C CHEM BENCH

ACCELERATING CHEMICAL GENOMICS RESEARCH BY CHEMINFORMATICS

C-ChemBench

User Guide

TABLE OF CONTENTS

INTRODUCTION

About this Document

User Guide document is intended for users of C-ChemBench web server. This document describes how to use the web service and contains a brief description of methods implemented in it.

About C-ChemBench web service

The C-ChemBench provides: a web-based user interface for Quantitative Structure-Activity Relationship (QSAR) analysis; access to data mining; library design; building the models; calculation of MolconnZ and Atom Pair descriptors; making predictions on sets of compounds and activity of the molecules. The C-ChemBench web server is available online: http://ceccr.ibiblio.org/c-chembench.

1 GETTING STARTED

1.1 Logging in to C-ChemBench web service

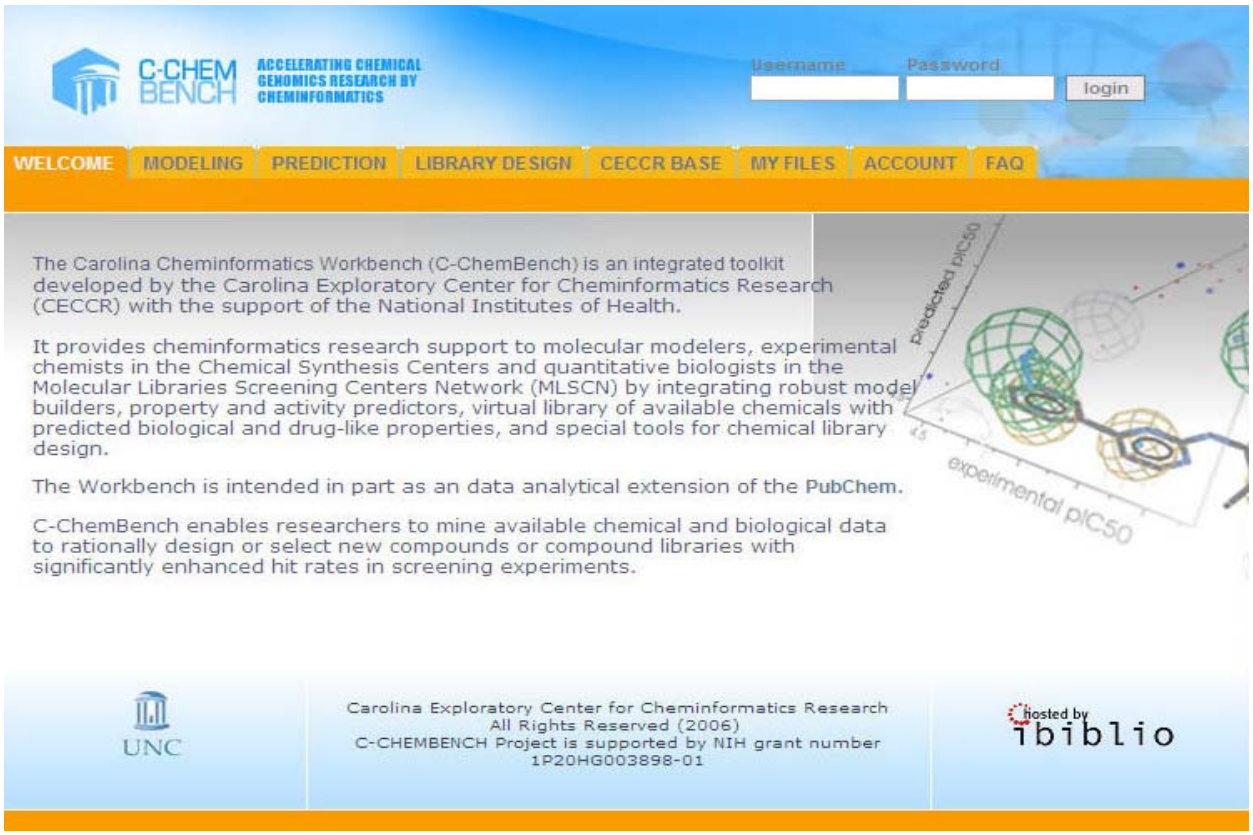
To start the C-ChemBench, do the following:

1. Start Web browser Firefox or Internet Explorer.

Note *C-ChemBench supports Firefox and Internet Explorer.*

2. In the Address bar, enter the C-ChemBench URL address: http://ceccr.ibiblio.org/c-chembench. The main window of the C-ChemBench and the ***Welcome*** tab opens (see Figure 1).
3. To login to the C-ChemBench, enter your **Username** and **Password**, and click ***login***.

Figure 1 C-ChemBench main window



Note *Only if you are registered in C-ChemBench you have the rights to log on to the system.*

1.2 Registering to C-ChemBench web service

1. To register to the system, expand ***Account*** in the main C-ChemBench menu (see Figure 2).

2. In the appropriate fields type in: **First name, Last name, Name of organization, Position in organization, Address, City, State/Province, Zip Code, Phone Number** (Example: (001)919-966-3459)**, Email** (Please, use your organization email account)**, User Name** (Note, that the user name should be at least 4 characters).

   Select from the drop down list the following information: **Type of organization, Country, Work Bench.**

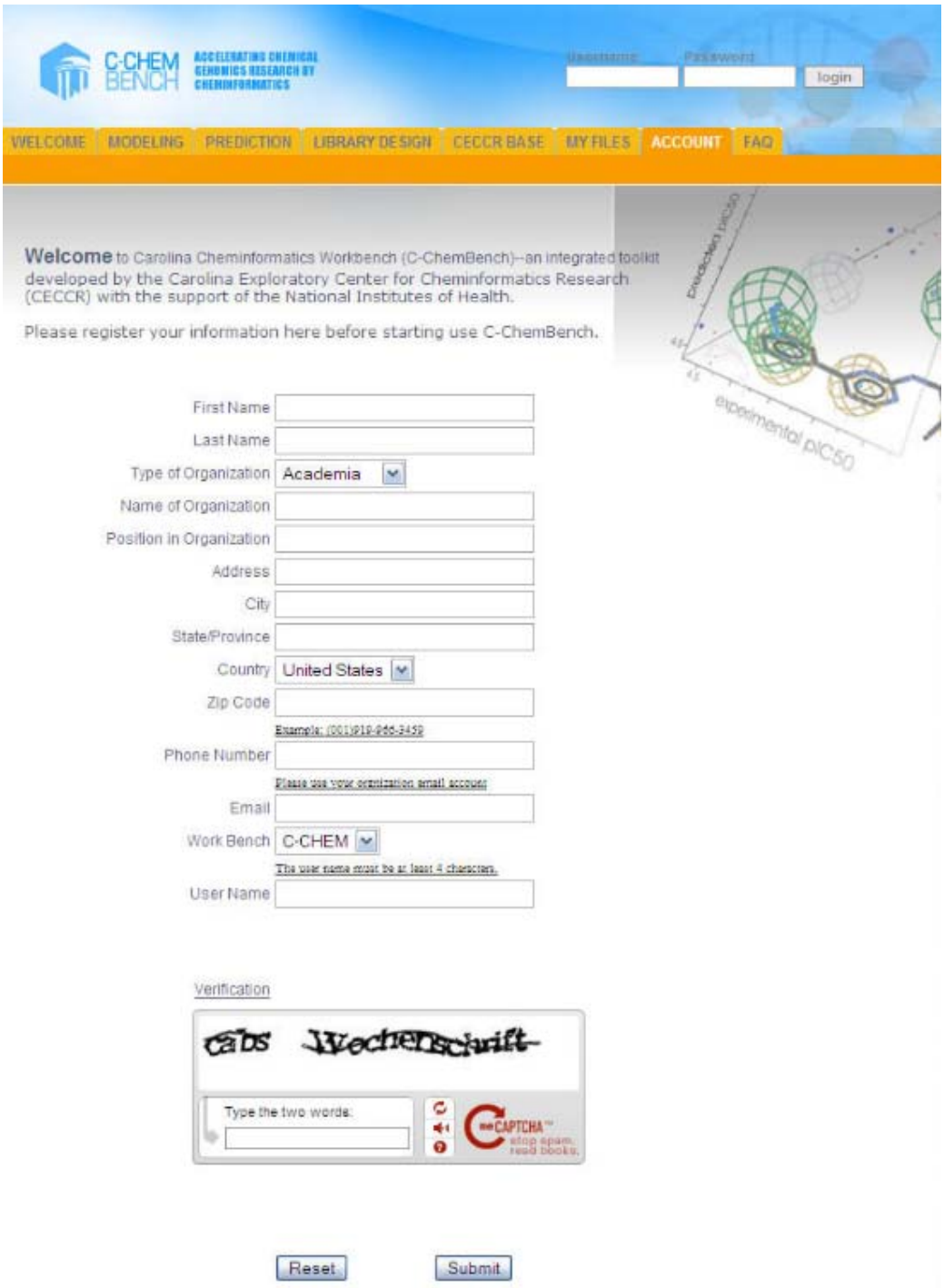
3. Type in the two words you see in the box (in order and separated by a space) in the **Verification** field. If you are not sure what the words are, click the ***Get the new challenge*** reload button  . Visually impaired users can click the ***Get the audio challenge*** audio button  to hear a set of digits that can be entered instead of the visual challenge. If you still have a problem, click ***Help*** button  .)

