

**CEFIC LRI RfP EEM-9**  
**Building blocks for a future**  
**(Q)SAR Decision Support System:**  
**Databases, applicability domain and structure conversions**  
(Codename AMBIT)

Nina Nikolova-Jeliazkova  
IPP - Bulgarian Academy of Science, Sofia, Bulgaria

Joanna Jaworska  
P&G Brussels, Belgium

April 5, 2005

## 1.Introduction

### 1.1.Background

This report concerns EEM9- 1a Applicability domain objective and describes **AmbitDisclosure**, the QSAR applicability domain identification software developed. It is appropriate to all modeling approaches and is based on

- ◆ The projection of the training set in the model descriptor space.
- ◆ Structural similarity, assessed by fingerprints

#### 1.1.1.Implementation

All the applications within the AMBIT project are implemented in Java.

The basic cheminformatics functionality relies on the open source Java library - [The Chemistry Development Kit, CDK](#).

### 1.2.Codename origin

A major part of the project is concerned with the applicability domain of QSAR models, i.e. how to determine whether a QSAR model can be applied for a given chemical compound. Since it means finding the scope of a QSAR model, the working name of the project became AMBIT<sup>1</sup>

### 1.3.Requirements

**AmbitDisclosure** is a Java application, this mean that:

- ◆ It is not an operating system dependent, i.e. it could run on any environment (Windows, Linux, etc.) where Java is installed (see Appendix A. Redhat Linux screenshot);

---

<sup>1</sup> A dictionary entry for the word "ambit":

Definition: [n] an area in which something acts or operates or has power or control: "the range of a supersonic jet"; "the ambit of municipal legislation"; "within the compass of this article"; "within the scope of an investigation"; "outside the reach of the law"; "in the political orbit of a world power"

Synonyms: compass, orbit, range, reach, scope

## Building blocks for QSAR Decision Support System

- ◆ Java Runtime Environment (JRE) has to be installed in order to run the application. It is not a restrictive requirement, since most modern computers have Java installed by default.

### *1.4.Start AmbitDisclosure*

**AmbitDisclosure** application is deployed in two ways – as a standalone application, which can be downloaded from <http://luna.acad.bg/nina/projects/> and as a **Java Web Start** application, which could be started directly from the web page. In both cases the main executable is the java archive **AmbitDisclosure.jar**. Additional libraries are required, but 1) in the case of Java Web Start, they are downloaded automatically and 2) in the case of download all necessary files are packed into a zip file to be downloaded.

### **What is Java Web Start**

Java Web Start (JWS) is a software technology that allows standalone Java software applications to be deployed with a single click on a Web page or (in Windows) from desktop icons or the Start menu. Java Web Start ensures the most current version of the application will be deployed as well as the correct version of the Java Runtime Environment (JRE). Downloaded applications are cached on the user machine and can be launched offline. Additionally, Java Web Start inherits the Java platform's complete security architecture.

Since Java Web Start is, in itself, a Java application, the software is **platform independent** and can be supported on any client system that supports the Java 2 platform. Java Web Start performs an update automatically when a client application is launched, downloading the latest code from the Web while simultaneously loading the application from a previous cache (provided that a cache exists). Java Web Start also provides a Java Application Manager utility, allowing end-users to organize their Java applications as well as providing a variety of options, such as clearing the cache of downloaded applications, specifying the use of multiple JREs, and setting HTTP proxies.

For more technical details see <http://java.sun.com/products/javawebstart/>.

### *1.5.Standalone application*

Download the application from

<http://luna.acad.bg/nina/projects/downloads.html> . The archive content is listed below:

1	05.04.2005 18:42	29	AmbitDisclosure.bat
2	04.04.2005 19:55	287 952	AmbitDisclosure.jar

## Building blocks for QSAR Decision Support System

3	03.02.2005 15:05	122 006 base.jar
4	03.02.2005 15:05	75 909 cdk-apps.jar
5	24.02.2005 18:41	49 730 cdk-core.jar
6	03.02.2005 15:05	106 490 cdk-experimental.jar
7	24.02.2005 18:41	422 288 cdk-extra.jar
8	24.02.2005 18:41	285 002 cdk-io.jar
9	24.02.2005 18:41	3 008 cdk-libio.jar
10	24.02.2005 18:41	23 329 cdk-render.jar
11	24.02.2005 18:41	141 211 cdk-standard.jar
12	03.02.2005 15:05	815 072 cmlAll.jar
13	05.04.2005 18:44	3 952 898 download.zip
14	03.02.2005 15:05	35 164 Jama-1.0.1.jar
15	03.02.2005 15:05	377 404 jcommon-0.9.6.jar
16	03.02.2005 15:05	993 307 jfreechart-0.9.21.jar
17	03.02.2005 15:05	91 293 jgrapht-0.5.3.jar
18	03.02.2005 15:05	291 443 pmrllib.jar
19	03.02.2005 15:05	114 306 vecmath1.2-1.14.jar
20	05.04.2005 18:37	<DIR> logs
21	05.04.2005 18:39	<DIR> data
22	01.04.2005 r. 19:38	8 705 data\Debnath_smiles.csv
23	04.04.2005 r. 17:57	1 795 data\Glende_smiles.csv

Unzip the archive, retaining the directory structure. Then click on **AmbitDisclosure.jar** file to start the application.

## Building blocks for QSAR Decision Support System

### ***1.6.Start AmbitDisclosure via Java Web start***

Java Web Start must be installed on every user machine that will be used to launch Java applications from the Web. When a user attempts to launch a Web-based application using Java Web Start, the Web browser will launch Java Web Start to start downloading the appropriate files. This is not a restriction, since Java Web Start is included in the recent versions of JRE.

If Java Web Start is installed on the local machine, then the desired application will launch and proceed normally. On the other hand, if Java Web Start is not installed, the user will be prompted to download the program. Once the user has agreed to download Java Web Start and the file has been downloaded, the user must run the program to install Java Web Start.

Below we assume the user has Java Web Start installed and is able to start the AmbitDisclosure through the link

<http://luna.acad.bg/nina/projects/ambit/AmbitDisclosure.jnlp>

- ◆ Initially, Java loading box appears:



- ◆ The splash screen of AmbitDisclosure follows:



- ◆ At this time a number of libraries are loaded. The time will depend on the network connection, but the next time AmbitDisclosure is launched all the necessary files will be read from the cache and the application will start without delay.
- ◆ The main screen of AmbitDisclosure appears.

# Building blocks for QSAR Decision Support System

## 1.7. Main Screen Layout

The following screenshot displays the main screen layout of AmbitDisclosure application.

