)
- ☑ Structural alerts (ToxAlerts)

Select alerts:

Publication	All articles
Endpoint	Extended Functional groups

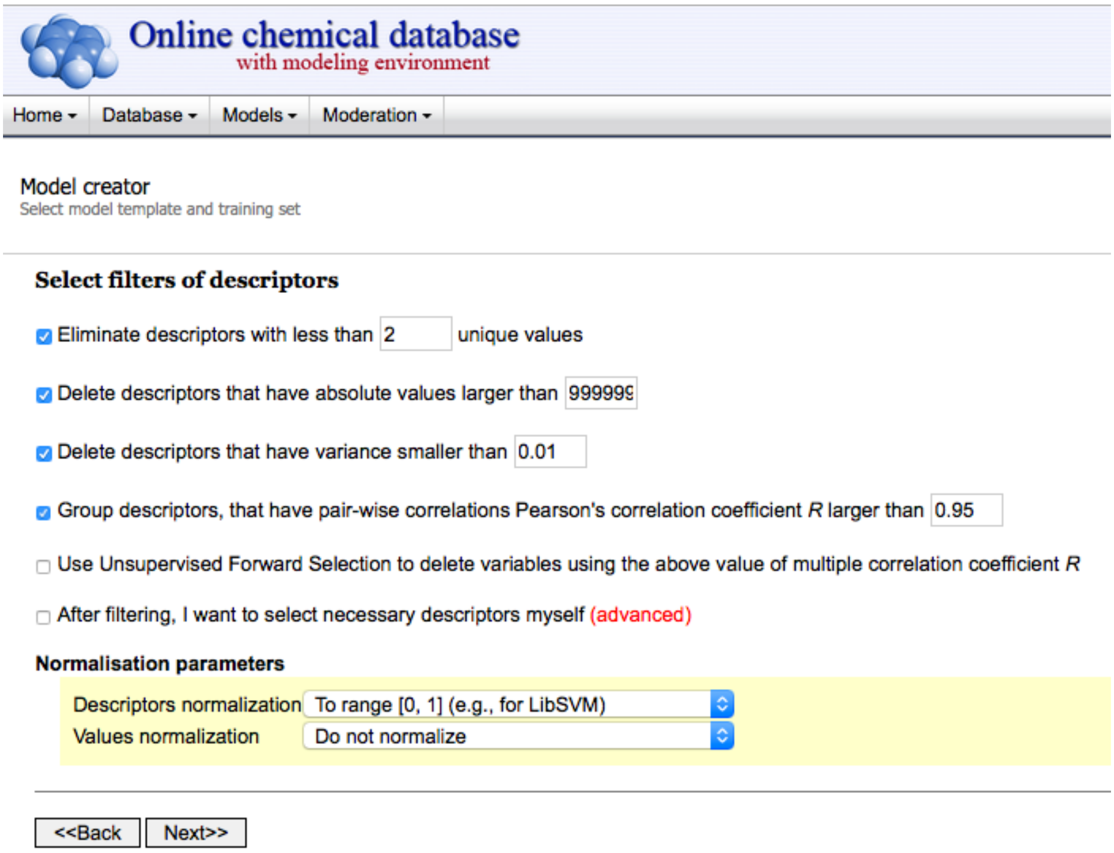
□ Only approved alerts

Special descriptors (scaffolds, fingerprints):

- □ SIRMS
- □ Scaffold Hunter Descriptors
- □ Chemaxon Scaffolds
- □ Silicos-It Scaffolds
- □ ECFP Fingerprints
- □ MolPrint Fingerprints

☑ Forbid NaN and Infinite descriptor values

2.4 Selection of Extended Functional Groups (EFGs) as descriptors.



Model creator
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*

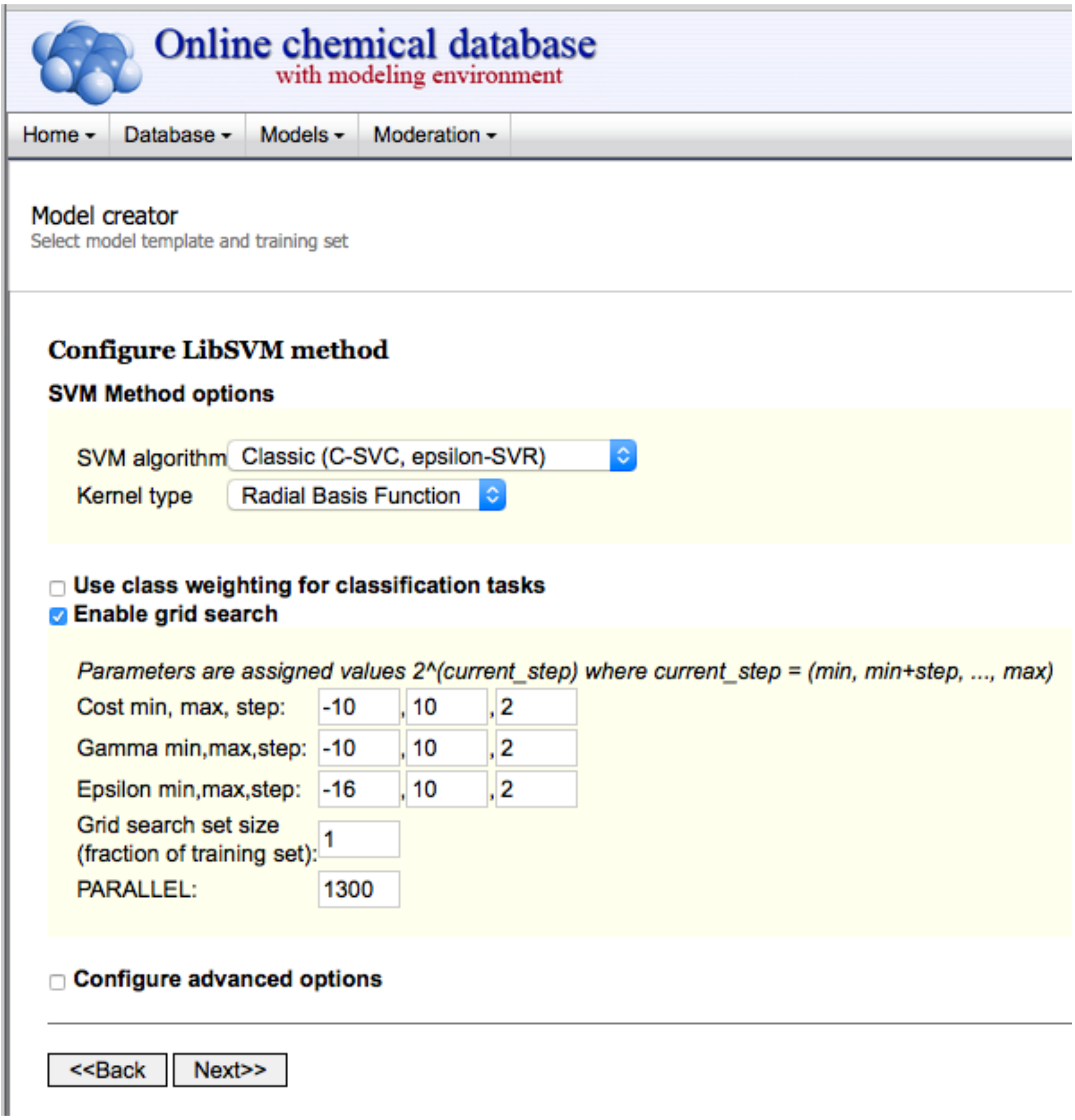
□ After filtering, I want to select necessary descriptors myself (advanced)

