

GUSAR

User Guide

**V.V.Poroikov, D.A.Filimonov, A.A.Lagunin, A.V.Zakharov
& Associates**

CONTENTS

CHAPTER 1 ABOUT THIS MANUAL

Organization of this Document	4
Abbreviations	5
Copyright Notice	6
How to Contact Us.	7

CHAPTER 2 GUSAR SOFTWARE PRODUCT

About GUSAR	8
Hardware and Software System Requirements	9

CHAPTER 3 HOW GUSAR WORKS

Input Data Formats	10
Prediction Results	12

CHAPTER 4 GETTING STARTED WITH GUSAR

Getting Started with GUSAR Interface	17
Workflow of QSAR model(s) creation and external test prediction .	21

CHAPTER 5 GUSAR INTERFACE AND FUNCTIONS

Main window	22
Viewer window	24
Opening SD files	31
Add data to SAR base	32
Save results of LOO	33
Opening SAR base	34
Delete data from SAR base	35
Save SAR base	36
Activity Selection window	37
Model Selection window	40
Create QNA models	44

Create MNA models	45
Create combinatorial models	46
Prediction by GUSAR	47
Y randomization.	49
Leave-Many-Out Options	50
Options of models	51
CHAPTER 6 TROUBLE SHOOTING	56
CHAPTER 7 TERMS AND DEFINITIONS	
Basis of QSAR	57
Molecular Descriptors	57
Multilevel Neighborhoods of Atoms (MNA) Descriptors	57
Quantitative Neighborhoods of Atoms (QNA) Descriptors.	59
Physico-Chemical Descriptors.	62
Methods	65
Self-Consistent Regression (SCR)	65
Radial Basis Function (RBF)	65
Use of Nearest Neighbours.	67
Consensus method.	68
Applicability Domain Assessment	70
Similarity.	70
Leverage.	70
Accuracy of three nearest neighbours' predictions	70
Validation Methods	72
Leave-Many-Out procedure	72
Y-Randomization procedure	72
REFERENCES	73

CHAPTER 1

ABOUT THIS MANUAL

ORGANISATION OF THIS DOCUMENT

This document consists of eight chapters.

- **Chapters 1 - 2** contain general information about the program.
- **Chapters 3 - 4** describe how to work with the program.
- **Chapter 5** contains detailed description of GUSAR interface.
- **Chapter 6** describes what to do in case of trouble shooting.
- **Chapter 7** contains theoretical introduction to using terms and definitions.
- **References** contain the list of publications related to GUSAR.

ABBREVIATIONS

AD – Applicability Domain

CSV - Comma-Separated Values file: A CSV file is used for the digital storage of data structured in a table of lists form.

LOO CV – Leave-One-Out Cross-Validation

MNA - Multilevel Neighbourhoods of Atoms

PASS - Prediction of Activity Spectra for Substances

QNA - Quantitative Neighbourhoods of Atoms

QSAR - Quantitative Structure-Activity Relationships

QSPR - Quantitative Structure-Property Relationships

RMSE – Root Mean Square Estimation

SAR - Structure-Activity Relationships

SCR – Self-Consisted Regression

SDfiles – Structure-data files: An SDfile contains structures and/or data for any number of molecules. SDfile is the format for import/export of chemical data

SPR – Structure-Property Relationships

COPYRIGHT NOTICE

Copyright © 2006-2014 by V.V. Poroikov, D.A. Filimonov, A.A. Lagunin, A.V. Zakharov & Associates. All rights reserved.

ISIS/Base (MDL ISIS/Base earlier), ISIS/Draw (MDL ISIS/Draw earlier) are registered trademarks of Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA. <http://accelrys.com>.

All other product names are trademarks or registered trademarks of their respective holders.

No part of this document may be reproduced by any means except of permitted in written by V.V. Poroikov, D.A. Filimonov, A.A. Lagunin, A.V. Zakharov & Associates.

HOW TO CONTACT US

If you have any questions about GUSAR program, please contact us by E-mail:
support@way2drug.com.

GUSAR home page:

<http://www.way2drug.com/GUSAR/>

CHAPTER 2

GUSAR software product

ABOUT GUSAR

GUSAR software was developed to create QSAR/QSPR models on the basis of the appropriate training sets represented as SDfile contained data about chemical structures and endpoint in quantitative terms.

GUSAR is a commercially available software product.

HARDWARE AND SOFTWARE SYSTEM REQUIREMENTS

Processor	x86 family - Intel® Pentium® or compatible.
Operating environment	Microsoft® Windows® XP/Vista/7.
Memory	1 Gb of RAM (2 Gb or more recommended).
Hard disk	minimum 100 MB free hard disc space.
Display	1024x768 or higher resolution.

Mouse or other compatible pointing device is recommended.

Chemical structure information is represented as SDfiles (formats of Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA. <http://accelrys.com>), which can be exported from many chemical Database Management Systems. The only requirement: these files should conform to the Symyx V2000 and/or V3000 standard.

CHAPTER 3

HOW GUSAR WORKS

GUSAR is commercially available computer program for analysis of quantitative structure-activity/structure-property relationship (QSAR/QSPR) on the basis of the structural formulas of the compounds and data on their activity/property, and prediction of activity/property for new compounds. GUSAR can be easily applied to different routine QSAR tasks, for building many models, and for prediction by these models of the different quantitative values simultaneously.

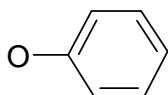
Theoretical bases of GUSAR work are described in Chapter 7. If any question arise during the reading of the User Guide, please, consult with the Chapter 7.

If you will not find the answer on your question, please, contact us:
support@way2drug.com

INPUT DATA FORMATS

The software GUSAR uses SD files as external sources of structure and activity data to prepare both the SAR Base and set of substances to be predicted. SD files can be exported from many molecular editors or databases such as ISIS/Base (<http://accelrys.com>).

The example of the record in SD file with experimental results of Algae acute toxicity (EC50) for Phenol molecule:



Phenol

```

-ISIS- 01180614322D

7 7 0 0 0 0 0 0 0 0 0999 V2000
-0.7107 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.2130 -0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.2130 0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.7174 0.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.7824 -0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.7824 0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.2892 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 2 0 0 0 0
1 3 1 0 0 0 0
1 4 1 0 0 0 0
2 5 1 0 0 0 0
3 6 2 0 0 0 0
5 7 2 0 0 0 0
6 7 1 0 0 0 0
M END
> <ID> (1)
1

```

```
> <EC50> (1)
-1.4600000000000000e+000
```


```
> <Toxicity> (1)
active
```

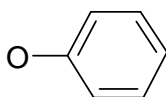
```
$$$$
```

GUSAR allows to create both quantitative (QSAR/QSPR) and classification (SAR/SPR) models. To achieve quantitative model it is necessary to use field with quantitative value of modeling property in the SD file. For creation of classification models it is necessary to use especial prepared field in the SD file. In this case, the field of modeling property in the SD file has to be represented by the following records: active / inactive or 1 / 0. As it can be seen from the figure presented above the field > <EC50> can be used for creation of quantitative models and the field > <Toxicity> can be used for creation of classification models.