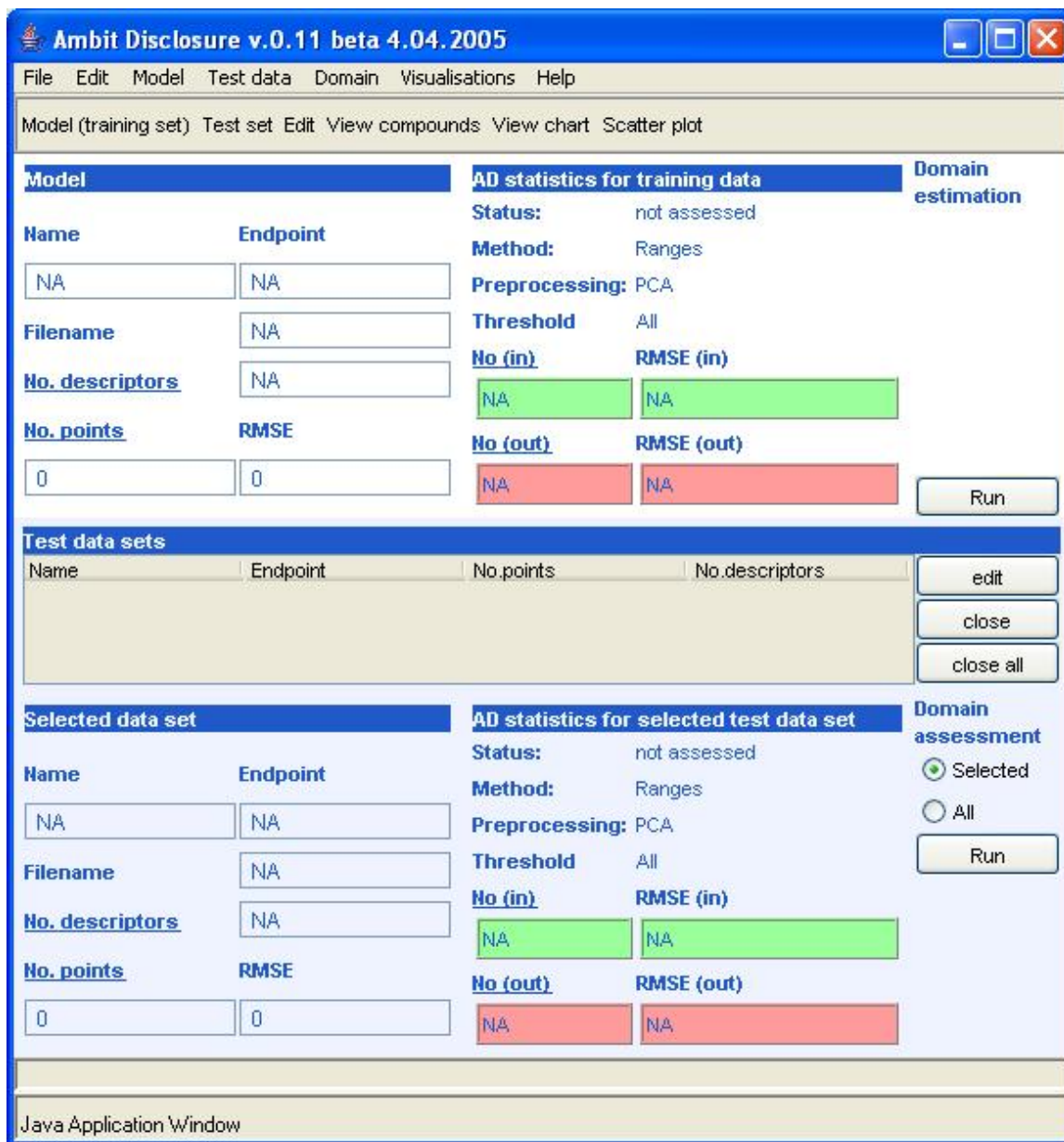


Figure 1. AmbitDisclosure application main screen

The screen comprises a main window with title bar, menu bar, data area, button bar and status bar.

## Building blocks for QSAR Decision Support System

The title bar contains the name of the application and the standard function for maximizing, minimizing and closing the window. The menu bar allows user to navigate between application options.

The main screen is organized around the application functionality. The ultimate objective is to be able to load a QSAR training data set (**a model**), estimate its applicability domain and then be able to assess if compounds from another data set (**the test set**) fall within applicability domain of the model.

The upper part of the screen contains information about the training data set, the bottom one contains information about (multiple) test data sets.

### 1.7.1. Screen layout for the training data set

This screen consists of 3 parts, the data set screen, applicability domain statistics and a panel with a single button, controlling applicability domain estimation.

#### 1.7.1.1. The data set screen

This screen displays the basic information about the data set. (name, endpoint, filename of the file from which data was loaded, number of descriptors, number of compounds, root mean square error). This screen is read only (i.e. data can not be modified).

Mouse clicking on the underlined text generally invokes some relevant information. In particular clicking on the text “No. descriptors” will display popup dialog with editable list of descriptors (see 4.1.View compounds). Clicking on “No. points” will display popup dialog with list of compounds (see 4.2.View descriptors)

Model	
Name	Endpoint
NA	NA
Filename	NA
<u>No. descriptors</u>	NA
<u>No. points</u>	RMSE
0	0

Figure 2. The training data set screen

#### 1.7.1.2. The applicability domain statistics

The next screen part on the right is occupied with applicability domain method information and statistics.

## Building blocks for QSAR Decision Support System

AD statistics for training data	
Status:	not assessed
Method:	Ranges
Preprocessing:	PCA
Threshold	All
No (in)	RMSE (in)
NA	NA
No (out)	RMSE (out)
NA	NA

Figure 3. The training data set statistics screen

### 1.7.1.3. The domain estimation screen

This screen consists of a single button which is used to estimate domain of the training set. Note that the applicability domain of the test data set can not be assessed if training set domain was not estimated. The estimation status is shown on the applicability domain statistics screen (Figure 3)



Figure 4. The domain estimation screen

### 1.7.2. Screen layout for the test data sets

The test data sets screen layout is similar to the one for training data set, but has additional box, displaying the list of test data sets. In contrast to training data set, which could be loaded only one at a time; there could be multiple test data sets.

## Building blocks for QSAR Decision Support System

Name	Endpoint	No. points	No. descriptors
------	----------	------------	-----------------

Figure 5. An empty list of test data sets

Name	Endpoint	No. points	No. descriptors
data\Debnath_smiles....	Pred	18	4
data\Debnath_smiles....	Pred	0	4

Figure 6. A list with two test data sets

Test data sets are added by **File / New / Test data set** or **File / Open / Test data set** menu items. Test data sets could be edited or closed by the corresponding buttons at the right of the list.

Mouse click on the lists select the clicked data set and updates the information on the selected test data set screen (the bottom panel)

### 1.7.3. Screen layout for the selected test data set

Selected data set		AD statistics for selected test data set		Domain assessment
Name	Endpoint	Status:	assessed	<input checked="" type="radio"/> Selected
ebnath_smiles.csv	Pred	Method:	Ranges	<input type="radio"/> All
Filename		Preprocessing:	PCA	Run
No. descriptors	4	Threshold	100.0%	
No. points	RMSE	No. (in)	RMSE (in)	
18	0	10	1,68	
		No. (out)	RMSE (out)	
		8	2,92	

Figure 7. The selected test data set screen

This is similar to the upper panel displaying the training data set, with the only difference that there is no **“Domain Estimation”** screen, but there is **“Domain Assessment”** screen.