Normalisation parameters

Descriptors normalization: To range [0, 1] (e.g., for LibSVM)
Values normalization: Do not normalize

2.5 The descriptor unsupervised filtering options

The normalization to [0,1] interval was selected. Other pre-processing options were used as default values.



Model creator
Select model template and training set

Configure LibSVM method

SVM Method options

SVM algorithm: Classic (C-SVC, epsilon-SVR)
Kernel type: Radial Basis Function

□ **Use class weighting for classification tasks**

☑ **Enable grid search**

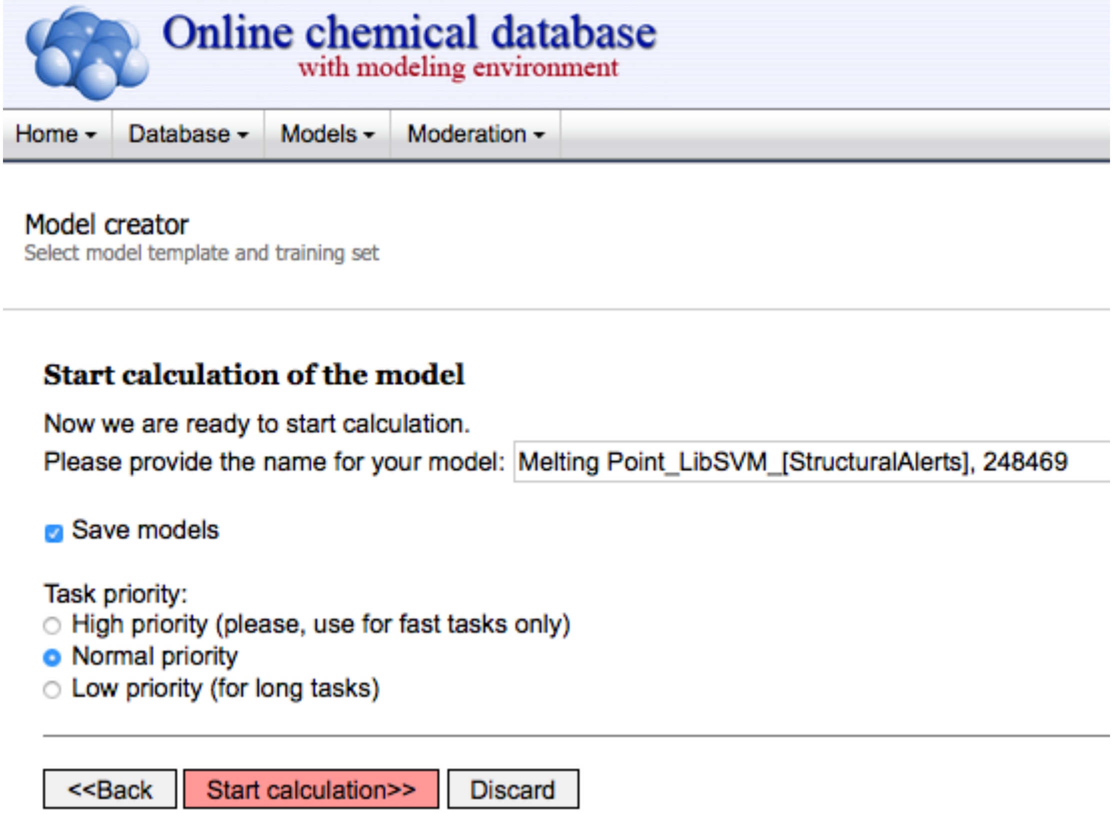
Parameters are assigned values 2^(current_step) where current_step = (min, min+step, ..., max)

Cost min, max, step:	-10	10	2
Gamma min,max,step:	-10	10	2
Epsilon min,max,step:	-16	10	2
Grid search set size (fraction of training set):	1		
PARALLEL:	1300		

□ **Configure advanced options**

2.6 Grid search to detect optimized parameters.

The calculations were configured to run in parallel on 1300 servers (option PARALLEL) and using the whole training set (fraction = 1).