PREDICTION RESULTS

A result of GUSAR prediction is saved in two types of SD file or CSV file.

The example of the record in SD files with GUSAR prediction results of Algae acute toxicity (EC50) for Phenol molecule (after the press of the button ):



Phenol

```


-ISIS- 01180614322D
7 7 0 0 0 0 0 0 0 0 0999 V2000
-0.7107 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.2130 -0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.2130 0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.7174 0.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.7824 -0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.7824 0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.2892 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 2 0 0 0 0
1 3 1 0 0 0 0
1 4 1 0 0 0 0
2 5 1 0 0 0 0
3 6 2 0 0 0 0
5 7 2 0 0 0 0
6 7 1 0 0 0 0
M END
> <ID> (1)
1
> <EC50> (1)
-1.4600000000000000e+000
> <GUSAR_EC50>
-1.055
> <GUSAR_EC50_AD>
0.958
$$$$

```

Where, the field "> <EC50>" contains the experimental data of EC50 in log10(mmol/kg). The field "> <GUSAR_EC50>" contains the predicted value of EC50 in log10(mmol/kg). The field "> <GUSAR_EC50_AD >" contains the value of Applicability Domain (AD). AD reflects a measure of similarity between the studied molecule and the training set for prediction results based on a single QSAR model (it varies from 0 to 1).

The values close to 0 means that a studied compound is out of AD. The values close to 1 means that a studied compound is in AD (AD > 0.7 is a cutoff value for correct prediction by default). For the consensus models this field contains the record "in AD" or "out AD". "in AD" means that the prediction result is in the applicability domain. "out AD" means that the prediction result is out of applicability domain. The detailed information about Applicability Domain is represented in Chapter 7. The cutoff value of AD can be changed by the user (see Chapter 5, Options of models, page 51).

The example of the record in SD files with GUSAR prediction results of *Algae* and *Vibrio*

Fischeri acute toxicity (EC50) for Phenol molecule (after the press of the button ):

```
-ISIS- 01180614322D

7 7 0 0 0 0 0 0 0 0 0999 V2000
-0.7107 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.2130 -0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.2130 0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.7174 0.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.7824 -0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.7824 0.8654 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.2892 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 2 0 0 0 0
1 3 1 0 0 0 0
1 4 1 0 0 0 0
2 5 1 0 0 0 0
3 6 2 0 0 0 0
5 7 2 0 0 0 0
6 7 1 0 0 0 0
M END
> <ID> (1)
1


> <EC50> (1)
-1.4600000000000000e+000

> <GUSAR_RESULTS_COUNT>
2

> <GUSAR_RESULTS>
-1.055 0.958 EC50 (Algae, log10(millimole))
3.099 out of AD EC50 (Vibrio fischeri, log10(micromole))

$$$$
```

Where, the field "> <GUSAR_RESULTS_COUNT>" contains information about the number of predicted activities, the field "> <GUSAR_RESULTS>" contains GUSAR predicted results for appropriate activity (s) (log10(EC50) for *Algae* and for *Vibrio fischeri* in the current example). First value of each line is a predicted value of EC50, second value is the value of AD.

The example of GUSAR prediction results of Algae acute toxicity (EC50) for Phenol molecule in CSV files (after the press of the button ):

GUSAR - Prediction of Values for Substances
 Copyright (C) 2013 V. Poroikov, D. Filimonov
 & Associates

Chemical Structure File: algae.sdf

Table of 1 predictable activities.

ID	EC50	AD
1	-1.0549	0.9575
2	-1.2095	0.9617
3	-1.4042	0.959
4	-0.9528	0.8163
5	-1.1643	0.8502
6	-0.9839	0.9384
7	-1.1103	0.9565
8	-1.1875	0.9327
9	-1.185	0.9646
10	-0.5627	0.6594
11	-0.6899	0.8552
12	-1.1418	0.01
13	-0.44	0.944
14	-0.3506	0.9744

Where, ID –structure identifier selected by the user; EC50 – GUSAR prediction result; AD – Applicability Domain estimation.

If prediction was made using more than one model then output data can be represented in two formats: Full and Short. Each type is related to different representation of applicability domain record. Full or Short type of representation of applicability domain record can be selected in the **Options of Models** window using **Model Output** tab.

The example of GUSAR prediction results after selection Full output data of Applicability Domain Record:

GUSAR - Prediction of Values for Substances
Copyright (C) 2013 V. Poroikov, D. Filimonov
& Associates

Chemical Structure File: algae.sdf

Table of 1 predictable activities.

ID	EC50	AD
1	-1.0549	in AD
2	-1.2095	in AD
3	-1.4042	in AD
4	-0.9528	in AD
5	-1.1643	out of AD
6	-0.9839	in AD
7	-1.1103	in AD
8	-1.1875	in AD
9	-1.185	out of AD
10	-0.5627	in AD
11	-0.6899	in AD
12	-1.1418	in AD
13	-0.44	in AD
14	-0.3506	in AD

Where, ID – structure identifier selected by the user; EC50 – GUSAR prediction result; AD – Applicability Domain estimation.

The example of GUSAR prediction results after selection Short output data of Applicability Domain Record:

ID	EC50
1	-1.0549
2	-1.2095
3	-1.4042
4	-0.9528
5	out of AD
6	-0.9839
7	-1.1103
8	-1.1875
9	out of AD
10	-0.5627
11	-0.6899
12	-1.1418
13	-0.44
14	-0.3506

Where, ID – structure identifier selected by the user; EC50 – GUSAR prediction result.

Short format of output data of applicability domain record in CSV file can be used for further processing of prediction results by other programs (KNIME, etc.).

Note!

If prediction was carried out for several activities, the second and third columns will be repeated for each additional activity with the appropriate names.