## Building blocks for QSAR Decision Support System

Test data domain assessment can be done only if the training data set domain has been already estimated by running **Domain estimation**.

If assessment is successful, the middle panel (**AD statistics for selected data set**) is updated with the corresponding status and statistics information.

Once the domain is estimated, compounds in- and out-of domain can be viewed (see 3.5.1.Compounds menu) and results can be exported to a text file (see 3.1.3.Save menu).

## 2.Getting started

### 2.1.Load training data set (a model)

Click on File / Open / Model (training set). A standard file open dialog appears.

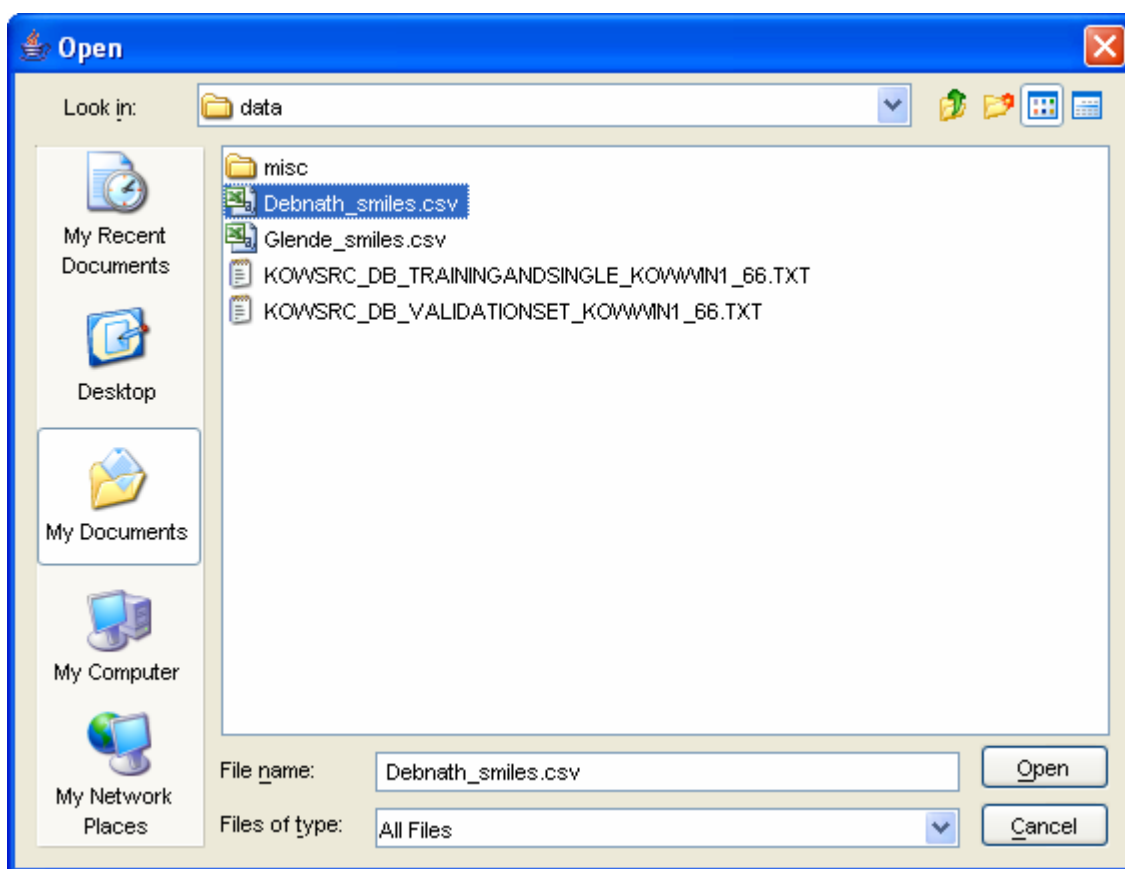


Figure 8. File open dialog

From data directory distributed along with the application select **“Debnath\_smiles.csv”**. This file can also be downloaded from <http://luna.acad.bg/nina/projects/ambit/data/>

## Building blocks for QSAR Decision Support System

The current version of the application permits only loading of CSV<sup>2</sup> files. The data in the CSV files are arranged in columns as on the figure below (Figure 9):

Code	Compound	CAS	SMILES	Obs	Pred	Dev.	log P	eLumo	eHomo	IL
1	2-bromo-7-aminofluorene		NC1=CC2=C(C	2.62	2.62	0	3.92	-0.405	-8.236	1
2	2-methoxy-5-methylaniline(p-cresidine)		COC1=CC=C(C	-2.05	-2.06	0.01	1.74	0.419	-8.449	0
3	5-aminoquinoline	611-34-7	NC1=CC=CC2	-2	-2.02	0.02	1.16	-0.395	-8.395	0
4	4-ethoxyaniline	156-43-4	CCOC1=CC=C	-2.3	-2.32	0.02	1.24	0.513	-8.182	0

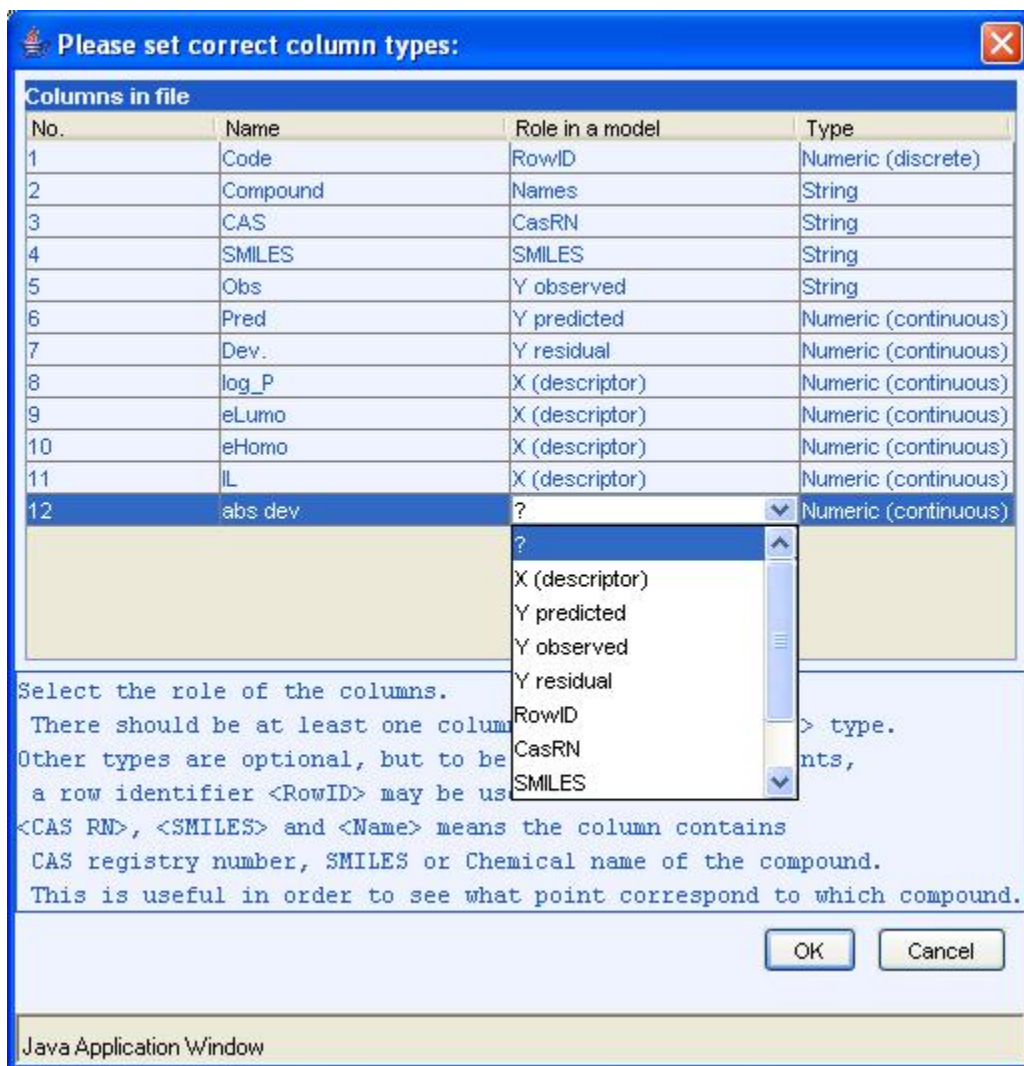
**Figure 9. An example of CSV file viewed with MSExcel**