Model creator
Select model template and training set

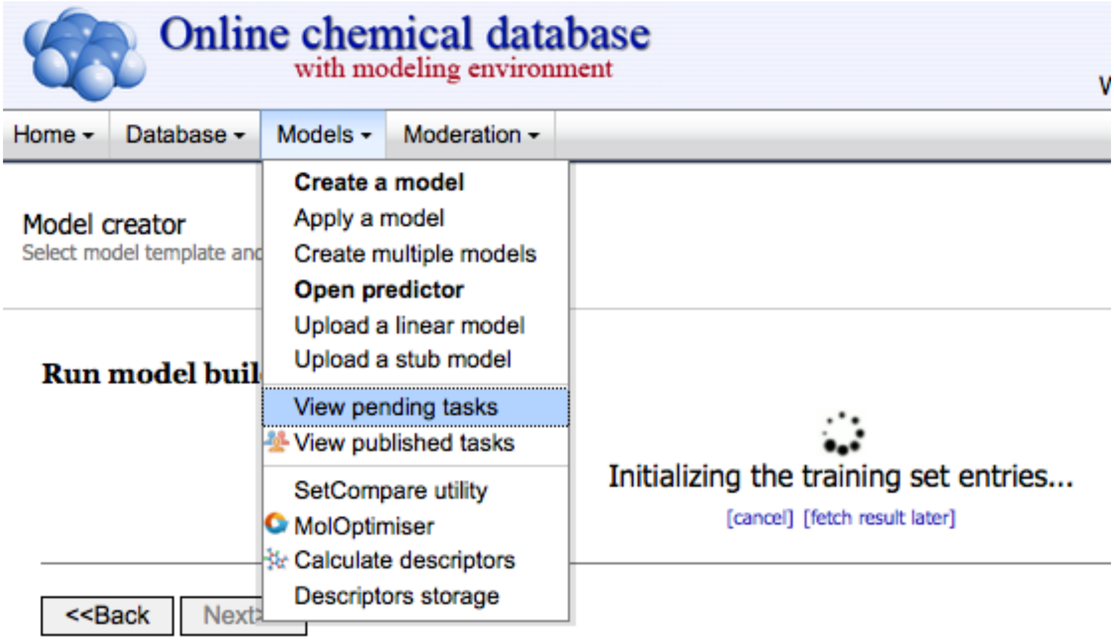
Start calculation of the model

Now we are ready to start calculation.
Please provide the name for your model: Melting Point_LibSVM_[StructuralAlerts], 248469

☑ Save models

Task priority:
○ High priority (please, use for fast tasks only)
● Normal priority
○ Low priority (for long tasks)

2.7 The calculations are launched.



Model creator

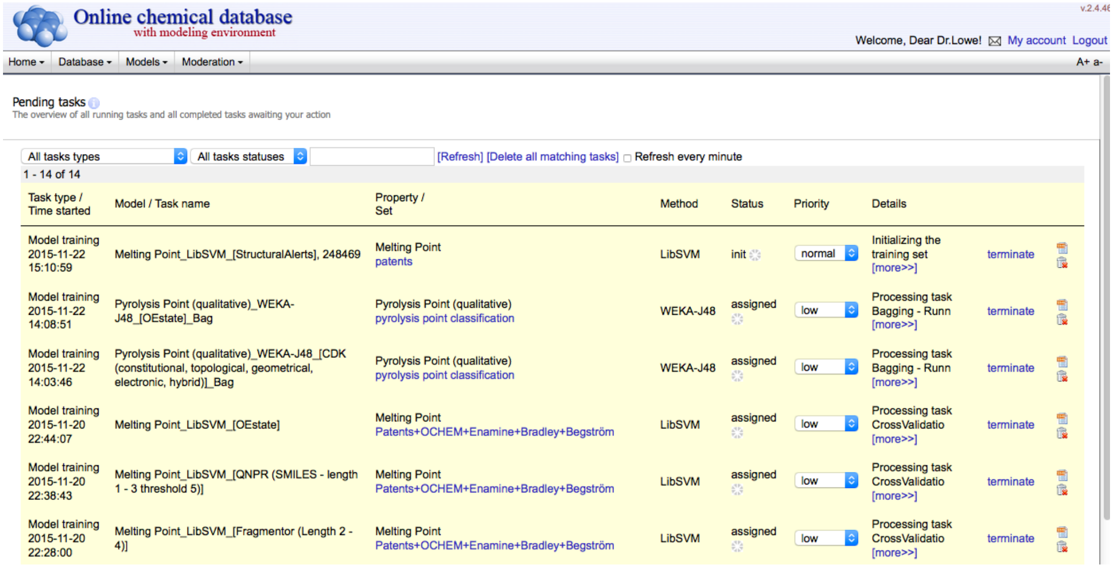
Select model template and

Run model buil

Initializing the training set entries...

[cancel] [fetch result later]

2.8 Selection of Pending task window to monitor the execution of tasks.



Pending tasks

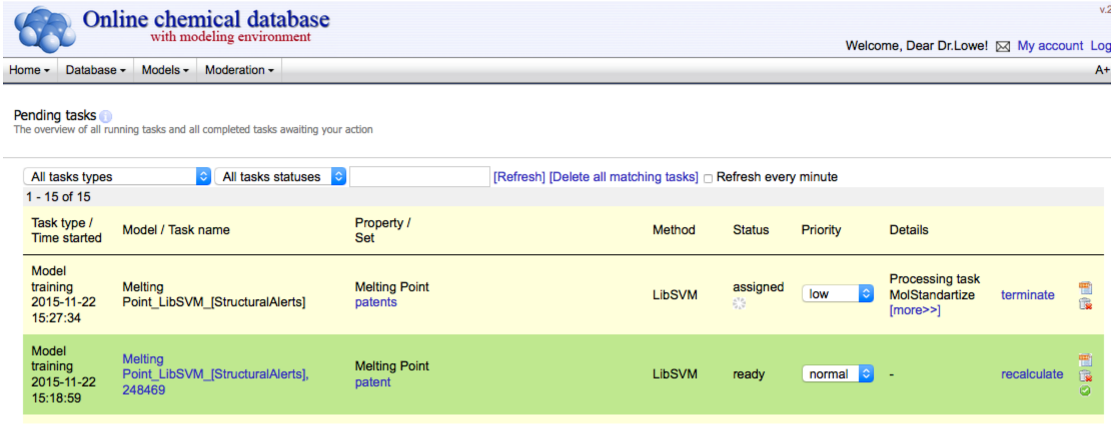
The overview of all running tasks and all completed tasks awaiting your action