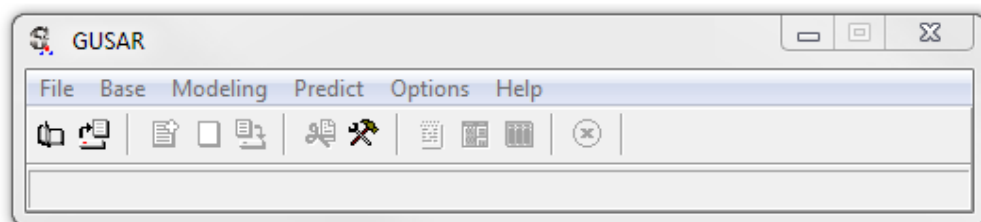
CHAPTER 4




GETTING STARTED WITH GUSAR


GETTING STARTED WITH GUSAR INTERFACE

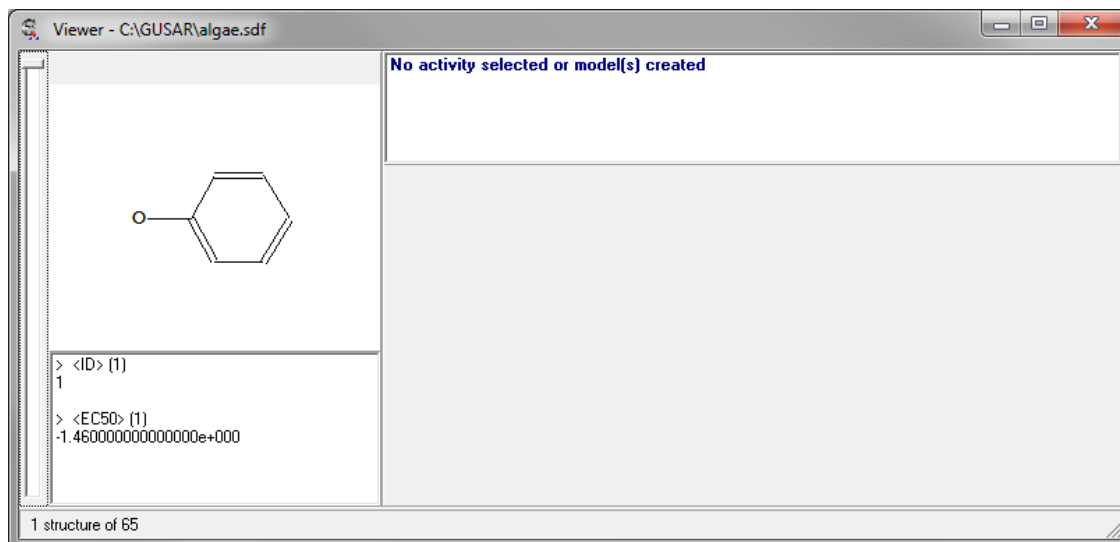
To start **GUSAR** - double-click GUSAR icon (shortcut );
or run **GUSAR.exe** from GUSAR folder.


The **Main** window of GUSAR interface appears and Descriptor Base loading is started.
After loading of Descriptor Base, the main window of GUSAR interface looks like this:

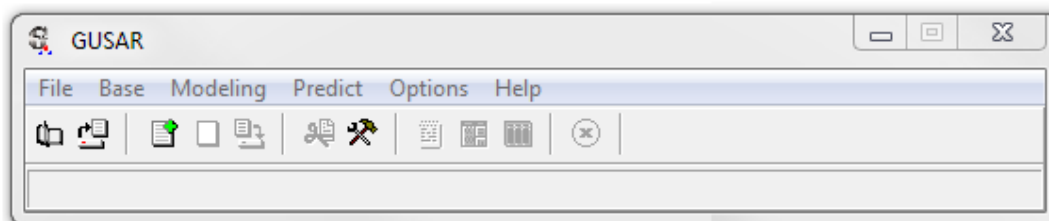


Three buttons are active by default. By these buttons you can load SD file with quantitative or qualitative data on the appropriate endpoint as a basis for creation of QSAR model (use the button  or the command "**File|Open SDF**"), load the previously GUSAR created QSAR model(s) (use the button  or the command "**Base|Open Base**") and determine parameters of QSAR modelling and/or prediction (use the button  or the command "**Options|Options of Models**").

After pressing the button  or running the command "**File|Open SDF**", the window with structures and text data from the loaded SD file will appear in **Viewer** window:




When SD files with data on chemical structure and endpoint data have been loaded, the command "**File|Add Data**" and the appropriate button  become available.




Use this command to start the calculation of descriptors for compounds from the loaded SD file and add them to SAR base. After application of this command, "**Select Field Names**" window will appear:


The window allows you to select a field for identification of structures ("**Field Name to Identification of Substances**"), a field with quantitative data of modelled endpoint ("**Input Field Name of Activity**"), a type of QSAR model (**Model Type**) (Continuous (i.e. quantitative) or Category (i.e. qualitative, classification) and determine the output name of the future QSAR model related with the modelled activity ("**Output Field Name of Activity**"). You may type the output name of activity different from those defined in the input field.

You also should determine if the values of modelled endpoint will be used "as is" or they will be transformed to decimal logarithm (choose **Log10(Value)** in "Use" field). By our opinion **Log10(Value)** in "Use" field should be used if the values of the modelled activity vary more than two Log10 units.

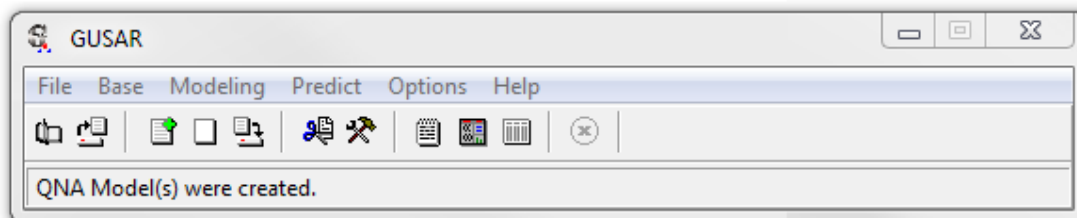
Press **Ok** button to start the calculation of descriptors for compounds from the loaded SD file and add them to SAR base. When this process will be finished, the button  and two commands: **Modeling|Create QNA Models** and **Modeling|Create MNA Models** will become available.







- Use the button  to clear SAR base (Delete all earlier created models and the appropriate training sets).
- Use the commands **Modeling|Create QNA Models** and **Modeling|Create MNA Models** to create QSAR models on the basis of QNA and MNA descriptors, respectively.

Note!

By default a single QSAR model is created. If you wish to create a consensus QSAR model, you should change the options of model before the modeling. The options of model are available by clicking on the button  or the command "**Options|Options of Models**" (Look Chapter 6 for details).

After creating of QSAR model, all buttons become active.




- Use the button  or the command "**Base|Save Base**" to save SAR Base with the created QSAR model(s);
- Use the button  or the command "**Base|Selection**" to select activities and appropriate earlier created QSAR models;
- Use the button  or the command "**Predict|Prediction to TXT**" to make a prediction and save the results in a special SD file format for a test set (represented by SD file) by the selected QSAR models;
- Use the button  or the command "**Predict|Prediction to SDF**" to make a prediction and save the results in SD file format for a test set (represented by SD file) by the selected QSAR models;
- Use the button  or the command "**Predict|Prediction to CSV**" to make a prediction and save the results in CSV format for a test set (represented by SD file) by the selected QSAR models;
- Press the button  to stop any process.


If they were saved as SDfile, GUSAR prediction results can be imported into the ISIS/Base or into some other chemical databases.

WORKFLOW OF QSAR MODEL(S) CREATION AND EXTERNAL TEST PREDICTION




Before the start of QSAR model creation you should make several procedures:

1. Load SD file with data about structures of chemicals and experimental data by **OPENING SDFILE** procedure (see page 31);
2. Add data from the loaded SD file to SAR Base by **ADD DATA TO SAR BASE** procedure (see page 32);
3. Determine parameters of creating QSAR model(s) by the command **Options|Options of Models** or the button  (optionally) (see page 51) or use parameters determined by default;

Now you are ready to create QSAR model(s):

4. Create QSAR model(s) on the basis of QNA and/or MNA descriptors by the commands **"Modeling|Create QNA Models"** (see page 44) and/or **"Modeling|Create MNA Models"** (see page 45);
5. Select the most prospective QSAR models by the use of the button  or the command **"Base|Selection"**.