The columns in a CSV file can have arbitrary names; therefore it is crucial to specify the role of each column in the data set. The dialog below (Figure 10) appears automatically after selecting file name from the open file menu.

---

<sup>2</sup> CSV files are comma-delimited text files, which can be read and write by Microsoft Excel. We suggest preparing your data in MS Excel and exporting as CSV file. Further versions of AmbitDisclosure will support additional file formats.

## Building blocks for QSAR Decision Support System



**Figure 10. Dialog to specify the roles of CSV columns in the data set.**

Please be accurate in specifying the columns' role in the model, as what will be read depends on your choice. No column is mandatory (even "**X (descriptor)**"), but the lack of some information can make impossible using some applicability domain methods.

- ◆ For example if no "**X (descriptor)**" columns are specified, descriptors based applicability domain estimation can not be performed.

## Building blocks for QSAR Decision Support System

- ◆ If no structure information for a compound (i.e. SMILES) is supplied, then it will be impossible to estimate applicability domain by fingerprints<sup>3</sup>.
- ◆ If information for “**Y predicted**”, “**Y observed**” or “**Y residual**” is missing, the applicability domain estimation can be performed, but RMSE statistics will not be available.
- ◆ Unknown or irrelevant columns can be marked with “**?**”.

Click “**OK**” when selection is finished. The training data set screen is updated with number of descriptors and number of points (compounds) as follows:

Model	
Name	Endpoint
Model	Pred
Filename	
<b><u>No. descriptors</u></b>	4
<b><u>No. points</u></b>	<b>RMSE</b>
88	0

Figure 11. AmbitDisclosure training data set screen with data set loaded.

### *2.2. Estimate applicability domain of the training data set*

Once training data set is loaded, applicability domain can be estimated. To do this, just click on the button “**Run**” on the domain estimation panel (upper right, Figure 4).

---

<sup>3</sup> Further versions of AmbitDisclosure will make possible to connect to the database and search compounds by CAS registry numbers and names. For now an online search from <http://luna.acad.bg/nina/projects/ambit/php/index.htm> could be used.

## Building blocks for QSAR Decision Support System

AD statistics for training data	
Status:	assessed
Method:	Ranges
Preprocessing:	PCA
Threshold	100.0%
No (in)	RMSE (in)
88	0,83
No (out)	RMSE (out)
0	0

Figure 12. AmbitDisclosure training data set statistics screen with estimated applicability domain statistics.

The default applicability domain method is **Ranges**. To change the method, use menu **Domain** / **Options** / **Method** (see 3.5.2.1.Method) and then click **Run** button again.

### *2.3.Export results to a text file*

Use **File** / **Save** / **Model** to save the results in a comma-delimited (CSV) text file. The standard File Save dialog appears.

## Building blocks for QSAR Decision Support System

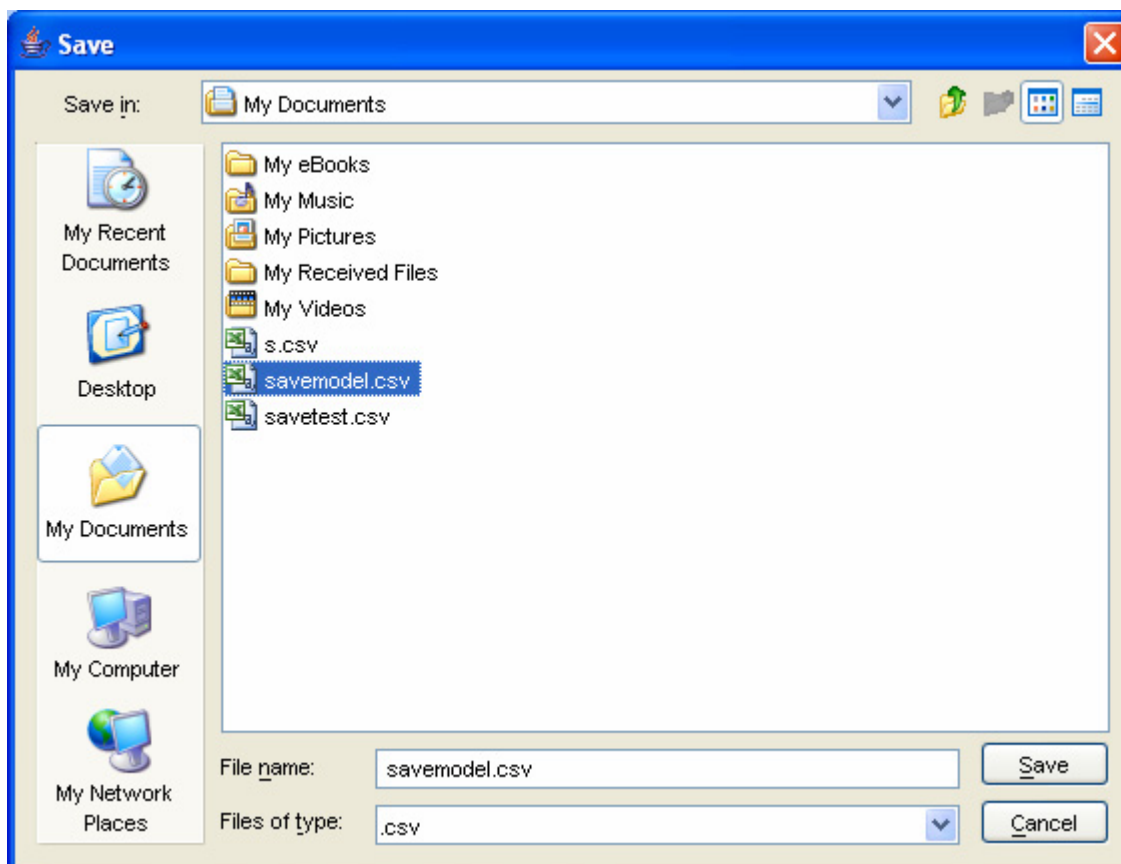


Figure 13. Save File dialog.