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

1 - 14 of 14

Task type / Time started	Model / Task name	Property / Set	Method	Status	Priority	Details		
Model training 2015-11-22 15:10:59	Melting Point_LibSVM_[StructuralAlerts], 248469	Melting Point patents	LibSVM	init	normal	Initializing the training set [more>>]	terminate	
Model training 2015-11-22 14:08:51	Pyrolysis Point (qualitative)_WEKA-J48_[OEstate]_Bag	Pyrolysis Point (qualitative) pyrolysis point classification	WEKA-J48	assigned	low	Processing task Bagging - Runn [more>>]	terminate	
Model training 2015-11-22 14:03:46	Pyrolysis Point (qualitative)_WEKA-J48_[CDK (constitutional, topological, geometrical, electronic, hybrid)]_Bag	Pyrolysis Point (qualitative) pyrolysis point classification	WEKA-J48	assigned	low	Processing task Bagging - Runn [more>>]	terminate	
Model training 2015-11-20 22:44:07	Melting Point_LibSVM_[OEstate]	Melting Point Patents+OCHEM+Enamine+Bradley+Begström	LibSVM	assigned	low	Processing task CrossValidatio [more>>]	terminate	
Model training 2015-11-20 22:38:43	Melting Point_LibSVM_[QNPR (SMILES - length 1 - 3 threshold 5)]	Melting Point Patents+OCHEM+Enamine+Bradley+Begström	LibSVM	assigned	low	Processing task CrossValidatio [more>>]	terminate	
Model training 2015-11-20 22:28:00	Melting Point_LibSVM_[Fragmentor (Length 2 - 4)]	Melting Point Patents+OCHEM+Enamine+Bradley+Begström	LibSVM	assigned	low	Processing task CrossValidatio [more>>]	terminate	

2.9 Monitoring of the calculations in the PendingTask browser.


Online chemical database
with modeling environment
Welcome, Dear Dr.Lowe! My account Log
Home ▾ Database ▾ Models ▾ Moderation ▾
A+

Pending tasks

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

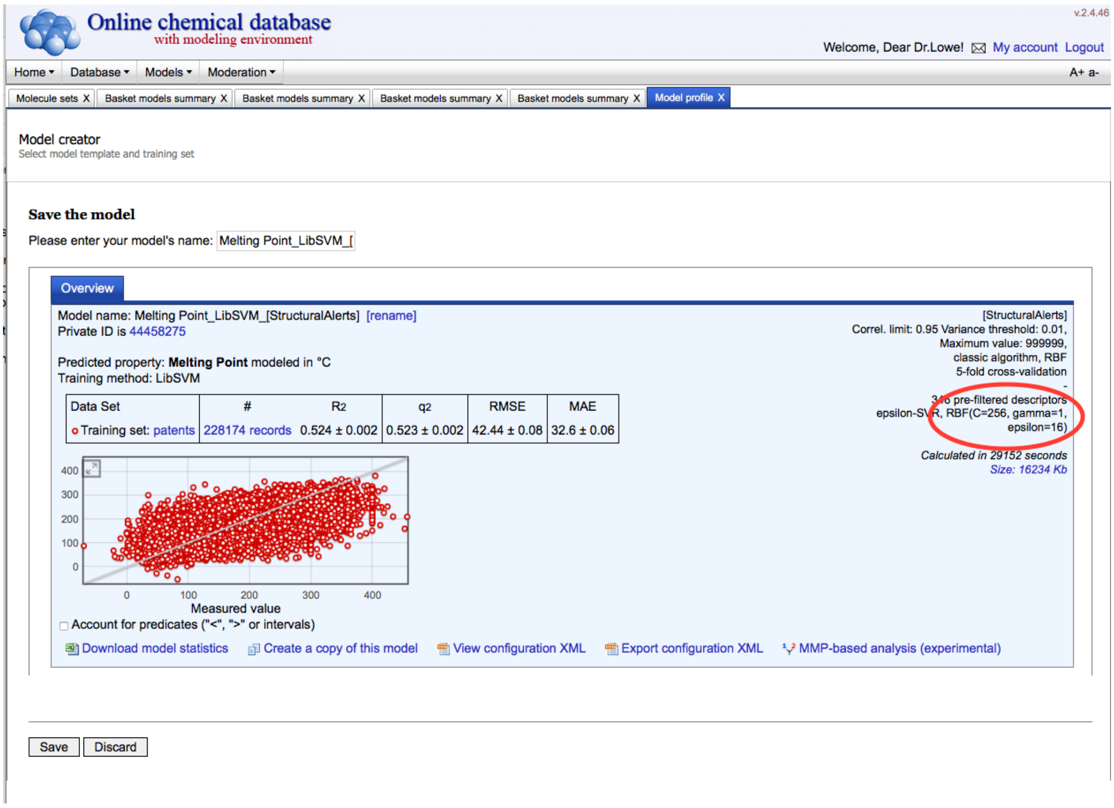
All tasks types | All tasks statuses | [Refresh] [Delete all matching tasks] □ Refresh every minute

1 - 15 of 15

Task type / Time started	Model / Task name	Property / Set	Method	Status	Priority	Details		
Model training 2015-11-22 15:27:34	Melting Point_LibSVM_[StructuralAlerts]	Melting Point patents	LibSVM	assigned	low	Processing task MolStandartize [more>>]	terminate	
Model training 2015-11-22 15:18:59	Melting Point_LibSVM_[StructuralAlerts], 248469	Melting Point patent	LibSVM	ready	normal	-	recalculate	

2.10 The model is calculated.

The optimal parameters for the training set were calculated. They could be exported using XML icon at the left side of model row.