4. Press Submit to save the information or Reset to discard it.

Note

*Your password will be sent to you by email. Once a user account has been set up for you, you will be able to login (by providing your **Username** and **Password** and by pressing **login** button in the right upper part of the main page), logout (by pressing **logout** button in the right upper part of the main page), change your password (for more details see section 7 Account), etc. by completing the appropriate form.*

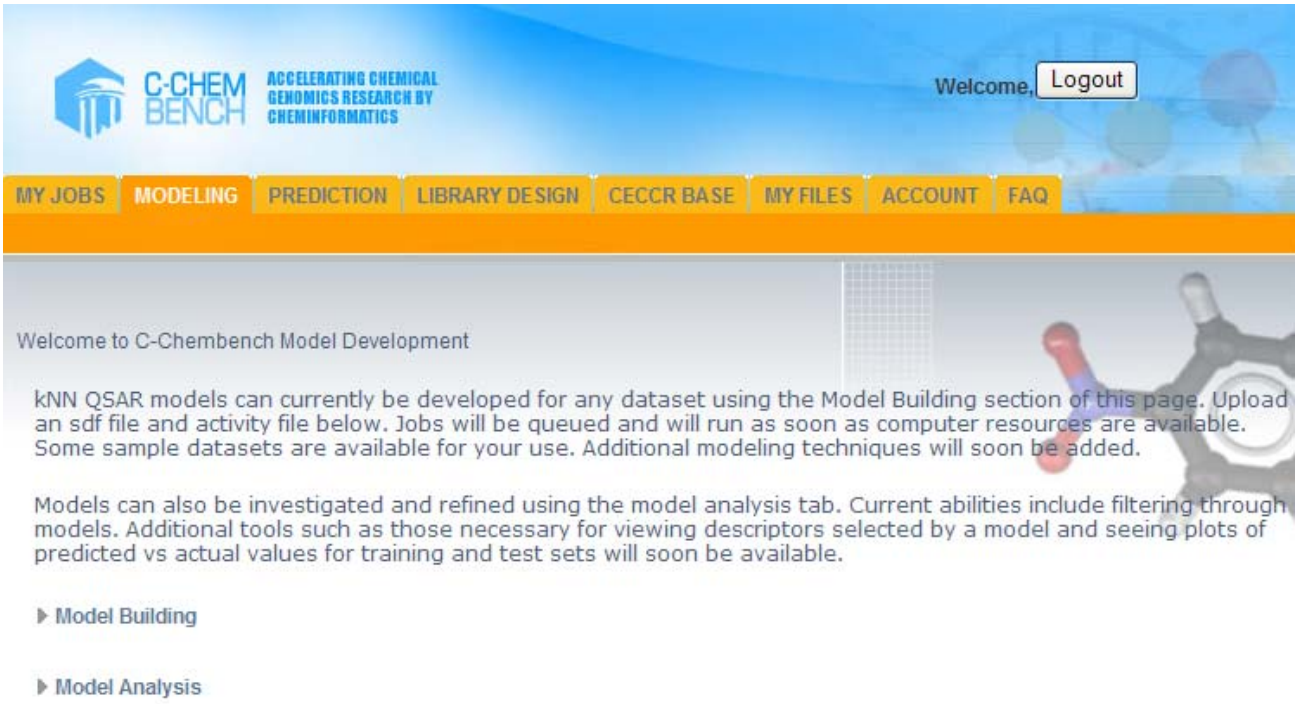
Figure 2 Account (for unregistered user) window



2 MODELING

After you are logged in to the system, the ***Modeling*** tab opens (see Figure 3). 'Modeling' gives you the possibility to build your own models and to analyze them. ***Welcome*** tab is changed onto the ***My Jobs*** tab.

Figure 3 Modeling (starting) window



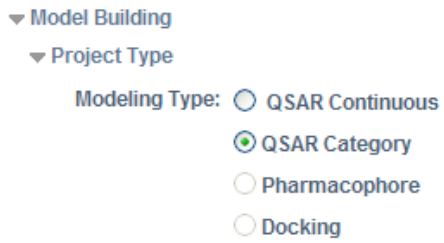
2.1 Model Building

1. To build a model, click ***Model Building*** (see Figure 4). It enables you to build the models step by step.

   You should specify the following properties: **Project Type**; **Data Set**; **Tool**; **Descriptors**; **Data Set Division**; **Job Name**.

2. Click ***Project Type*** and specify the ***Modeling Type*** by selecting the radio button:
   - QSAR Continuous – for predicting activities values which are continuous;
   - QSAR Category - for predicting activities values which are discrete.

Figure 4 Modeling Type dialog



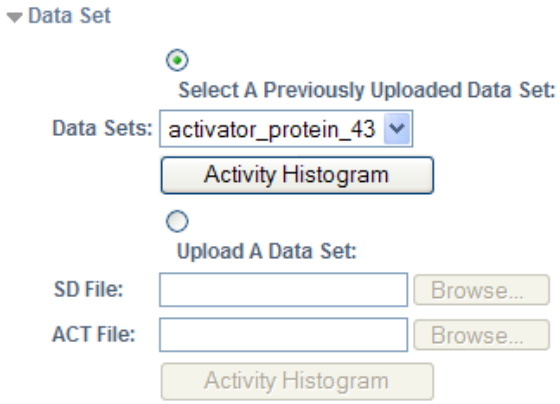
3. Click ***Data Set*** (see Figure 5) and choose the appropriate radio button:
   - Select A Previously Uploaded Data Set (You have the possibility to select public as well as your own files from the drop-down list)
   - Upload a Data Set (You have the possibility to upload your own files. User's dataset files should include a molecule file in SDF format and an activity file)

Note

SD File *is Structure Data Format file. For more details, see* http://www.epa.gov/ncct/dsstox/MoreonSDF.html.