The resulting CSV file looks like on Figure 14.

#	ID	CAS	NAME	SMILES	0.log_P	1.eLumo	2.eHomo	3.iL	Predicted	Observed	Error	Equation	Ranges	in domain
1	0		2-bromo-7-	NC1=CC2=	3.92	-0.405	-8.236	1	2.62	2.62	0		0	TRUE
2	1		2-methoxy	COC1=CC	1.74	0.419	-8.449	0	-2.06	-2.05	0.01		0	TRUE
3	2	611-34-7	5-aminoqu	NC1=CC=	1.16	-0.395	-8.395	0	-2.02	-2	0.02		0	TRUE
4	3	156-43-4	4-ethoxyar	CCOC1=C	1.24	0.513	-8.182	0	-2.32	-2.3	0.02		0	TRUE
5	4		1-aminona	NC1=C2C	2.25	-0.195	-8.108	0	-0.63	-0.6	0.03		0	TRUE

Figure 14. An example of exported results viewed within MSEXcel.

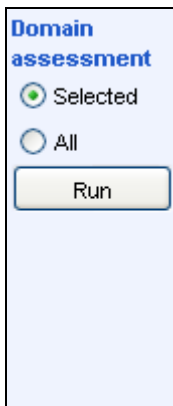
### 2.4. Load test data set

The procedure is similar to loading a model. Click on **File / Open / Test set**. A standard file open dialog appears and then a dialog to specify the role of CSV columns in the data set (as in Figure 10).

Upon successful load, the data set is loaded into the test data set list (Figure 5) and selected test data screen is updated (Figure 7).

### ***2.5. Assess applicability domain of the test data set***

To assess applicability domain of the selected data set, click on the **Run** button on the domain assessment panel. If **“All”** radio button is selected, all test data sets will be assessed, otherwise only the selected one.



**Figure 15. Test data set Domain assessment panel.**

Note that to be able to assess test dataset domain, the domain has to be estimated for the training data set (see above).

### ***2.6. View / Export results of the test data set domain estimation***

This is similar to the procedure already explained for the training set.

Use **Domain / Compounds / Test set** to view compounds in- and out-of domain

Use **File / Save / Test set** to export results to a CSV file.

## **3. Menu structure**

AmbitDisclosure main menu consists of File, Edit, Model, Test data, Domain, Visualisation and Help items. Sub items will be described below.



**Figure 16. AmbitDisclosure main menu**

## Building blocks for QSAR Decision Support System

### ***3.1.File***

#### **3.1.1.New**

##### **3.1.1.1.Model (Training set)**

Creates an empty training set.

##### **3.1.1.2.Test set**

Creates an empty test set and adds it to the test data sets list.

#### **3.1.2.Open**

##### **3.1.2.1.Model (Training set)**

Loads training data set from a file

##### **3.1.2.2.Test set**

Loads a test data set from a file and adds the resulting dataset to the test data sets list.

#### **3.1.3.Save**

##### **3.1.3.1.Model (Training set)**

Export the training data set to a text (CSV) file. Domain assessment results are also exported if available.

##### **3.1.3.2.Test set**

Export the selected test data set to a text (CSV) file. Domain assessment results are also exported if available.

### ***3.2.Edit***

Provide menu access to the standard text operations Cut, Copy and Paste.

### ***3.3.Model***

#### **3.3.1. Edit**

#### **3.3.2. View compounds**

## Building blocks for QSAR Decision Support System

No.	CAS	Name	Pred	YO...	log_P	eL...	eH...	IL
1		2-bromo...	2.62	2.62	3.92	-0...	-8...	1.0
2		2-methox...	-2.06	-2.05	1.74	0.419	-8...	0.0
3	611-...	5-aminoq...	-2.02	-2.0	1.16	-0...	-8...	0.0
4	156-...	4-ethoxya...	-2.32	-2.3	1.24	0.513	-8...	0.0
5		1-aminon...	-0.63	-0.6	2.25	-0...	-8...	0.0
6		4-aminofl...	1.1	1.13	2.7	-0...	-8...	1.0
7		2-aminoa...	2.65	2.62	3.26	-0...	-7...	1.0
8		7-aminofl...	2.83	2.88	3.72	-0...	-8...	1.0
9		8-aminoq...	-1.08	-1.14	1.79	-0...	-8...	0.0
10		1,7-diamin...	0.82	0.75	1.64	-1...	-8...	1.0
11		2-aminon...	-0.74	-0.67	2.28	-0...	-8...	0.0
12		4-aminop...	3.25	3.16	3.72	-0.85	-7...	1.0
13		3-amino-3...	-0.44	-0.55	2.68	-1...	-8...	0.0
14		2,4,5-trim...	-1.19	-1.32	2.41	0.581	-8...	0.0
15	6344...	3-aminofl...	1.02	0.89	2.7	-0...	-8...	1.0
16		3,3'-dichl...	0.67	0.81	3.51	-0...	-8...	0.0
17		2,4-dimet...	-2.05	-2.22	1.68	0.605	-8...	0.0
18		2,7-diamin...	0.31	0.48	1.47	0.0	-7...	1.0
19		3-aminofl...	3.49	3.31	4.2	-0...	-8...	1.0
20		2-aminofl...	1.73	1.93	3.14	-0...	-8...	1.0
21		2-amino-4...	-0.42	-0.62	2.68	-1...	-8...	0.0
22		4-aminobi...	-0.34	-0.14	2.86	0.048	-8...	0.0
23		3-methox...	-2.19	-1.96	1.52	0.583	-8...	0.0
24		2-aminoc...	0.84	0.6	2.3	0.025	-8...	1.0
25		2-amino-5...	2.26	2.53	4.26	0.0	-8...	0.0

**Molecule**

CAS: 611-34-7

Formula:

Name: 5-aminoquinoline

SMILES: NC1=CC=CC2=C1C=CC=N2

Structure diagram

Name	Value
Y predicted	-2.02
Y observed	-2.0
log_P	1.16
eLumo	-0.395
eHomo	-8.395

Prev Next Add Delete

OK Cancel

Figure 17. A screen, displaying training data set compounds, descriptor and predicted / observed values

### 3.3.3. View chart

If training data set is loaded, this scatter plot appears upon clicking on **Model / View Chart** menu item.