Model creator

Select model template and training set

Save the model

Please enter your model's name: Melting Point_LibSVM_[

Model name: Melting Point_LibSVM_[StructuralAlerts] [rename]
Private ID is 44458275

Predicted property: **Melting Point** modeled in °C
Training method: LibSVM

[StructuralAlerts]
Correl. limit: 0.95 Variance threshold: 0.01,
Maximum value: 999999,
classic algorithm, RBF
5-fold cross-validation
-
3/6 pre-filtered descriptors
epsilon-SVR, RBF(C=256, gamma=1,
epsilon=16)

Calculated in 29152 seconds
Size: 16234 Kb

Data Set	#	R2	q2	RMSE	MAE
○ Training set: patents	228174 records	0.524 ± 0.002	0.523 ± 0.002	42.44 ± 0.08	32.6 ± 0.06

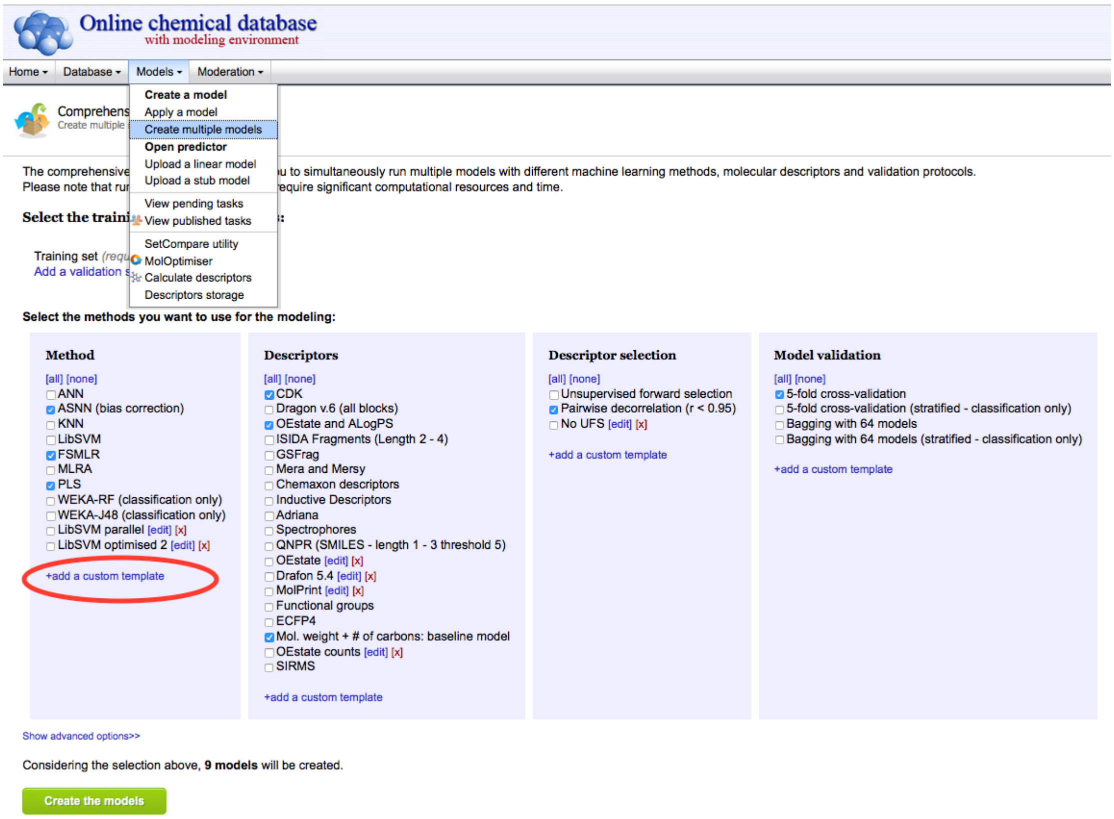
□ Account for predicates ("<", ">" or intervals)

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

Save | Discard

2.11 The model is open (green button on the previous plot) and saved.

The optimized parameters for LibSVM are shown. They are part of the model template, which can be exported using the “Export configuration XML” button. The exported template can be used to develop new models (see step 2.1 – “import an XML model template”) using the same setting.



The comprehensive u to simultaneously run multiple models with different machine learning methods, molecular descriptors and validation protocols.
Please note that rur equire significant computational resources and time.

Select the traini :

Training set *(requ*
Add a validation s

Select the methods you want to use for the modeling:

Method

[all] [none]
□ ANN
☑ ASNN (bias correction)
□ KNN
□ LibSVM
☑ FSMLR
□ MLRA
☑ PLS
□ WEKA-RF (classification only)
□ WEKA-J48 (classification only)
□ LibSVM parallel [edit] [x]
□ LibSVM optimised 2 [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
□ QNPR (SMILES - length 1 - 3 threshold 5)
□ OEstate [edit] [x]
□ Drafon 5.4 [edit] [x]
□ MolPrint [edit] [x]
□ Functional groups
□ ECFP4
☑ Mol. weight + # of carbons: baseline model
□ OEstate counts [edit] [x]
□ SIRMS

+add a custom template

Descriptor selection

[all] [none]
□ Unsupervised forward selection
☑ Pairwise decorrelation ($r < 0.95$)
□ No UFS [edit] [x]

+add a custom template

Model validation

[all] [none]
☑ 5-fold cross-validation
□ 5-fold cross-validation (stratified - classification only)
□ Bagging with 64 models
□ Bagging with 64 models (stratified - classification only)