Now you are ready to make prediction for the external test set on the basis of created and selected QSAR model(s):

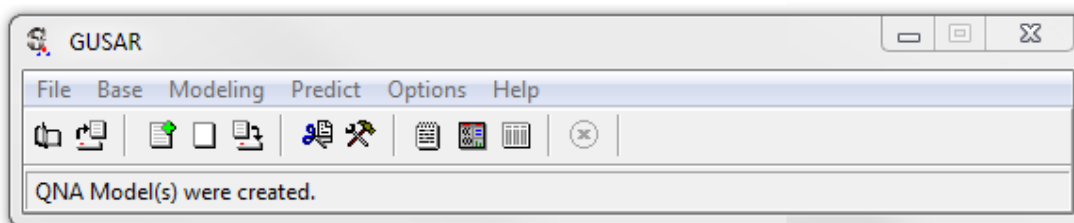
6. Use **"Predict|Prediction to TXT"**, **"Predict|Prediction to SDF"** and **"Predict|Prediction to CSV"** commands or appropriate buttons (, , ) to predict and save prediction results in special formats (see **PREDICTION RESULTS**, page 12).

CHAPTER 5

GUSAR INTERFACE and FUNCTIONS

MAIN WINDOW

The main window contains menu, speed buttons, progress bar and status panel.



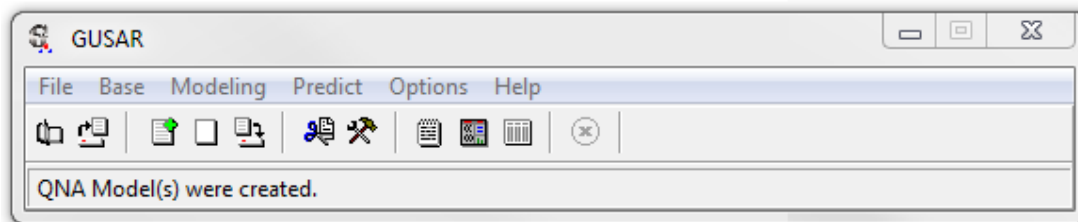
You can use either **menu commands** or the speed buttons to execute **GUSAR** system procedures.

The main menu items are **File, Base, Modeling, Predict, Options** and **Help**. All these item menus are described in more detail below.












Use the mouse, the keyboard (**F10**) or key combinations: **Alt+F, Alt+B, Alt+M, Alt+O, Alt+S, Alt+H** to choose the particular menu item.

- The speed buttons' hints will appear when the mouse cursor points to the button.
- Use Alt+F4 shortcut, **File|Exit (Alt+X)** menu command or button **X** in upper right corner of GUSAR main window to quit GUSAR.

An existing Descriptor Base is loaded by default. After the creation of any QSAR model(s) or loading of the existing QSAR models by the command "**Base|Open Base**", the most of the commands become available (see the Figure below).



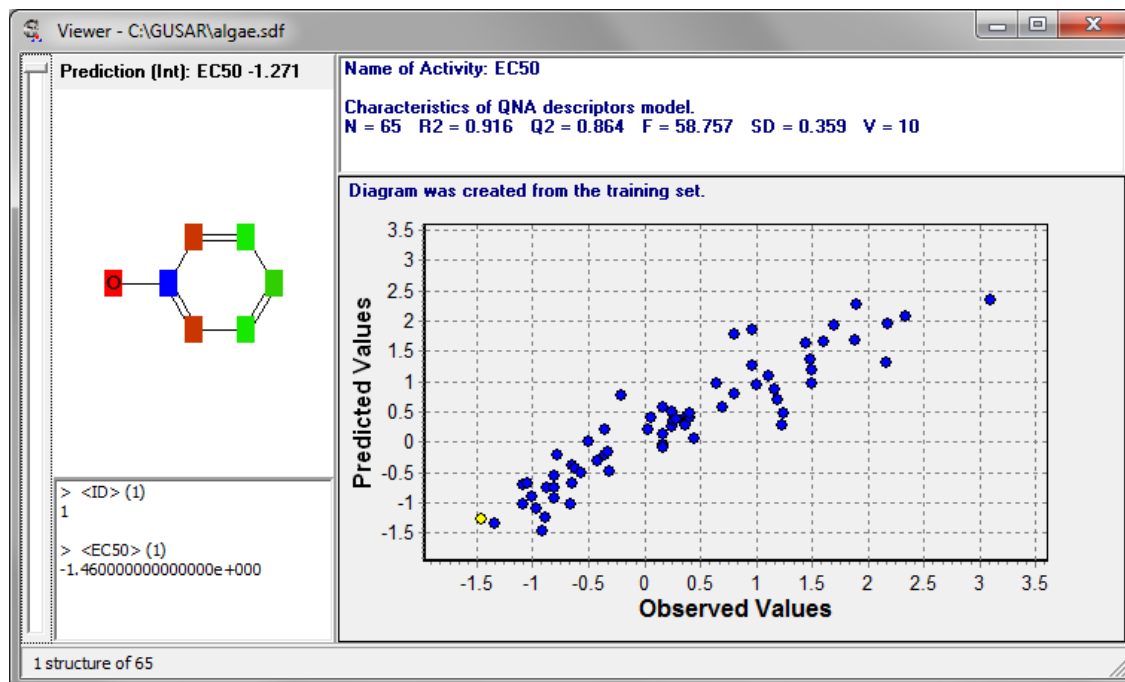
The buttons have the following functions:

-  - Opens SDfile.
-  - Opens SAR base with earlier created QSAR model(s).
-  - Calculates descriptors for compounds from the loaded SD file and add them to SAR base.
-  - Deletes all earlier created models and appropriate descriptors from SAR base.
-  - Saves SAR Base with the created QSAR model(s).
-  - Shows the existed QSAR models and allows to select the appropriate earlier created QSAR models.
-  - Determines parameters of creating QSAR model(s).
-  - Predicts and saves the structure the prediction results in a special format as an SDfile.
-  - Predicts and saves the structure and prediction results as an SDfile.
-  - Predicts and saves the prediction results as CSV file.
-  - Interrupts any current process (reading from file, saving into file, training and prediction).

VIEWER WINDOW

Viewer window displays the structures and data from the loaded SD file (left side) and information about QSAR model(s) for the selected activity (right side).

The results of QSAR modeling are represented in two views depending on the type of QSAR model (Continuous or Category). The window for continuous model is the following:



The top left side of the window displays the prediction result for each structure from the loaded SD file on the basis of the selected QSAR model(s) (EC50 value in the picture).

The top right side of the window displays the characteristics of the created model(s):

N - number of compounds in the training set;

R² - average R² of the models calculated for the appropriate training set;

Q² - average Q² of the models calculated for the appropriate training set;

F - Fisher coefficient;

SD - Standard deviation;

V - number of independent variables in the model.