## Building blocks for QSAR Decision Support System

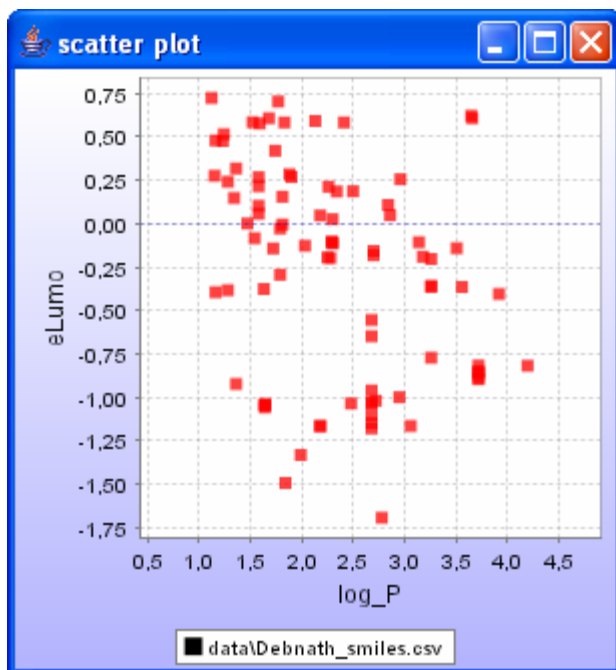


Figure 18. Scatter plot of a training data set

### 3.4. Test data

#### 3.4.1. Edit

The same as **Model/ Edit** menu, but displays selected data set.

A dialog box titled "Edit selected test data set" with a close button (X). It contains the following fields:

Data set	
Name	Endpoint
Test data set for the	Pred
Filename	data\my_data.csv
No. descriptors	4
No. points	RMSE
14	0

Buttons: OK, Cancel

Figure 19. Test data set edit

## Building blocks for QSAR Decision Support System

### 3.4.2. View compounds

The same as **Model/ View compounds** menu, but displays compounds from the selected test data set.

### 3.4.3. View chart

The same as **Model/ View chart** menu, but points on the chart reflect the selected test data set.

## 3.5. Domain

### 3.5.1. Compounds

This menu allows viewing test and training set compounds in- and out-of domain. The figure below displays the compounds from the demo test , which has been classified out-of-domain by the fingerprint method.

**Compounds out of domain**

#	No	Name
1	7	"1-Ethyl-2-aminofluorene",CCC1=C(N)C...
2	8	"1-iPropyl-2-aminofluorene",CC(C)C1=C...
3	9	"1-nButyl-2-aminofluorene",CCCC1=C(...
4	10	"1-tButyl-2-aminofluorene",CC(C)(C)C1...
5	12	"3-Ethyl-4-aminobiphenyl",CCC1=C(N)C...
6	13	"3-iPropyl-4-aminobiphenyl",CC(C)C1=C...
7	14	"3-nButyl-4-aminobiphenyl",CCCC1=C(...
8	15	"3-tButyl-4-aminobiphenyl",CC(C)(C)C1...
9	17	"3,5-Diethyl-4-aminobiphenyl",CCC1=CC...
10	18	"3,5-Dipropyl-4-aminobiphenyl",CC(C)C...

**Molecule**

CAS

Formula

Name: 3,5-Dipropyl-4-aminobiphenyl

SMILES: CC(C)C1=CC(=CC(C(C)C)=C1N)C2=CC=CC=C2

Structure diagram

OK Cancel

Figure 20. Test set compounds out-of-domain

## Building blocks for QSAR Decision Support System

### 3.5.2.Options

#### 3.5.2.1.Method

The options below are mutually exclusive and are used to select particular applicability domain assessment method. For method details see 5.Applicability domain estimation.

Note that this menu item only alters the method type. All results from previous estimation/assessments are lost.

Domain estimation and assessment should be done by pressing buttons **“Run”** (See 1.7.1.3. The domain estimation screen).

##### 3.5.2.1.1.Ranges

Sets descriptor ranges as an applicability domain estimation method.

##### 3.5.2.1.2.Euclidean distance

Sets Euclidean distance as an applicability domain estimation method.

##### 3.5.2.1.3.City-block distance

Sets City-block distance as an applicability domain estimation method.

##### 3.5.2.1.4.Probability density

Sets non-parametric probability density estimation as an applicability domain estimation method.

##### 3.5.2.1.5.Fingerprints

Sets fingerprints as an applicability domain estimation method.

#### 3.5.2.2.Data preprocessing

Principal Component Analysis (PCA) is the only option available here. See 5.1.Data preprocessing.

#### 3.5.2.3.Threshold

Allows the user to select the percent of points in the training set which will determine the domain of the model. The default is all points (100%).

The threshold usage differs slightly with different methods. (See 5.Applicability domain estimation)

## Building blocks for QSAR Decision Support System

### 3.5.2.4.Distance options

Distances in descriptor space are calculated from a point to a certain central point of the training data set. The mean of the data set is used as a central point by default, but other options are also available

- ◆ Mean (i.e.  $\sum (x_1..x_n)/n$ )
- ◆ Center (i.e.  $\max - \min / 2$ )
- ◆ Origin of the coordinate system (the point (0,0,..0) )

This setting has no effect on **Ranges**, **Probability** and **Fingerprint** methods, but could influence results for the distance methods.

### 3.5.2.5.Fingerprint comparison

This selects how fingerprint will be compared in order to estimate and assess applicability domain. Two options are available: “**Missing fragments**” and “**Tanimoto distance**”. For details see 5.3.Structure-based methods.

## 3.6. Visualizations

Up to now the same as **Model** / **View** chart

## 3.7.Help

### 3.7.1.Help

Not implemented yet.

### 3.7.2.Java

Some basic information about Java VM is displayed. The output looks like

## Building blocks for QSAR Decision Support System

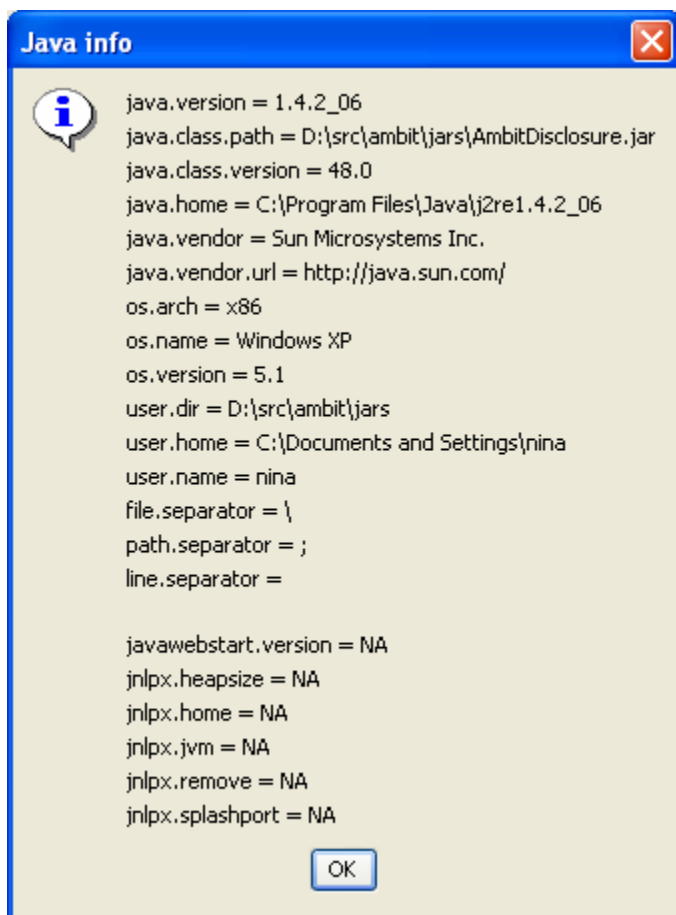


Figure 21. Java information

### 3.7.3. Demo :

There are two options here : **“Demo: Debnath mutagenicity model”** and **“Demo : test set for Debnath mutagenicity model”**