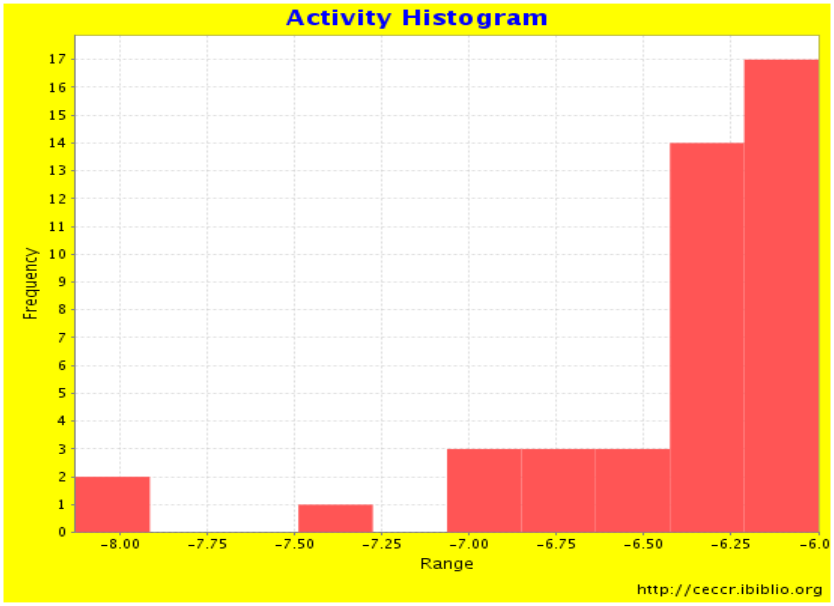
ACT File *(Activity File) has the following structure: One line per compound: compound_name activity_value (category); Categories must be non-negative consecutive whole numbers.*

Figure 5 Data Set dialog



To view the number of the molecules in each interval of activity, press ***Activity Histogram*** (see Figure 5) button**.** The 'Activity Histogram' window appears (see Figure 6). The horizontal axis shows the range of activities and its division into bins and the vertical axis shows the number of compounds with activities within bins.

Figure 6 Activity Histogram window

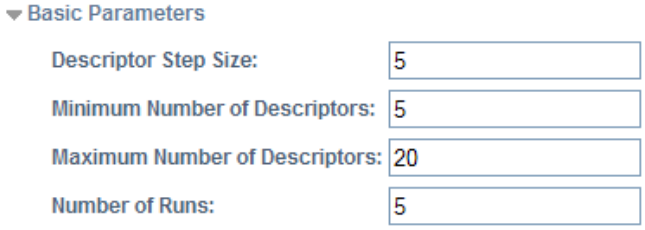


4. Click ***Tool*** (see Figure 7) and select ***KNN*** (k nearest neighbors QSAR method). You have the possibility to edit basic and advanced parameters. To specify them, click ***Basic Parameters*** (see Figure 8) and then ***Advanced Parameters*** (see Figure 9 and Figure 10).

Figure 7 Tool dialog

Basic parameters (see Figure 8) include: Descriptor step size, Minimum Number of Descriptors selected, Maximum Number of Descriptors selected, and the Number of Runs for each set of parameters.

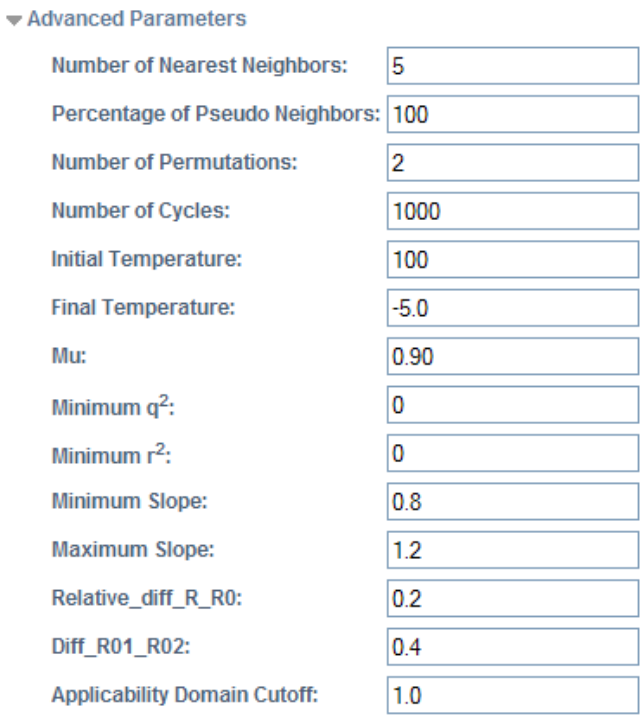
Figure 8 Basic Parameters dialog



If in the step 2 (see Figure 4) you have selected *QSAR Continuous* modeling type, the **Advanced parameters** will be the following (see Figure 9): Number of Nearest Neighbors, Percentage of

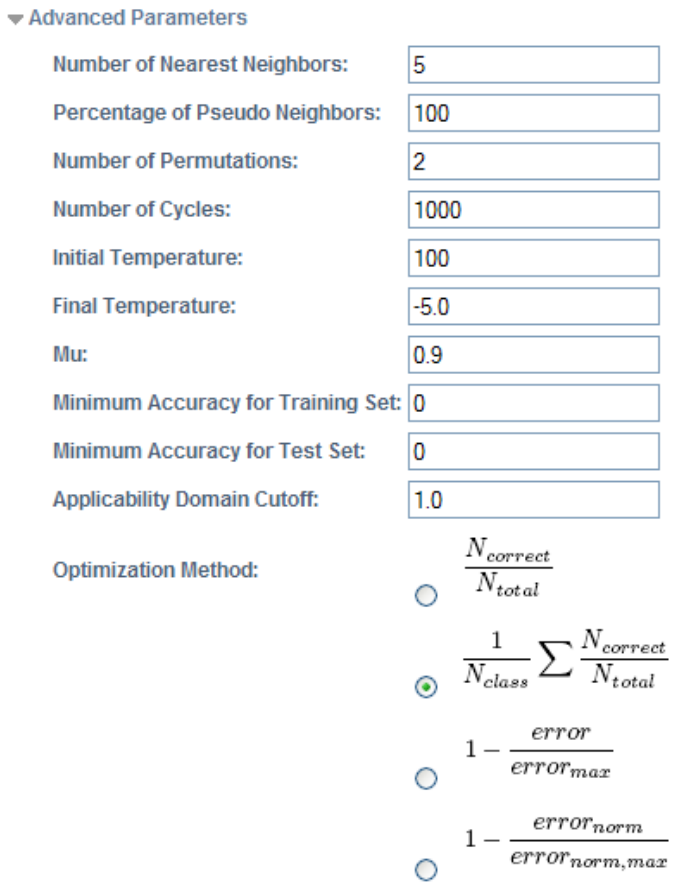
Pseudo Neighbors, Number of Permutations of selected descriptors, Number of Cycles, Initial Temperature, Final Temperature, Temperature Coefficient Mu, Minimum acceptable q^2, Minimum acceptable r^2, Minimum and Maximum slopes for regressions through the origin for predicted vs. observed and observed vs. predicted activities, coefficients of determination for regressions through the origin, Applicability Domain Cutoff model parameter.

Figure 9 Advanced Parameters (for Continuous kNN) dialog



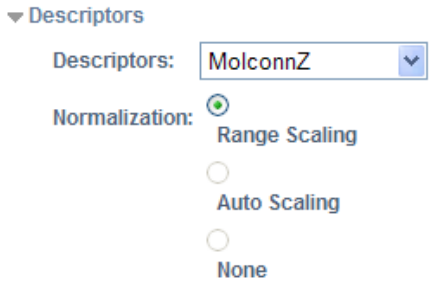
If in the step 2 (see Figure 4) you have selected *QSAR Category* modeling type, the **Advanced parameters** will be the following (see Figure 10): Number of Nearest Neighbors, Percentage of Pseudo Neighbors, Number of Permutations of selected descriptors, Number of Cycles, Initial Temperature, Final Temperature, Temperature Coefficient Mu, Minimum Accuracy of classification for Training Set, Minimum Accuracy of classification for Test Set, Applicability Domain Cutoff model parameter. Also, you have to specify the Optimization Method, by selecting the appropriate radio button.