QNA or MNA type of descriptors is displayed for single model.

If there were several training sets with the data for the same activity and several models were created, the characteristics would be displayed as average values. For example:

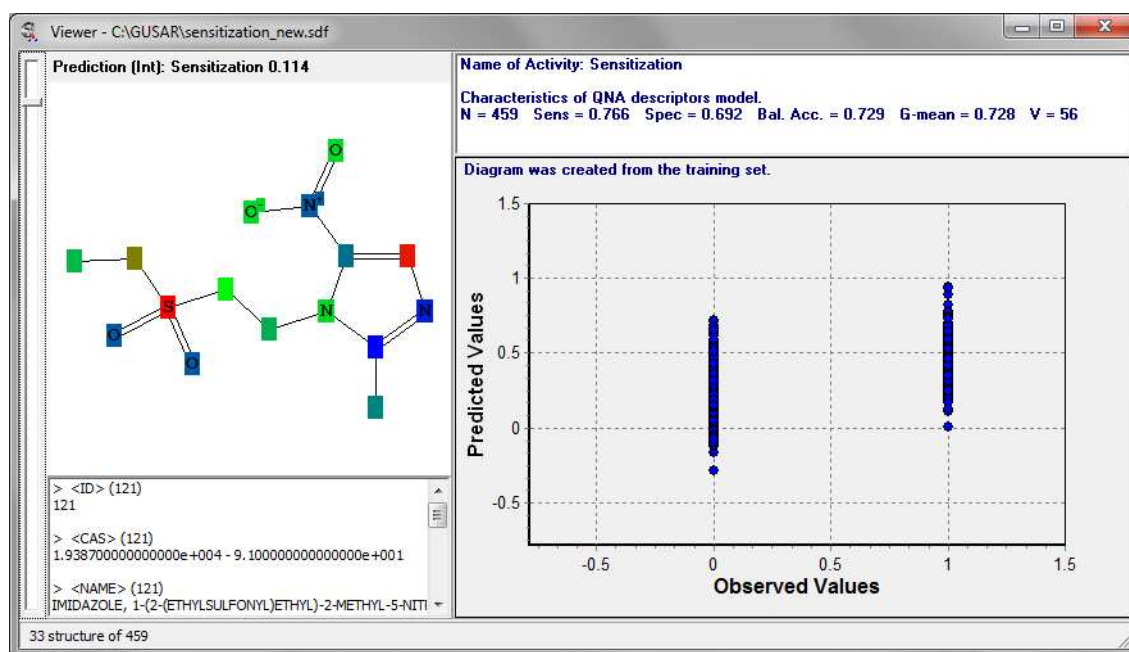
Name of Activity: Mouse SC LD50 log10(mmol/kg)

Characteristics of Consensus model from 2 models.
N = 1072-2500 R² m = 0.578 F m = 13.005SD m = 0.597 Q² m = 0.480 V m = 262

Note!

Use **Services|Selection** command or press the button  to select the desirable activity and models.

For classification QSAR model the result of modelling is represented at the following figure:



The points at 0 and 1 of observed value mean inactives and actives, respectively. The threshold between actives and inactives is calculated by adjusting decision threshold approach (see Chapter 7) during leave-one-out cross validation procedure. Sensitivity (Sens), Specificity (Spec), Balanced Accuracy (Bal. Acc.) and Geometric mean (G-mean) are characteristics of accuracy of the created model calculated by leave-one-out cross validation procedure. The following formulae are used for calculation of these characteristics:

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{FP}+\text{TN})$$

$$\text{Balanced Accuracy} = (\text{Sensitivity}+\text{Specificity})/2$$

Geometric mean = $(\text{Sensitivity} * \text{Specificity})^{1/2}$

where TP – true positive, TN – true negative, FP – false positive, FN – false negative.

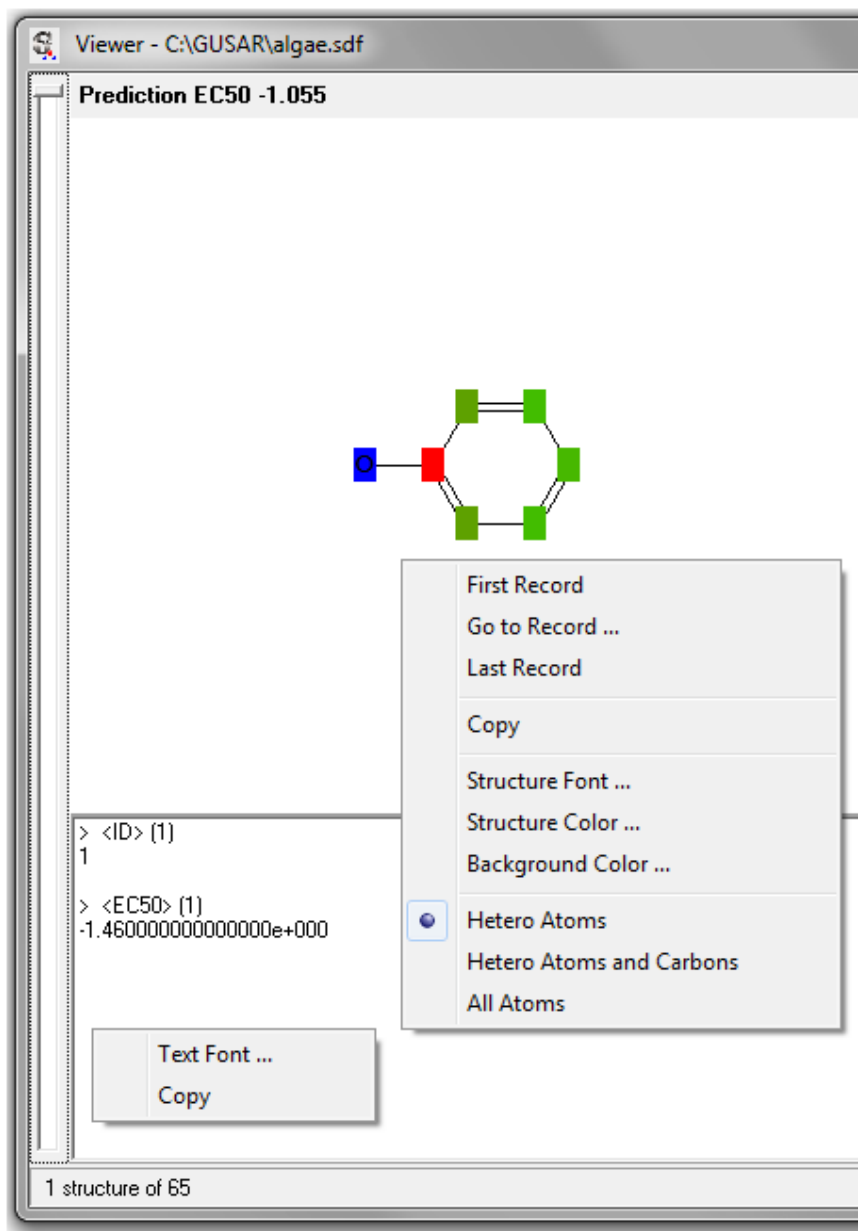
If QSAR model(s) was(were) created on the basis of QNA descriptors the contribution of each atom into the predicted value is displayed for a studied compound. At the user's display these contributions are reflected by different colors:

Green – no significant contribution;

Blue – decrease in the activity;

Red – increase in the activity.