This menu items are provided to easily load demo training and test data set. However, they rely on finding files

- ◆ For the model : “Debnath\_smiles.csv”
- ◆ For the test set: “Glende\_smiles.csv”.

in a folder **data**, located just below the file **AmbitDisclosure.jar**.

The files can be downloaded from <http://luna.acad.bg/nina/projects/ambit/data/>.

This menu will not be able to run if the AmbitDis

## 4. Other options

These options are generally accessible by clicking on the underlined text anywhere on the interface

### 4.1. View compounds

Compounds from the training set can be viewed by selecting the menu item **Model/ View compounds** as well as the underlined text **“No. Points”** on the training data set screen. A window as in Figure 17 appears.

Compounds from the test set can be viewed by selecting the menu item **Test data/ View compounds** as well as the underlined text **“No. Points”** on the selected test dataset screen. A window as in Figure 17 appears.

### 4.2. View descriptors

This dialog is invoked by clicking on the underlined text **“No. Descriptors”** on a dataset panel. Descriptors can be modified / deleted and added.

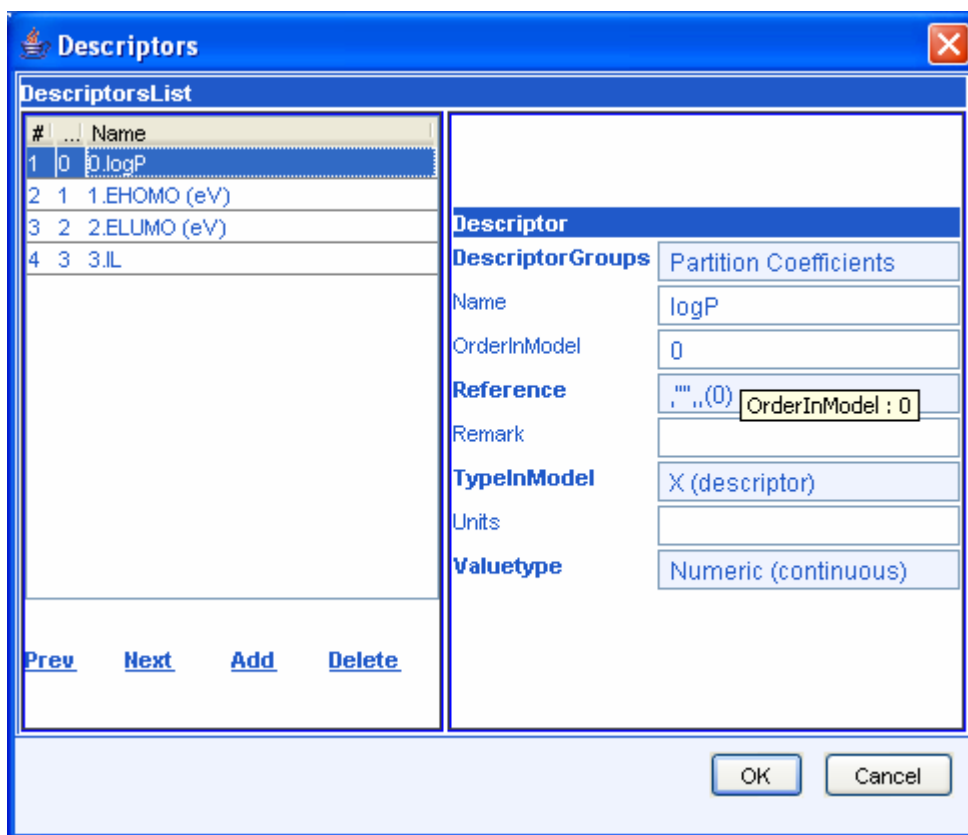


Figure 22. AmbitDisclosure view descriptors

## 5. Applicability domain estimation

### 5.1. Data preprocessing

Data standardization and PCA are performed if **Domain/Options/Data Preprocessing/PCA** menu item is checked, otherwise data is taken as it is. No data preprocessing is done for the structure-based methods.

### 5.2. Descriptor-based methods

AmbitDisclosure provides four applicability domain assessment methods : ranges, Euclidean distance, city-block distance, probability density.

See the review of applicability domain methods at

[http://ecb.jrc.it/DOCUMENTS/QSAR/AD\\_methods.zip](http://ecb.jrc.it/DOCUMENTS/QSAR/AD_methods.zip)

### 5.3. Structure-based methods

The implementation of structure –based applicability domain in AmbitDisclosure is based on fingerprints. Since the AMBIT project uses the open source cheminformatics Java library - [The Chemistry Development Kit, CDK](#)., fingerprint implementation uses CDK' fingerprint generation procedure [Fingerprinter](#). It follows the ideas of [Daylight fingerprint theory](#), i.e.

- ◆ for a given molecule all possible paths for a predefined length (default is 7) are generated,
- ◆ the path is submitted to a hash function which uses it as a seed to a pseudo-random generator
- ◆ the hash function outputs a set of bits
- ◆ the set of bits thus produced is added (with a logical OR) to the fingerprint.

Fingerprint length (the number of bits) is user defined, but 1024 bit (the default) fingerprints are generally used.

Applicability domain assessment with fingerprints is done in several steps.

- ◆ Fingerprints are calculated for each compound in the training data set. Compounds for which no structural information is available are ignored.

## Building blocks for QSAR Decision Support System

- ◆ Fingerprint profile is calculated, i.e. the information stored is how many compounds have bit  $i$  set to one,  $i=1..1024$
- ◆ Consensus fingerprint is calculated. The exact procedure depends on fingerprint distance type selected (see 3.5.2.5.Fingerprint comparison)

### 5.3.1.Fingerprints (Missing fragments)

A consensus fingerprint bit is set to one if the corresponding bit is set to one by at least N compounds. The number N is determined by user-defined threshold (percent of compounds). See 3.5.2.3.Threshold.

Then the number of fingerprint bits zero in the consensus fingerprint, but set to one in the query compound fingerprint is calculated. This serve as a measure for applicability domain (compounds with zero number belongs to the applicability domain of the model, others do not).

### 5.3.2.Fingerprints (Tanimoto distance)

A consensus fingerprint bit is set to one if the corresponding bit is set to one by at least one compound.