Figure 10 Advanced Parameters (for Category kNN) dialog



5. Click ***Descriptors*** (see Figure 11) and select from the drop-down list 'MolconnZ' descriptors. You have the possibility to normalize descriptors, by selecting 'Range Scaling' radio button.

Figure 11 Descriptors dialog

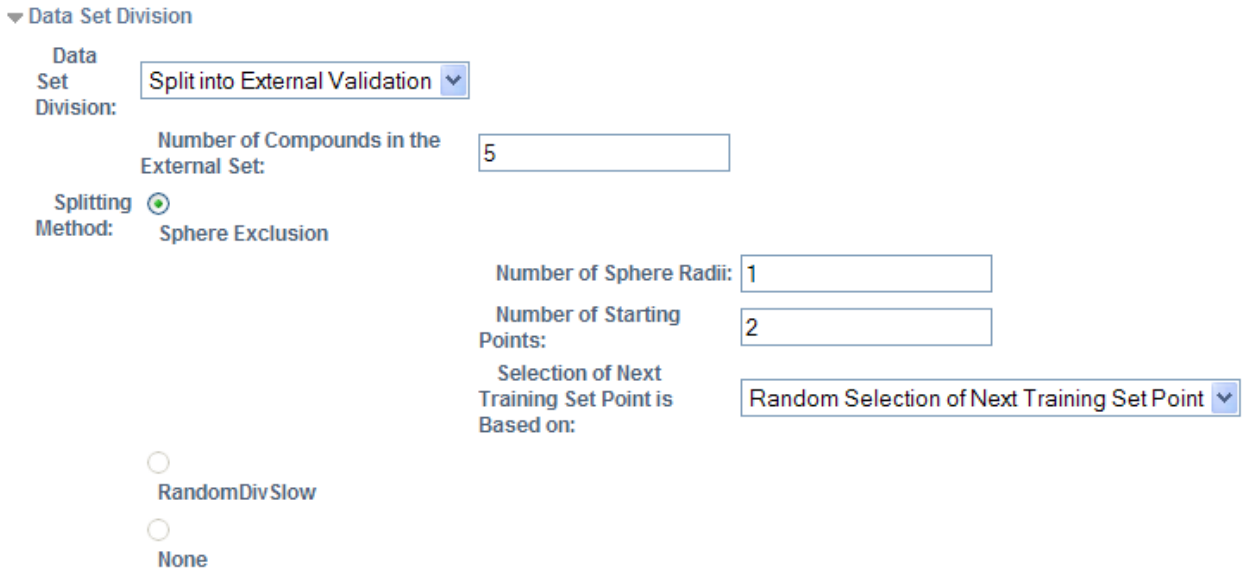


6. Click ***Data Set Division*** (see Figure 12) and select 'Split into External Validation' from the drop-down list of **Data Set Division**. The External Validation Data Set will be selected from the entire dataset randomly.

Enter the **Numbers of Compounds in the External Set** and choose the **Splitting Method**, by selecting 'Sphere Exclusion' radio button. Provide **Number of Sphere Radii** and **Number of Starting Points** included in the training set. Select from the drop-down list the **Selection of Next Training Set Point** that is based on the following:

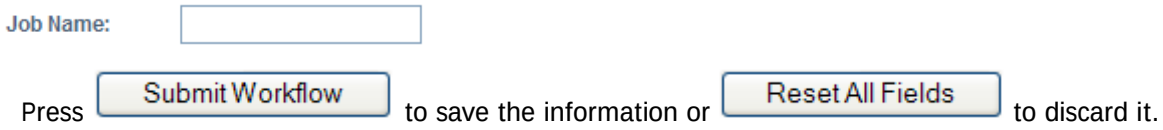
- Random Selection of Next Training Set Point (In this case, the next point of the training set is selected randomly.)
- Minimum Sphere Center Distances (In this case, the next point of the training set is selected as the point closest to the center of one of the previous spheres.)
- Maximum Sphere Center Distances (In this case, the next point of the training set is selected as the point closest to one of the sphere centers among farthest from the centers of all of the previous spheres.)

Figure 12 Data Set Division dialog



7. Enter the ***Job Name*** (see Figure 13).

Figure 13 Job Name dialog



2.2 My Jobs

1. After pressing ***Submit Workflow*** button (see Figure 13), ***My Jobs*** tab opens (see Figure 14). The list of active tasks and tasks in the queue appears in the table.

   My Active Tasks contain the following information:

- Name – the name of the job;
- Submitted – date and time of submitting the data;
- Status – the status of the job ('started' or 'finished');
- Started – date and time of starting to process data;
- Finished – date and time of finishing to process data.

Tasks in the Queue contain the following information:

- Name – the name of the job;
- Owner – your name (if it is your job) or 'other user';
- Submitted – date and time of submitting the data;
- Status – the status of the job ('ready' or 'permission required').

Note *'Permission Required' Status appears if the files you submitted have more than 200 compounds or 10 000 models. In this case, the administrator will be notified and may give you permission to perform the processing of data.*

2. Select the finished job, by clicking on its appropriate name.

To check the status of your job, press ***Refresh Status of Your Tasks*** button.