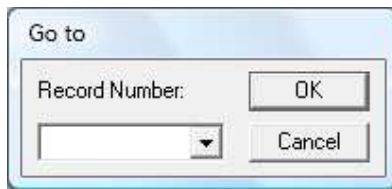
The left part of the window contains a scroll-bar for navigation on the loaded SD file and two pop-up menus that are activated by the clicking of the right mouse button on the appropriate fields (the structure field and the text field).



The bottom part of the window displays a number of a current record and total number of records in the loaded SD file.

The structure field pop-up menu contains the following command:

- Choose **First Record** command to go to the first record.
- Choose **Go to Record** command to go to the chosen record. The **Go to** dialog box will appear. You should type the Record Number and press **OK** to go to this record.

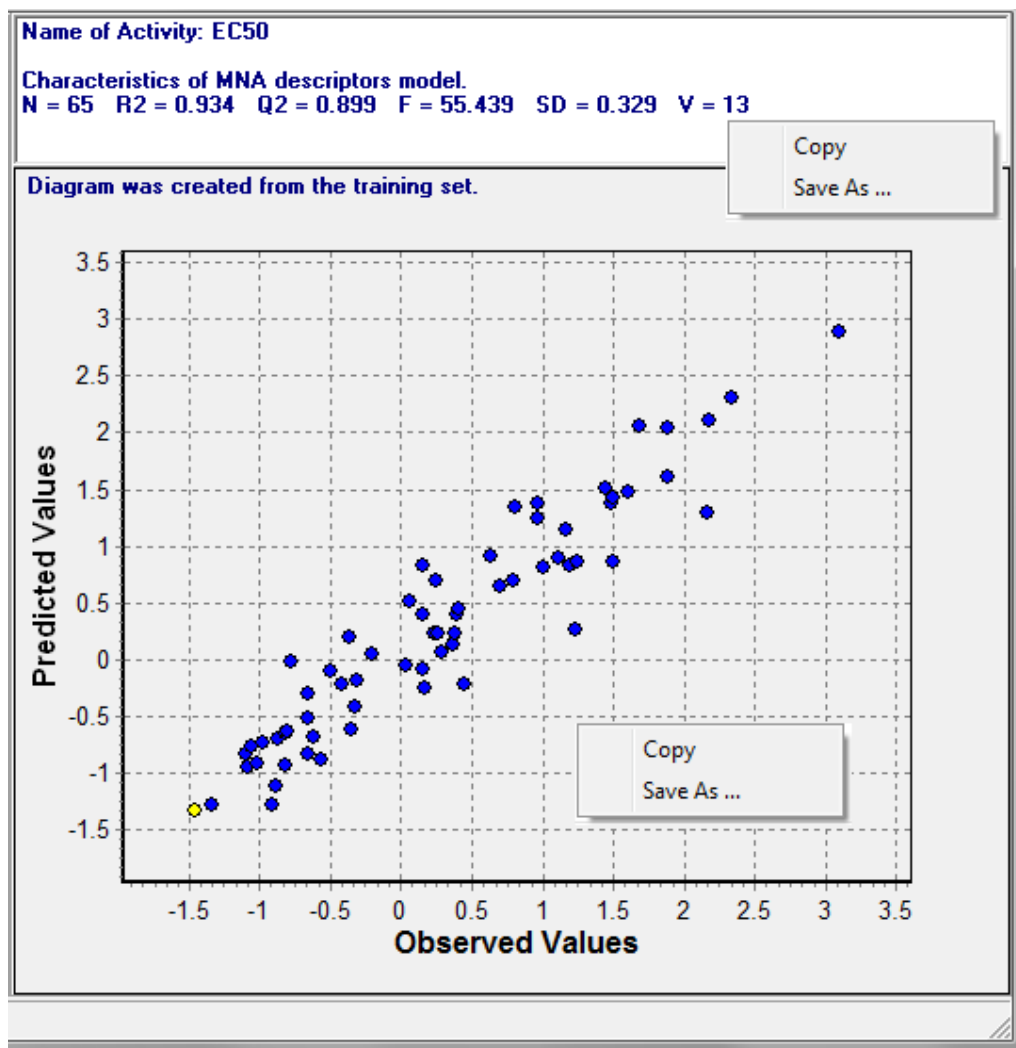


- Choose **Last Record** command to go to the last record.
- Choose **Copy** command to copy the current structure to the clipboard. It may be pasted to any Microsoft Word document as a bitmap image.
- Choose **Structure Font ...** command to change font of the atom symbols.
- Choose **Structure Color ...** command to change color of the structure's bonds.
- Choose **Background Color ...** command to change color of the background.
- Choose **Hetero Atoms**, **Hetero Atoms and carbons** or **All Atoms** options to modify presentation of structure formula.

The text field pop-up menu contains the following command:

- **Text Font ...** command changes font for text information.
- **Copy** command copies a text from the text field to the clipboard.

The right side of **Viewer** window contains the text field with description of QSAR model(s) (top part) and a diagram of correlation between the observed and predicted values for compounds from the training set of the model (yellow point shows the values for a current structure). The right side also contains two pop-up menus.



The text field pop-up menu contains the following command:

- **Copy** command copies a text from the text field to the clipboard.
- **Save as ...** command save a text from the text field to the TXT file.


The diagram pop-up menu contains the following command:

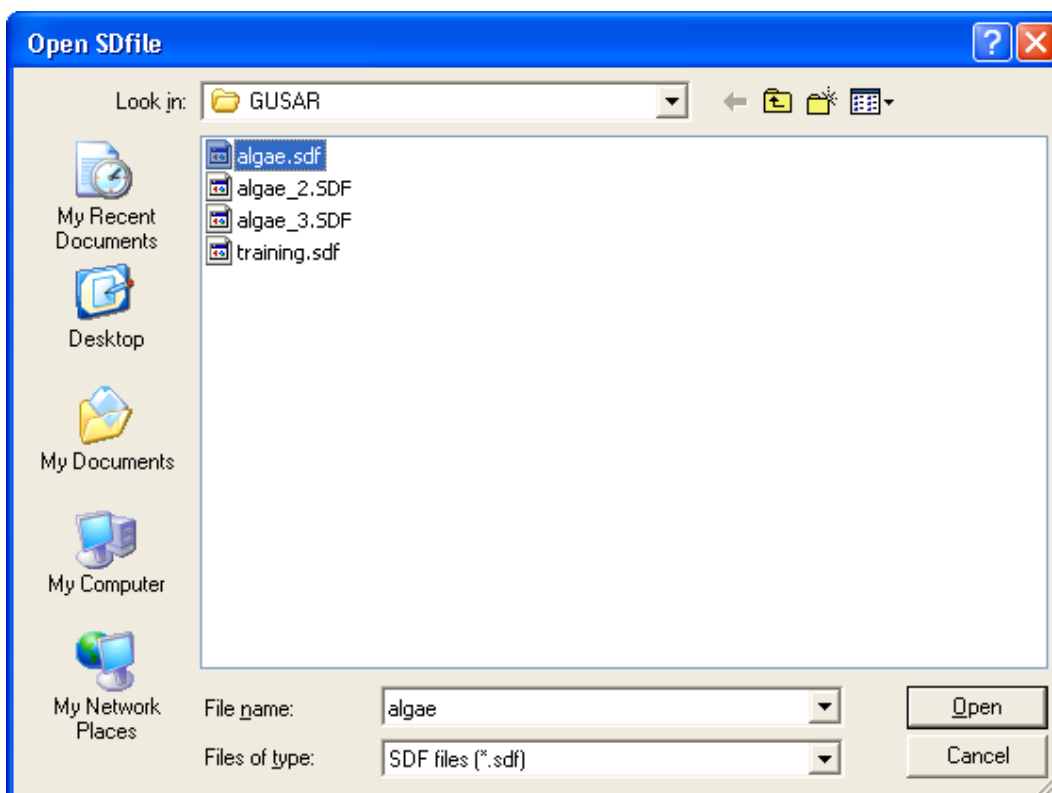
- **Copy** command copies diagram to the clipboard. It may be pasted to any Microsoft Word, Excel or PowerPoint document as a bitmap.
- **Save as ...** command save a bitmap to the .bmp file.

The text field with description of QSAR model(s) contains information about the name of selected activity (EC50 in the last picture) and characteristics of the model or consensus models:

- N** – the number of compounds in the training set(s);
- R2** – the square of the regression coefficient;
- Q2** – the cross-validated R^2 ;
- F** – the value of Fisher's statistics;
- SD** – the standard deviation;
- V** – the number of variables in the final regression equation.

OPENING SDFILE

Use **File|Open SDF** menu command or the button  to open SDfile (*.sdf) with structures of chemical compound and quantitative or qualitative data on activity (property). The Open dialog box appears:




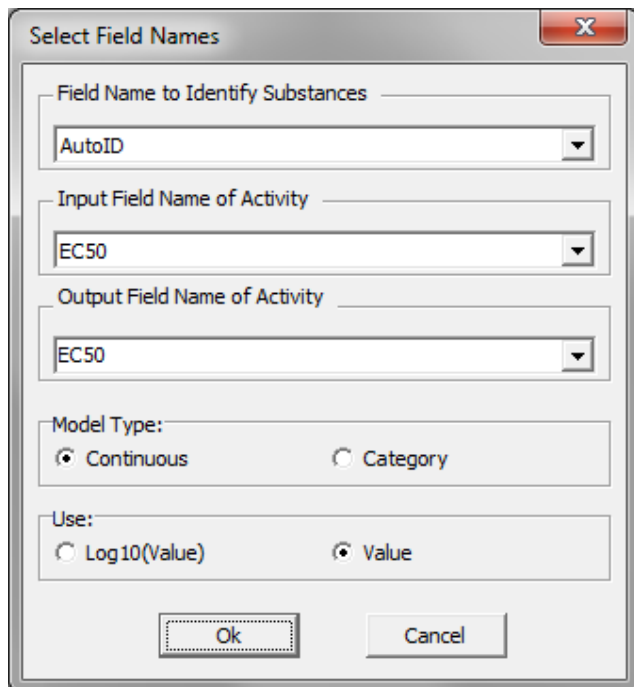
The **Viewer** window with structures and text data from the loaded SD file will appear. Its description is given below.

Note!

You can add several SDfiles to create QSAR model(s) for the same activity. For that you should use **File|Open SDF** and **File|Add Data** menu commands sequentially several times before the creation of QSAR model(s).

ADD DATA TO SAR BASE

Use the command "**File|Add Data**" or the appropriate button  to start the calculation of QNA and MNA descriptors for compounds from the loaded SD file and add them to SAR base together with data about the modelled activity. After application of this command, "**Select Field Names**" window will appear:

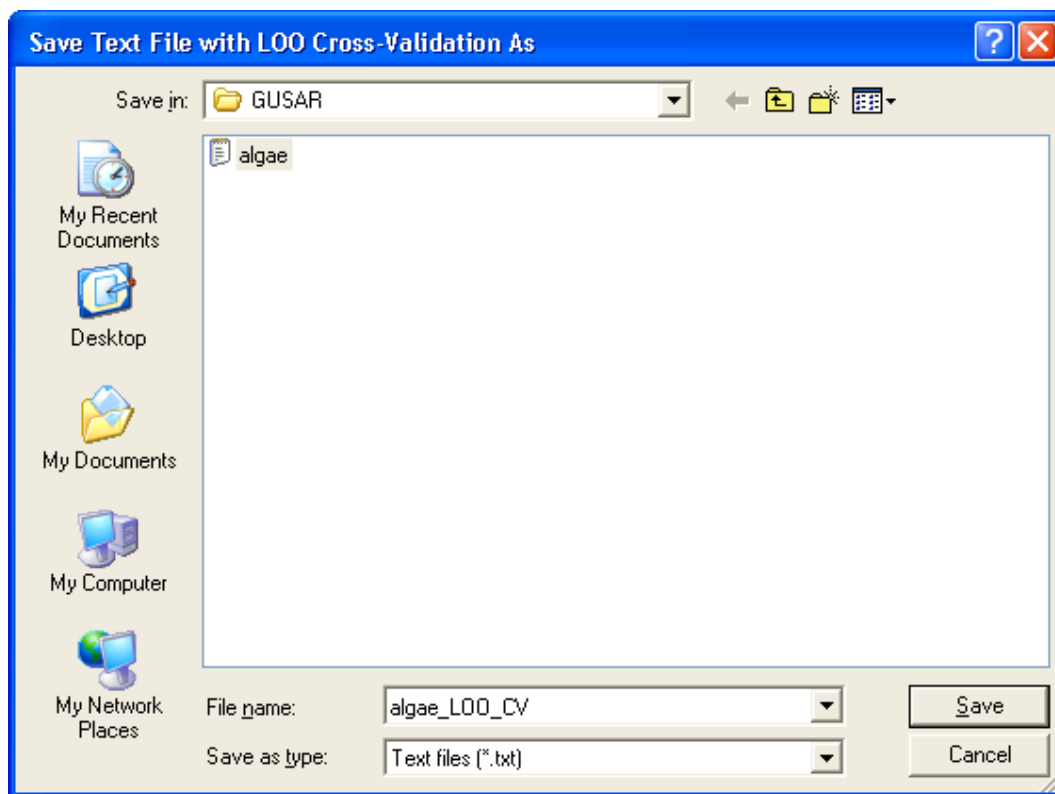


The window allows you to select a field for identification of structures ("**Field Name to Identify Substances**"), a field with quantitative data of modelled endpoint ("**Input Field Name of Activity**"), a type of QSAR model (**Model Type**) (Continuous (i.e. quantitative) or Category (i.e. qualitative, classification) and determine the output name of the future QSAR model related with the modelled activity ("**Output Field Name of Activity**"). You may type the output name of activity different from those appeared in the input field.