The (1-Tanimoto) distance is calculated between each compound of the training set and the consensus fingerprint. The values are sorted and the highest value for the set of (percent\*N) compounds is taken as a threshold, where N is the number of compounds in the training set and percent is the user defined value as in 3.5.2.3.Threshold.

A query compound is classified in-the-domain if the (1-Tanimoto) distance between its fingerprint and the consensus fingerprint is below the estimated threshold; otherwise it is classified out-of-domain.

## Appendix A. Redhat Linux screenshot

This screenshot is provided as an illustration that **AmbitDisclosure** is able to run on different platforms. Here the application was started by Java Web Start using Mozilla Browser on Redhat Linux.

# Building blocks for QSAR Decision Support System

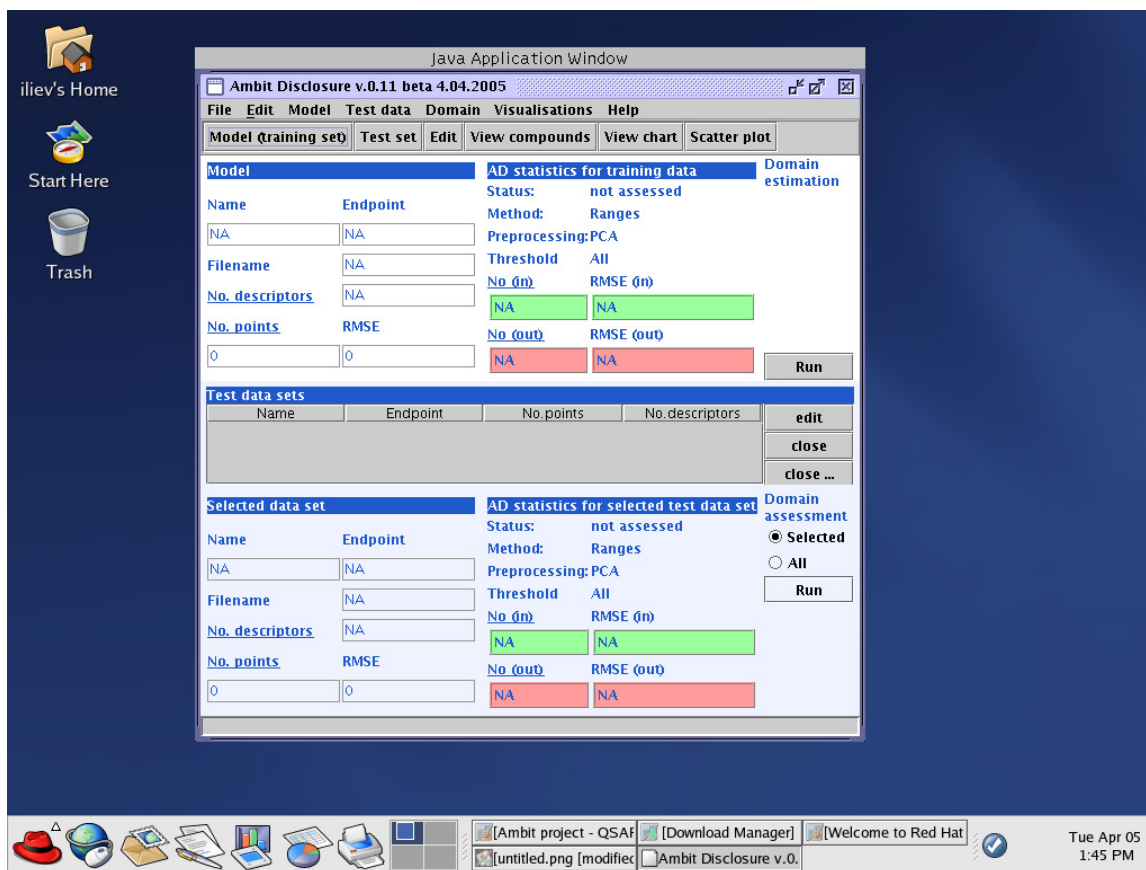


Figure A.1. AmbitDisclosure running on Redhat Linux

## Table of Contents

1 .Introduction.....	2
1.1 .Background.....	2
1.1.1 .Implementation .....	2
1.2 .Codename origin.....	2
1.3 .Requirements .....	2
1.4 .Start AmbitDisclosure.....	3
What is Java Web Start.....	3
1.5 .Standalone application.....	3
1.6 .Start AmbitDisclosure via Java Web start .....	5
1.7 .Main Screen Layout.....	6
1.7.1 . Screen layout for the training data set.....	7
1.7.1.1 . The data set screen .....	7
1.7.1.2 . The applicability domain statistics.....	7
1.7.1.3 . The domain estimation screen .....	8
1.7.2 . Screen layout for the test data sets.....	8
1.7.3 . Screen layout for the selected test data set.....	9
2 .Getting started.....	10
2.1 .Load training data set (a model) .....	10
2.2 .Estimate applicability domain of the training data set.....	13
2.3 .Export results to a text file .....	14
2.4 .Load test data set.....	15
2.5 .Assess applicability domain of the test data set.....	16
2.6 .View / Export results of the test data set domain estimation.....	16

## Building blocks for QSAR Decision Support System

3 .Menu structure .....	16
3.1 .File .....	17
3.1.1 .New .....	17
3.1.1.1 .Model (Training set) .....	17
3.1.1.2 .Test set .....	17
3.1.2 .Open .....	17
3.1.2.1 .Model (Training set) .....	17
3.1.2.2 .Test set .....	17
3.1.3 .Save .....	17
3.1.3.1 .Model (Training set) .....	17
3.1.3.2 .Test set .....	17
3.2 .Edit .....	17
3.3 .Model .....	17
3.3.1 . Edit .....	17
3.3.2 . View compounds .....	17
3.3.3 . View chart .....	18
3.4 .Test data .....	19
3.4.1 . Edit .....	19
3.4.2 . View compounds .....	20
3.4.3 . View chart .....	20
3.5 .Domain .....	20
3.5.1 .Compounds .....	20

## Building blocks for QSAR Decision Support System

3.5.2 .Options.....	21
3.5.2.1 .Method.....	21
3.5.2.2 .Data preprocessing.....	21
3.5.2.3 .Threshold.....	21
3.5.2.4 .Distance options.....	22
3.5.2.5 .Fingerprint comparison.....	22
3.6 .Visualizations.....	22
3.7 .Help.....	22
3.7.1 .Help.....	22
3.7.2 .Java.....	22
3.7.3 . Demo :.....	23
4 .Other options.....	24
4.1 .View compounds.....	24
4.2 .View descriptors.....	24
5 .Applicability domain estimation.....	25
5.1 .Data preprocessing.....	25
5.2 .Descriptor-based methods.....	25
5.3 .Structure-based methods.....	25
5.3.1 .Fingerprints (Missing fragments).....	26
5.3.2 .Fingerprints (Tanimoto distance).....	26
Appendix A. Redhat Linux screenshot.....	26