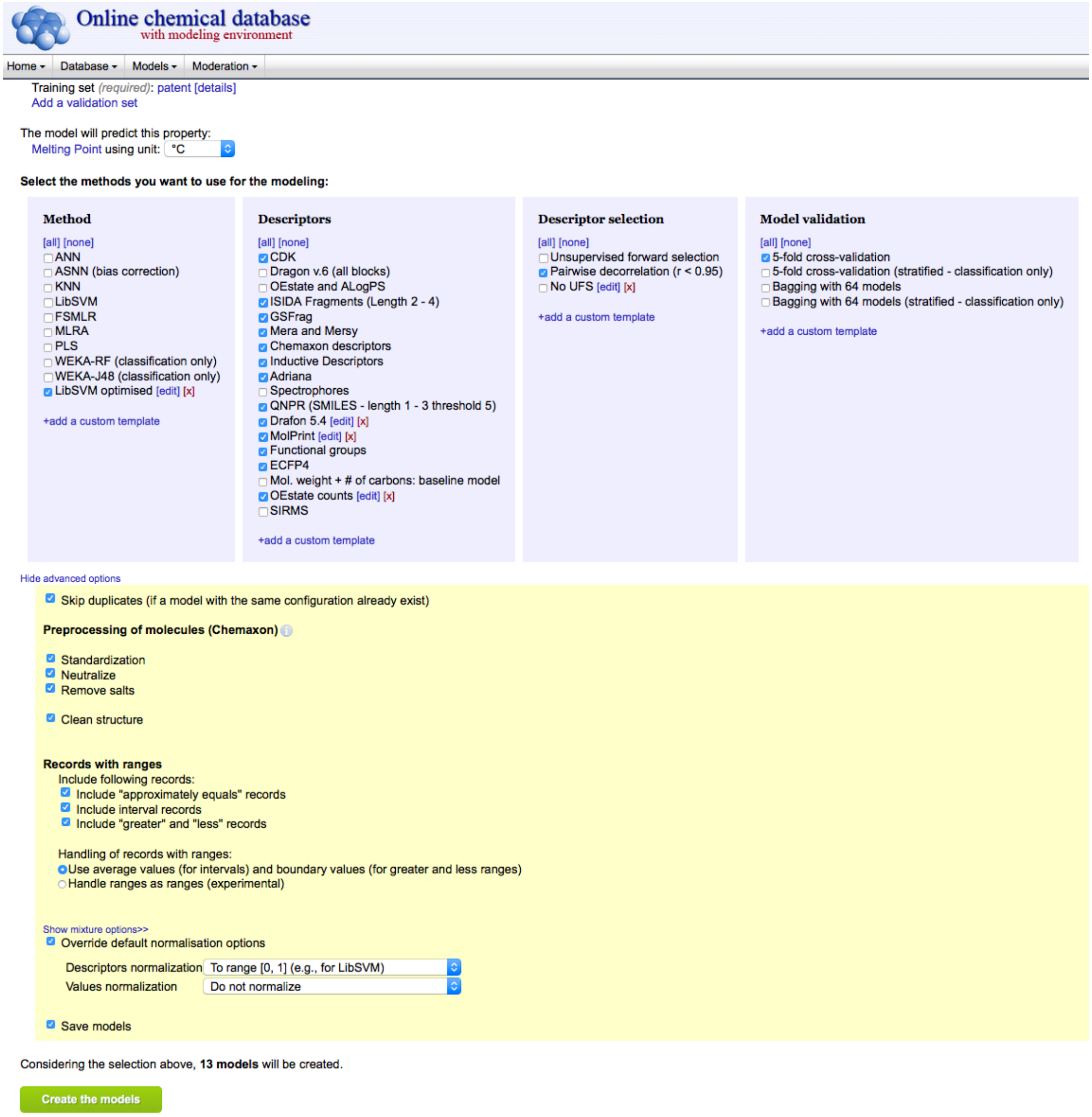
+add a custom template

Show advanced options>>

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

2.12 The optimal parameters of the model are used to create a new method template.

The “Comprehensive Modeling” mode was opened and a template was created using the “add a custom template on “Models/Create multiple models” menu. The previously selected model with optimized parameters was used as the template. An experienced user can also manually edit the XML file in the browser. In this example template with optimized LibSVM parameters was created. In a similar way new templates for descriptors, unsupervised descriptor selection and model validation can be created. They could be combined to create multiple models.



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

The model will predict this property:
Melting Point using unit: °C

Select the methods you want to use for the modeling:

Method
[all] [none]
- □ ANN
- □ ASNN (bias correction)
- □ KNN
- □ LibSVM
- □ FSMLR
- □ MLRA
- □ PLS
- □ WEKA-RF (classification only)
- □ WEKA-J48 (classification only)
- ☑ LibSVM optimised [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
- ☑ QNPR (SMILES - length 1 - 3 threshold 5)
- ☑ Drafon 5.4 [edit] [x]
- ☑ MolPrint [edit] [x]
- ☑ Functional groups
- ☑ ECFP4
- □ Mol. weight + # of carbons: baseline model
- ☑ OEstate counts [edit] [x]
- □ SIRMS

+add a custom template

Descriptor selection
[all] [none]
- □ Unsupervised forward selection
- ☑ Pairwise decorrelation (r < 0.95)
- □ No UFS [edit] [x]

+add a custom template

Model validation
[all] [none]
- ☑ 5-fold cross-validation
- □ 5-fold cross-validation (stratified - classification only)
- □ Bagging with 64 models
- □ Bagging with 64 models (stratified - classification only)

+add a custom template

Hide advanced options

☑ Skip duplicates (if a model with the same configuration already exist)

Preprocessing of molecules (Chemaxon)

- ☑ Standardization
- ☑ Neutralize
- ☑ Remove salts

☑ Clean structure

Records with ranges
Include following records:
- ☑ Include "approximately equals" records
- ☑ Include interval records
- ☑ Include "greater" and "less" records

Handling of records with ranges:
- ● Use average values (for intervals) and boundary values (for greater and less ranges)
- ○ Handle ranges as ranges (experimental)

Show mixture options>>
☑ Override default normalisation options

Descriptors normalization: To range [0, 1] (e.g., for LibSVM)
Values normalization: Do not normalize