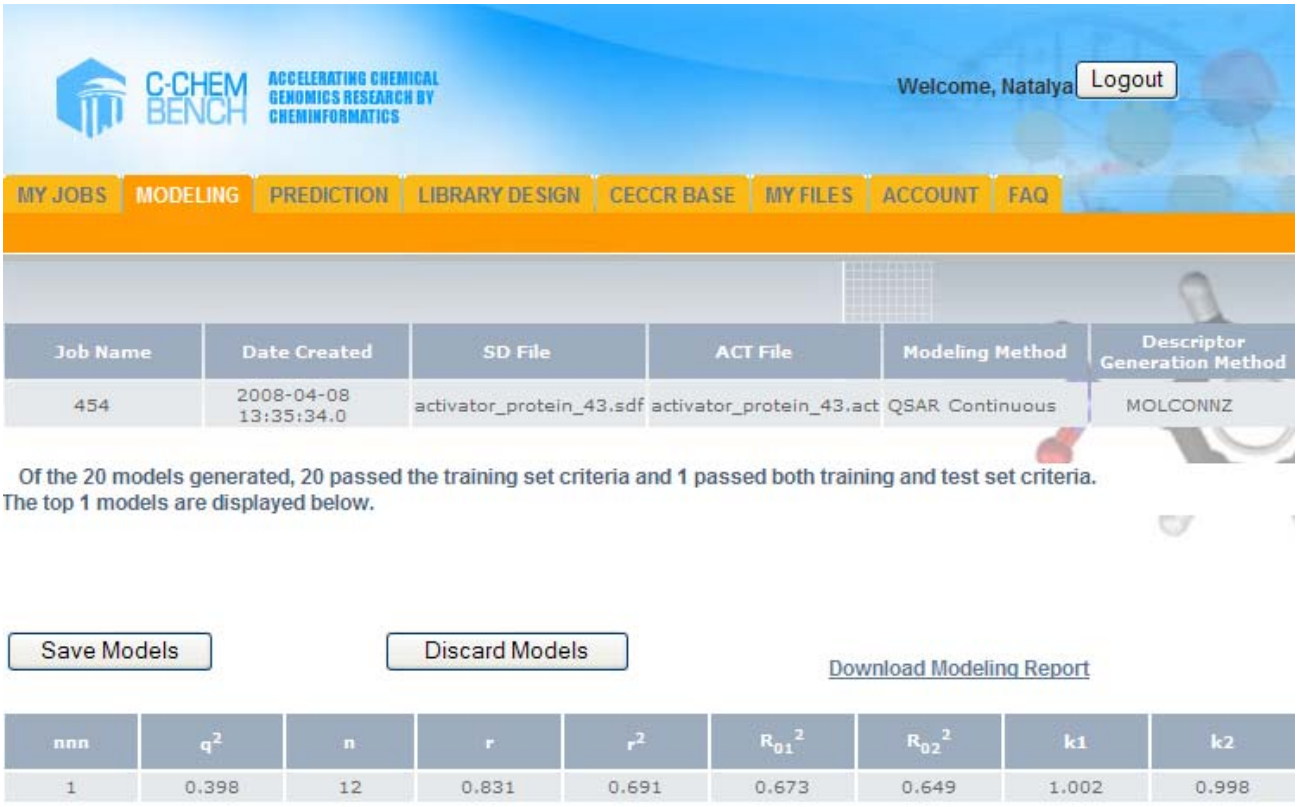
Figure 14 My Jobs window

2.3 Model Analysis

1. After clicking on the name of the finished job at 'My Jobs' tab (see Figure 14), the ***Modeling*** tab opens (see Figure 15). On the 'Modeling' page you can: find the information about the number of generated models and which passed training and test set criteria; view the list of properties with the following attributes:
    - Job Name – the name of the job given by you;
    - Date Created – date and time the model was built;
    - SD File – structure file used for model generation;
    - ACT File – activity file used for model generation;
    - Modeling Method – either QSAR Continuous or QSAR Category;

- Descriptor Generation Method – (MolconnZ).

Figure 15 Modeling (after generation) window



2. To download the zip file version of the modeling report, click ***Download Modeling Report*** button.
3. To view the results of randomization, click ***Y Randomization Results*** (see Figure 16).

   Y Randomization is a statistical QSAR model validation technique where the performance of the model built with original data is compared to that of models built for multiple artificial datasets with randomly shuffled activities. Ideally, there will be no models with high values (> 0,6) of both q^2 (training set) and R^2 (test set) found.

   The top few models are displayed with the following properties:

   - nnn – number of nearest neighbors used for making the activity prediction with this model;
   - Training Accuracy = $\frac{\#\ of\ correct\ predictions}{\#\ of\ total\ compounds}$ for training set
   - Normalized Training Accuracy =
     $= \frac{1}{\#\ of\ classes}\left(\frac{\#\ of\ correct\ class1}{\#\ of\ total\ compound\ class1} + \frac{\#\ of\ correct\ class2}{\#\ of\ total\ compound\ class2}\right)$

NTA = $\frac{1}{\#of\ classes}\sum_{N=1:\#\ of\ classes}\left(\frac{\#of\ correct\ predictions\ for\ class\ N}{\#of\ total\ compounds\ for\ class\ N}\right)$ for training set

- Test Accuracy = $\frac{\#of\ correct\ predictions}{\#of\ total\ compounds}$ for test set
- Normalized Test Accuracy =

$$= \frac{1}{\#of\ classes}\left(\frac{\#of\ correct\ class1}{\#of\ total\ compound\ class1} + \frac{\#of\ correct\ class2}{\#of\ total\ compound\ class2}\right)$$

NTA = $\frac{1}{\#of\ classes}\sum_{N=1:\#\ of\ classes}\left(\frac{\#of\ correct\ predictions\ for\ class\ N}{\#of\ total\ compounds\ for\ class\ N}\right)$ for test set

Figure 16 Y Randomization Results dialog

Y Randomization Results

Y-randomization is a statistical QSAR model validation technique where the performance of the model built with original data is compared to that of models built for multiple artificial datasets with randomly shuffled activities. Ideally, there will be no models with high values(> 0.6) of both q^2 (training set) and R^2 (test set) found.
For your data, 20 models for randomized datasets were built and 0 models were found to have high prediction accuracy.

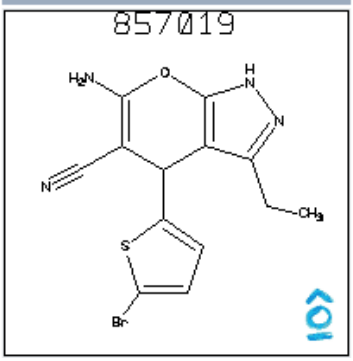
nnn	q^2	n	r	r^2	R_{01}^2	R_{02}^2	k1	k2

You have the possibility to view the **External Validation Results** (see Figure 17) with the following properties:

- Comp_ID – ID of the compound;
- Structure – the two-dimensional structure of compound;
- Observed Value – the value of the activity for the compound as submitted in the ACT file;
- Predicted Value – the value of the activity for the compound as predicted by the consensus of models;
- Residual – difference between the Observed Value and Predicted Value;
- # of Models – number of models for which this compound was within the applicability domain.

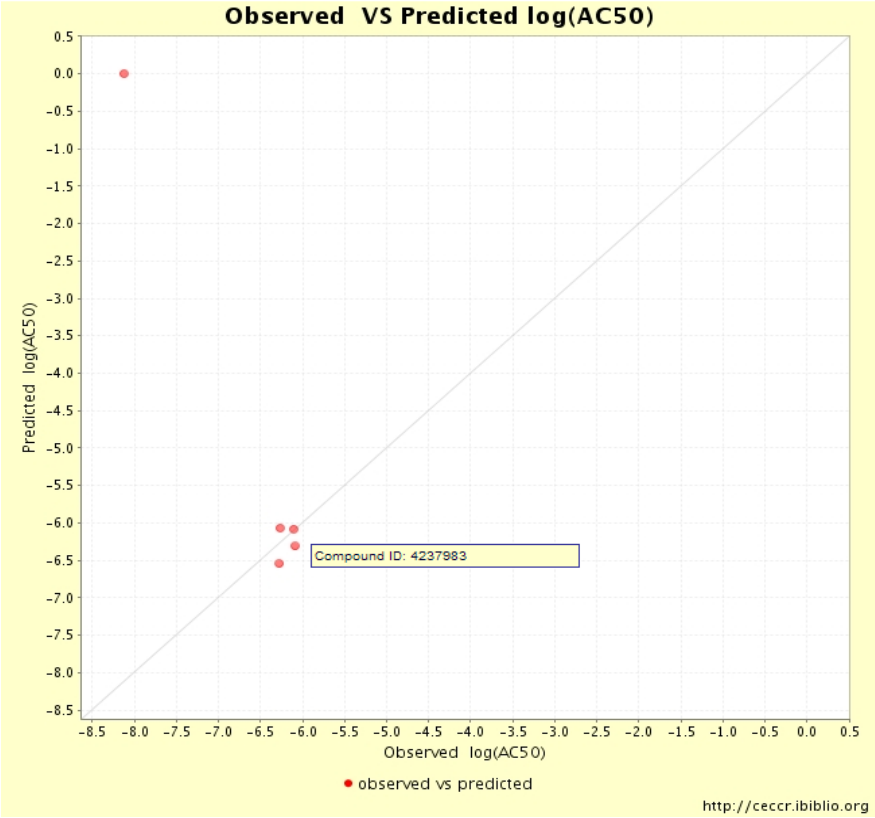
Figure 17 External Validation Results dialog

External Validation Results: Chart View

Comp_ID	Structure	Observed Value	Predicted Value	Residual	# of Models
857019		-6.09	-6.31	0.22	1

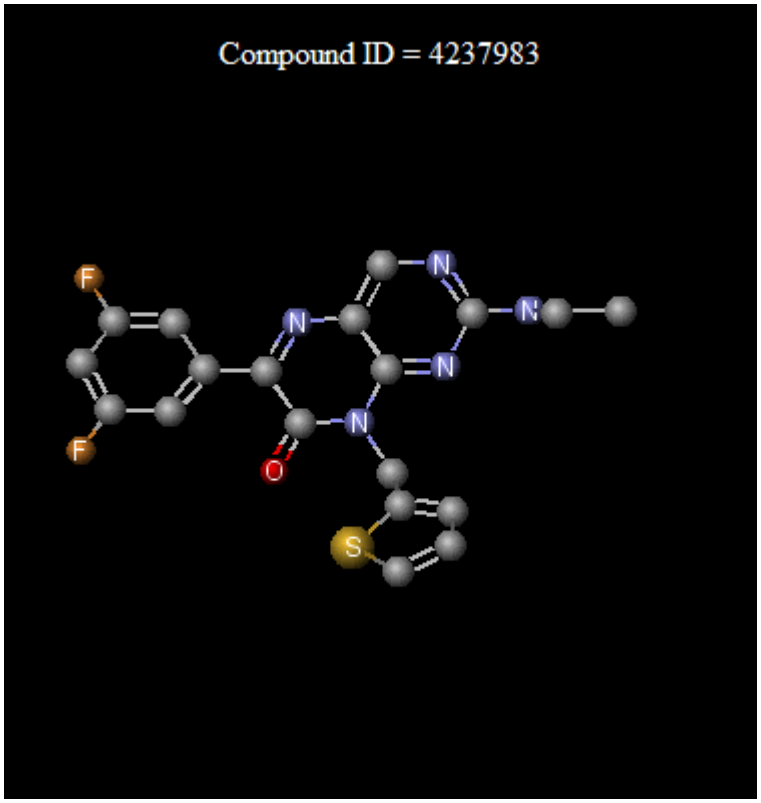
To view the chart of the predicted versus actual activities for the external molecules, click ***Chart View***. The 'Observed vs Predicted log' window appears (see Figure 18). The horizontal axis shows the observed activity as uploaded in your activity file and the vertical axis shows the predicted activity by the models, which you generated. Each red circle represents a single molecule.

Figure 18 Observed vs Predicted log window



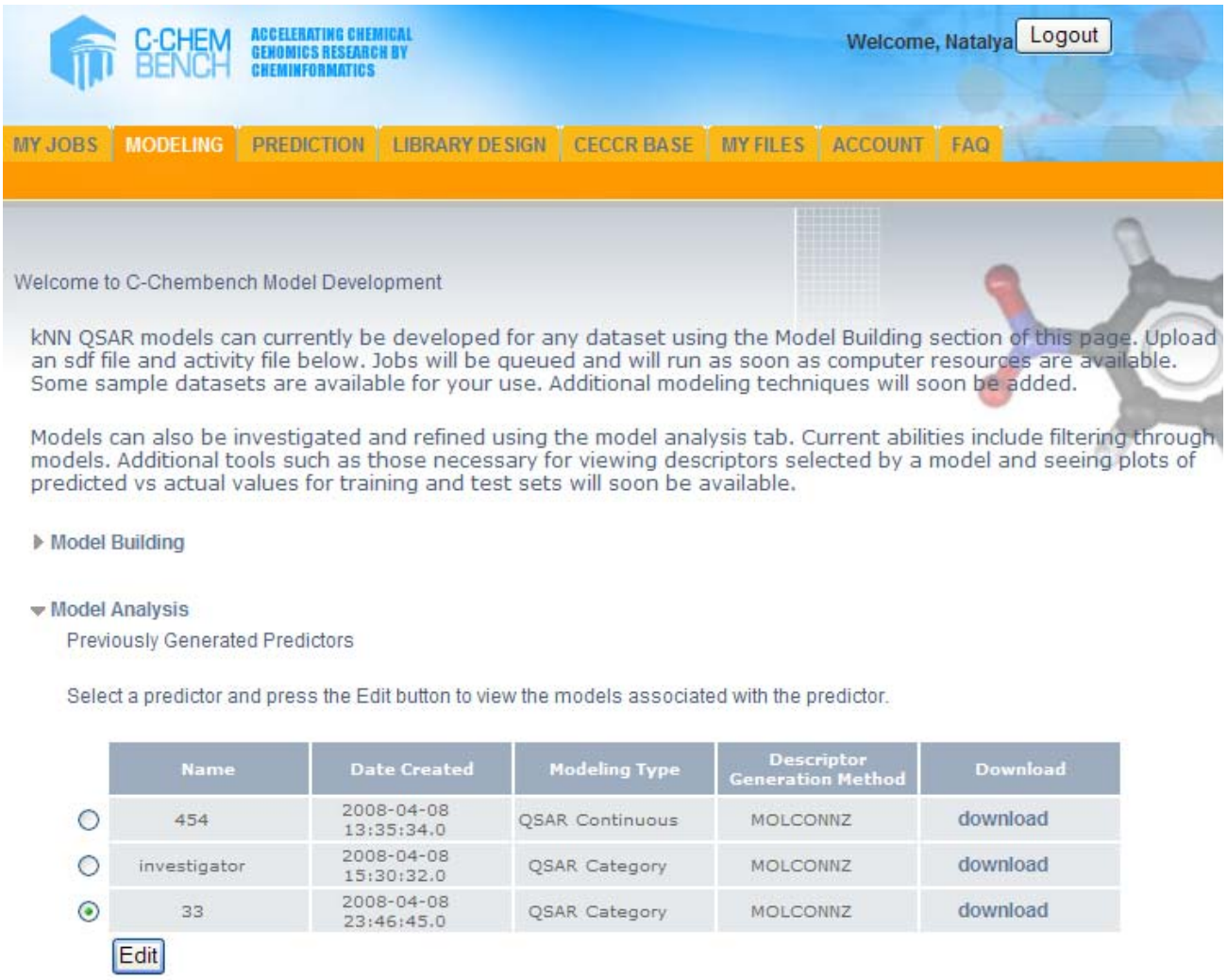
If you point the mouse cursor at the red circle, the compound ID appears. To view the 3-D model of the compound structure, click on the corresponding red circle. The 'Compound ID' window appears (see Figure 19).

Figure 19 Compound ID window



4. Press ***Save Models*** button to save the information or ***Discard Models*** button to discard it (see Figure 15).
5. After you saved the model, you have the possibility to view the list of all predictors, which were generated previously, by clicking ***Model Analysis*** (see Figure 20).
6. The system reflects the following predictor features: **Name**, **Date Created**, **Modeling Type**, **Descriptor Generation Method**, and **Download**.
7. To download the zip file version of the model, click ***download.*** Every necessary file for your model as well as all files generated for external validation will be downloaded.
8. Choose the necessary predictor, by selecting the appropriate radio button.
9. To view the models associated with the predictor and to be able to edit information, press ***Edit*** button.