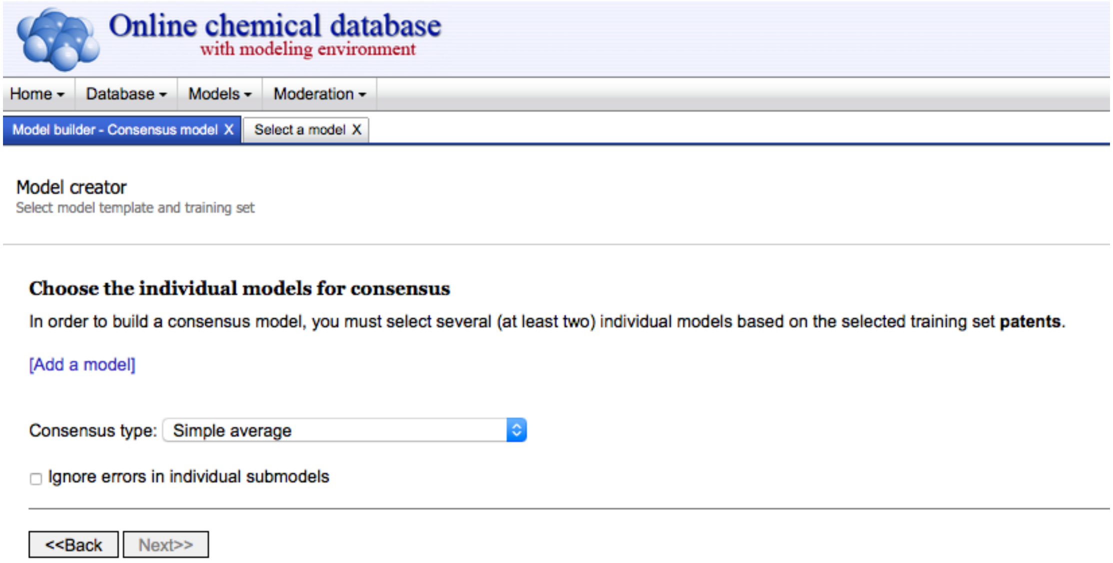
☑ Save models

Considering the selection above, **13 models** will be created.

3. The development of multiple models is started using model template with optimized LibSVM parameters.

For this article we used default settings for descriptors, descriptor selection and validation protocols as shown in the Figure above.

Models calculation was monitored in the Pending Task menu (see 2.8 and 2.9). Once models were finished, they were stored (each model individually, see step 2.11). After that they were used to develop a Consensus model using the "Consensus model" template and the same steps (2.1 and 2.2) used for the LibSVM model.



Model creator

Select model template and training set

Choose the individual models for consensus

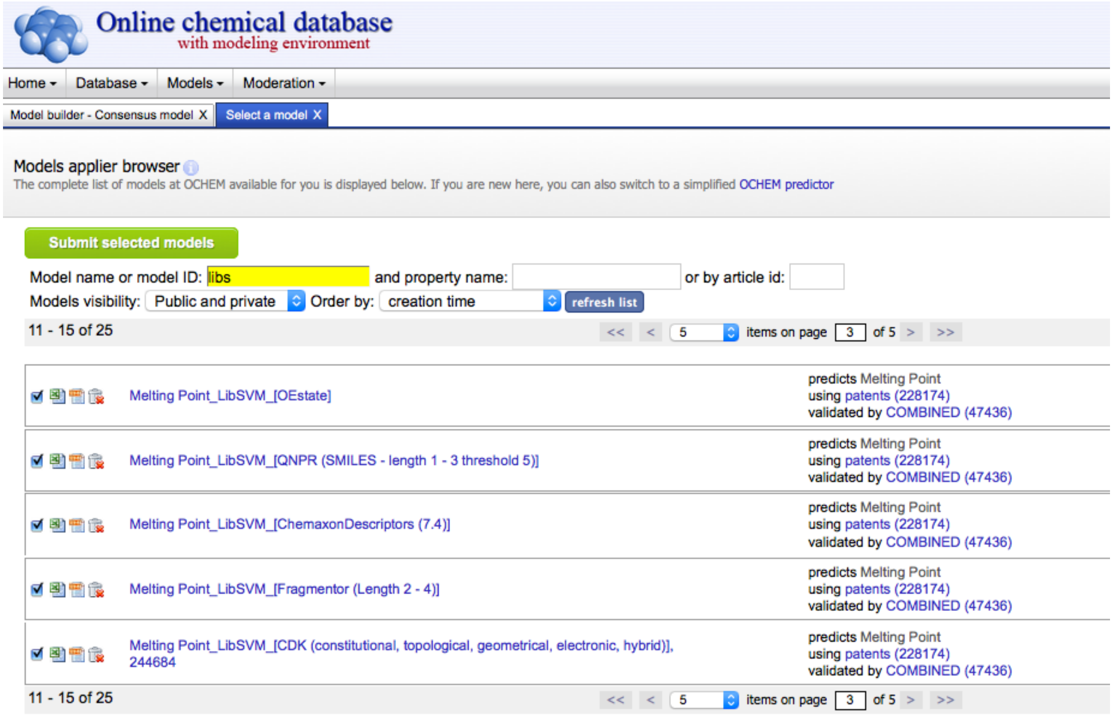
In order to build a consensus model, you must select several (at least two) individual models based on the selected training set **patents**.

[Add a model]

Consensus type: Simple average

□ Ignore errors in individual submodels

3.1 Consensus model developed using the "Simple average" option.



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

Model name or model ID: libs and property name: or by article id:

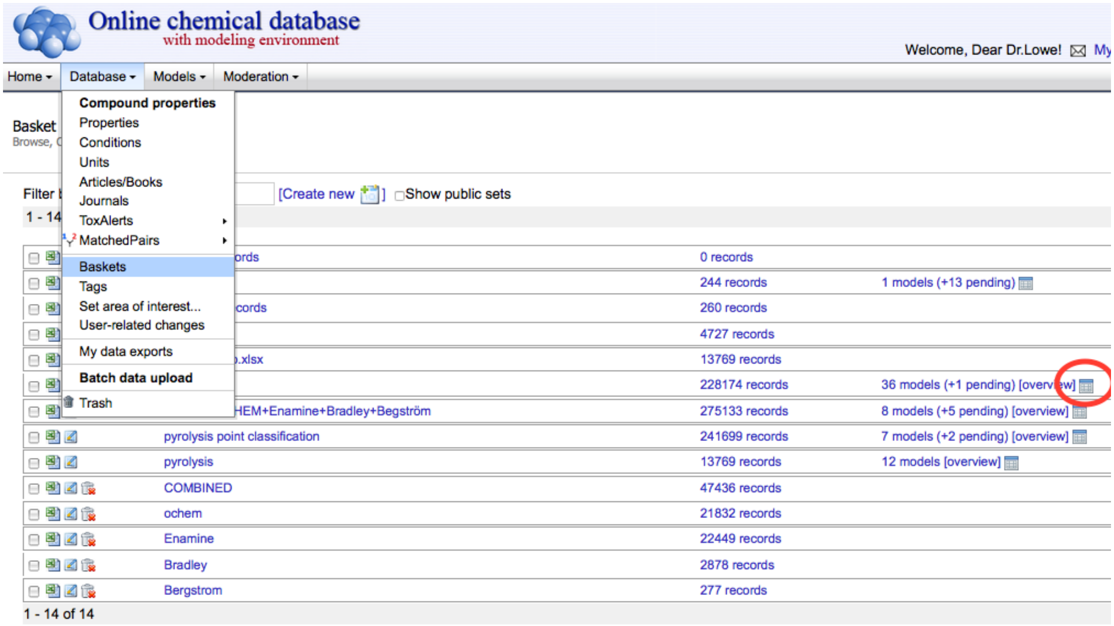
Models visibility: Public and private Order by: creation time

11 - 15 of 25 — 5 items on page 3 of 5

☑	Melting Point_LibSVM_[OEstate]	predicts Melting Point using patents (228174) validated by COMBINED (47436)
☑	Melting Point_LibSVM_[QNPR (SMILES - length 1 - 3 threshold 5)]	predicts Melting Point using patents (228174) validated by COMBINED (47436)
☑	Melting Point_LibSVM_[ChemaxonDescriptors (7.4)]	predicts Melting Point using patents (228174) validated by COMBINED (47436)
☑	Melting Point_LibSVM_[Fragmentor (Length 2 - 4)]	predicts Melting Point using patents (228174) validated by COMBINED (47436)
☑	Melting Point_LibSVM_[CDK (constitutional, topological, geometrical, electronic, hybrid)], 244684	predicts Melting Point using patents (228174) validated by COMBINED (47436)

11 - 15 of 25 — 5 items on page 3 of 5

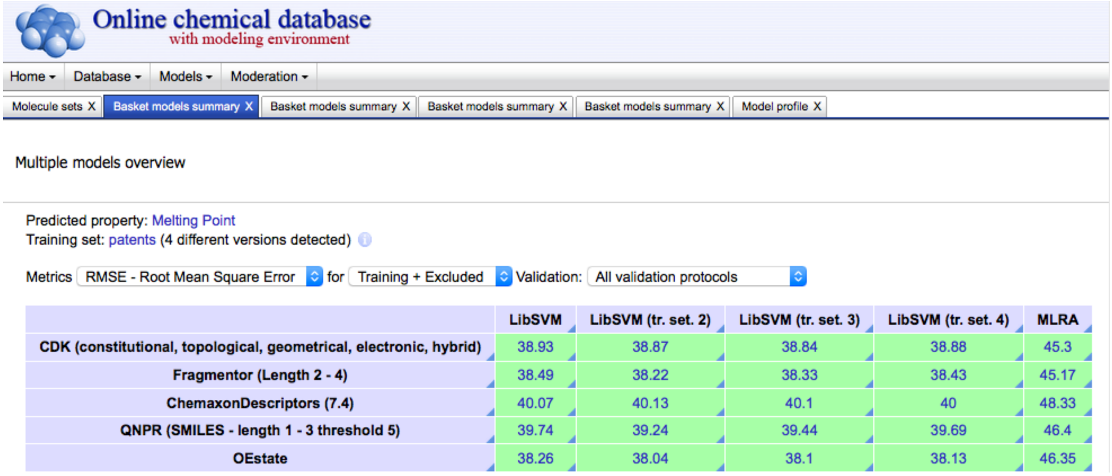
3.2 Selection of saved models for averaging in "Consensus model".



ords	0 records	
	244 records	1 models (+13 pending)
cords	260 records	
	4727 records	
).xlsx	13769 records	
	228174 records	36 models (+1 pending) [overview]
HEM+Enamine+Bradley+Begström	275133 records	8 models (+5 pending) [overview]
pyrolysis point classification	241699 records	7 models (+2 pending) [overview]
pyrolysis	13769 records	12 models [overview]
COMBINED	47436 records	
ochem	21832 records	
Enamine	22449 records	
Bradley	2878 records	
Bergstrom	277 records	

1 - 14 of 14

Step 4. The calculated models are accessed in the table – like Comprehensive Modeling (CM) view from “Baskets” web page.



Multiple models overview

Predicted property: Melting Point
Training set: patents (4 different versions detected)

Metrics: RMSE - Root Mean Square Error for Training + Excluded Validation: All validation protocols

	LibSVM	LibSVM (tr. set. 2)	LibSVM (tr. set. 3)	LibSVM (tr. set. 4)	MLRA
CDK (constitutional, topological, geometrical, electronic, hybrid)	38.93	38.87	38.84	38.88	45.3
Fragmentor (Length 2 - 4)	38.49	38.22	38.33	38.43	45.17
ChemaxonDescriptors (7.4)	40.07	40.13	40.1	40	48.33
QNPR (SMILES - length 1 - 3 threshold 5)	39.74	39.24	39.44	39.69	46.4
OEstate	38.26	38.04	38.1	38.13	46.35

Step 4.1 An overview of the developed models in the Comprehensive Modeling (CM) window.

Models developed with individual sets of descriptors are shown. The columns correspond to sets formed by the exclusion of outlying molecules with different p-values. The models developed using MLRA calculate with much lower accuracy.