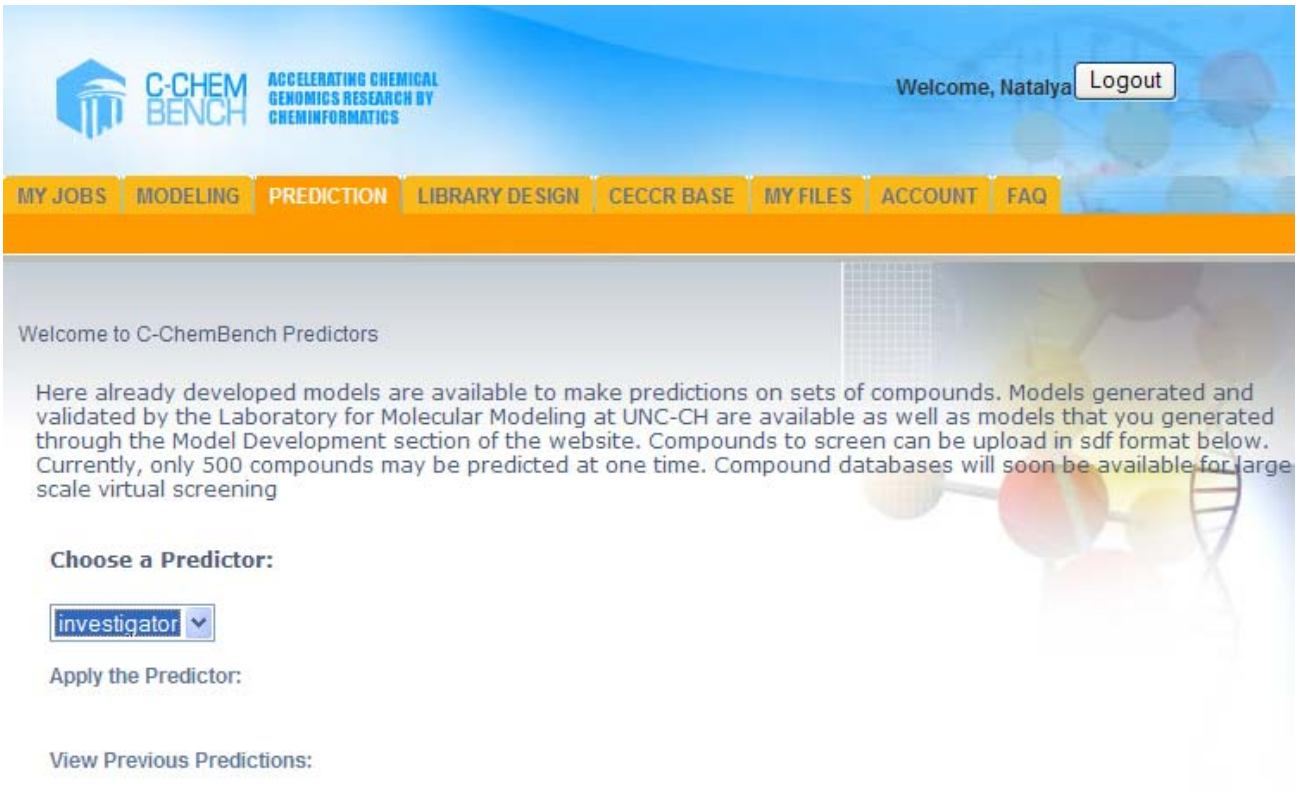
Figure 20 Modeling (Predictor list) window



3 PREDICTION

1. Expand ***Prediction*** in the main C-ChemBench menu (see Figure 21). The already developed models are available to make predictions on sets of compounds.
2. Select a predictor from the drop down list.

Figure 21 Prediction (starting) window



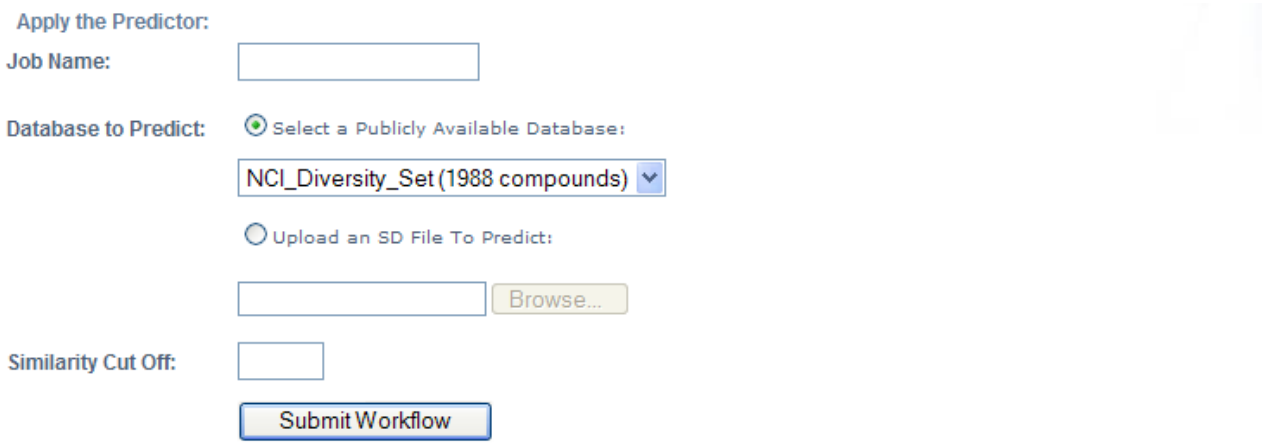
3.1 Predictor Applying

1. Click ***Apply the Predictor*** (see Figure 22).
2. Enter the ***Job Name.***
3. Select a database for prediction by choosing the appropriate radio button:
    - Select a Publicly Available Database (You have the possibility to select public as well as your own files from the drop-down list)
    - Upload an SD File to Predict (You have the possibility to upload your own files. User's dataset files should include a molecule file in SDF format and an activity file)

4. Enter the ***Similarity Cut Off*** value — it is the parameter which defines the model applicability domain. The higher the Z Cutoff is, the larger the applicability domain is and the more compounds from external datasets are predicted. Do not make it more than 3.

5. Press Submit Workflow button to save the information.

Figure 22 Apply the Predictor dialog



Note *After pressing **Submit Workflow** button, **My Jobs** tab opens. For more details, see section* 2.2My Jobs.

3.2 Prediction Analysis

1. After clicking on the name of the finished job at 'My Jobs' tab (see Figure 14), the ***Prediction*** tab opens (see Figure 23). On the 'Prediction' page you can view the list of properties with the following attributes:

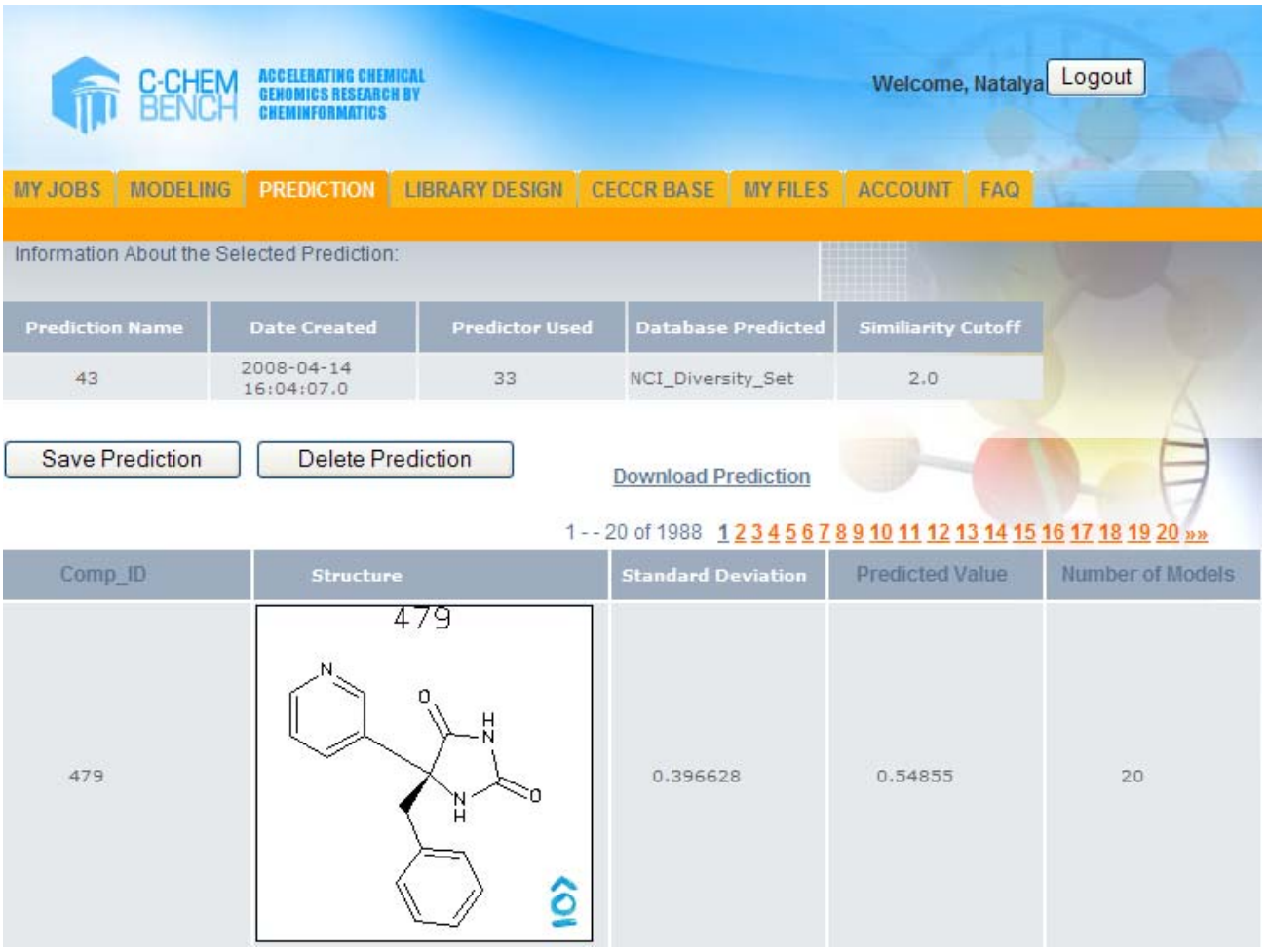
    - Prediction Name - the name of the prediction given by you;
    - Date Created - date and time the prediction was made;
    - Predictor used – the name of predictor used for the prediction;
    - Database Predicted – the name of the chemical database on which predictions were made;
    - Similarity Cutoff – the similarity cutoff used for this prediction.

    You have the possibility to view the results with the following properties:

    - Comp_ID – ID of compounds;
    - Structure – the two-dimensional structure of the predicted compound;
    - Standard Deviation – the standard deviation in the predicted value based on individual model predictions;

- Predicted Value – the value of the activity for the compound as predicted by the consensus of models;
- Number of Models – number of models for which this compound was within the applicability domain.

Figure 23 Prediction (after generation) window



2. To view the current prediction result table, click ***Download Prediction***.
3. Press ***Save Prediction*** button to save the information or ***Delete Prediction*** button to discard it (see Figure 23).
4. After you saved the prediction, you have the possibility to view the list of all predictions, which were generated previously, by clicking ***View Previous Predictions*** (see Figure 24).
5. The system reflects the following predictor features: **Name**, **Date Created**, **Predictor**, **Database Predicted**, and **Download**.

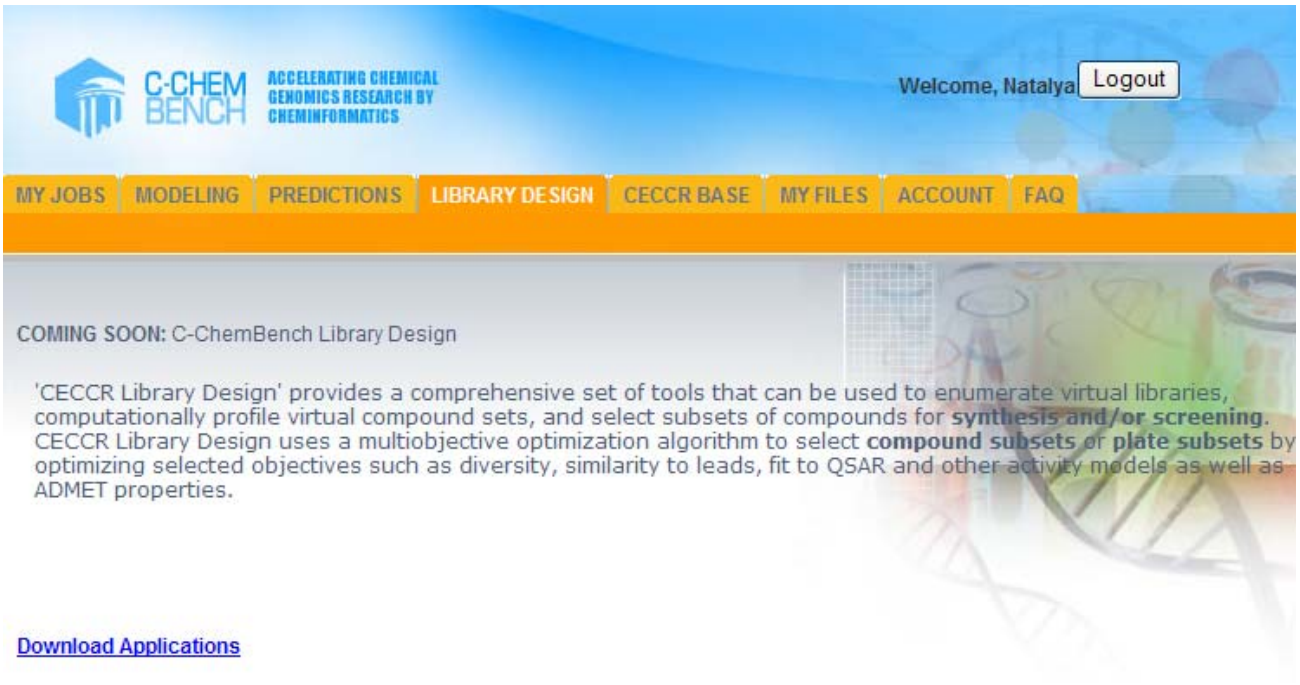
6. To download the prediction results, click ***download.*** All necessary files for your model as well as all files generated while making your predictions will be downloaded.
7. Choose the necessary prediction, by selecting the appropriate radio button.
8. To view the predictions associated with the predictor, press ***View*** button.

Figure 24 Prediction (including results) window

4 LIBRARY DESIGN

1. Expand ***Library Design*** in the main C-ChemBench menu (see Figure 25). 'Library Design' provides an access to tools which allow the generation of virtual chemical libraries.
2. Click ***Download Applications***.

Figure 25 Library Design window



3. The list of downloaded applications appears with the following features: **Application Name**, **Description**, **Download** (see Figure 26).
4. To download the zip file version of the applications used in model generation, click ***download.***

Figure 26 Library window

5 CECCR BASE

1. Expand ***CECCR Base*** in the main C-ChemBench menu (see). 'CECCR Base' provides an access to the data base, which contains chemical information and biological pathways. Base provides an access to a large chemical database containing virtual activities and providing links to biological information.

6 MY FILES

1. Expand ***My Files*** in the main C-ChemBench menu (see). 'My Files' allows management of all the files which were uploaded by you in the 'Data Set' section of 'Model Building' tab (see Figure 5).

7 ACCOUNT

1. Expand ***Account*** in the main C-ChemBench menu (see). The 'Account' tab opens, where you have the possibility either to register to the system (for more details see section 1.2Registering to C-ChemBench web service), or, in case if you are already registered user, to change your password.
2. To change your password, enter: your **Old Password** and **New Password**.
3. Enter your new password once more in the **Confirm Password** field.
4. Press ***Submit*** button to save the information or ***Reset*** button to discard it.

Figure 27 Account (for registered user) window