You should also determine if the modelled values will be used as is or they will be transformed to decimal logarithm (choose **Log10(Value)** in "**Calculate**" field). Press **Ok** button to start the calculation of descriptors for compounds from the loaded SD file and add them to SAR base. **Log10(Value)** in "**Calculate**" field should be used if the values of the modelled activity vary more than two Log10 units.

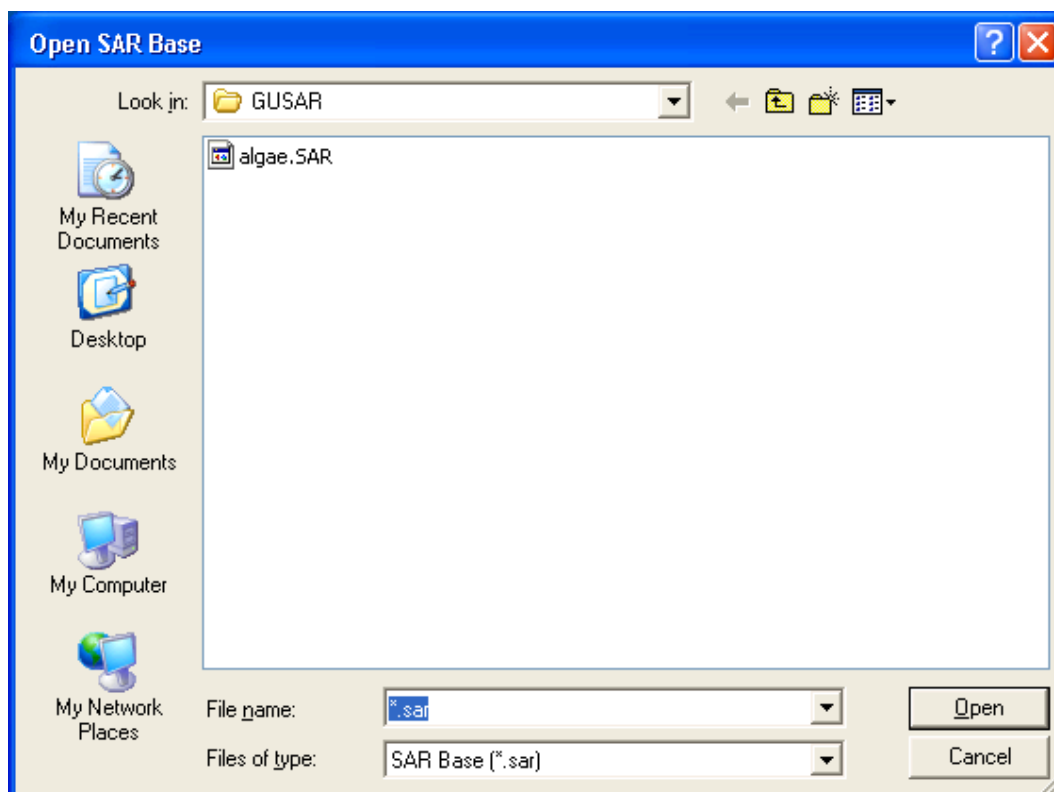
SAVE RESULTS OF LOO CV

Use **File|Save results of LOO** menu command to save the results of leave-one-out cross-validation procedure in a TXT file. This procedure is executed during the creation of QSAR model(s) at calculation of Q^2 . The window for determination of the name of the file with Leave-one-out cross-validation procedure results will appear:




OPENING SAR BASE


Use **Base|Open Base** menu command or press the button  to open the existing SAR Base with the earlier created QSAR models. The following window appears.




Note!

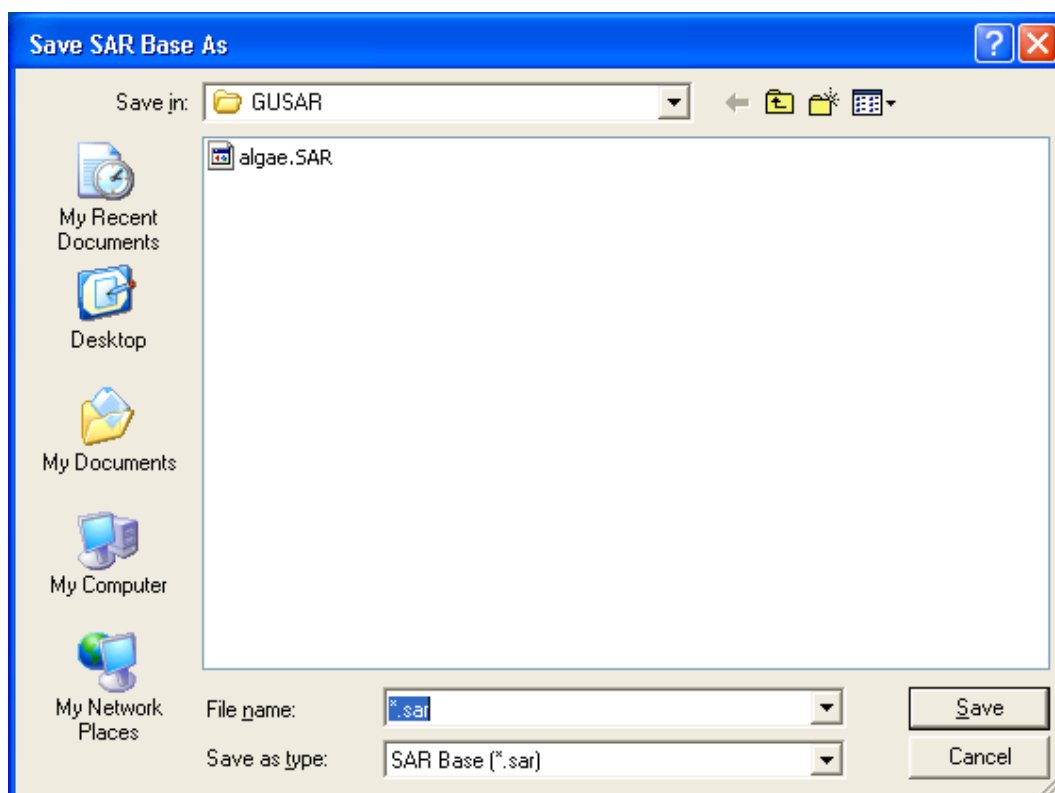
After the SAR Base loading use **Base|Selection** or press the button  to look at the loaded QSAR model(s) and select those that you need.

DELETE DATA FROM SAR BASE


Use the command "**Base|Clear Base**" or the appropriate button  to delete all information from the current SAR Base. After application of this command you can use **ADD DATA TO SAR BASE** procedure to start the creation of new QSAR model(s).

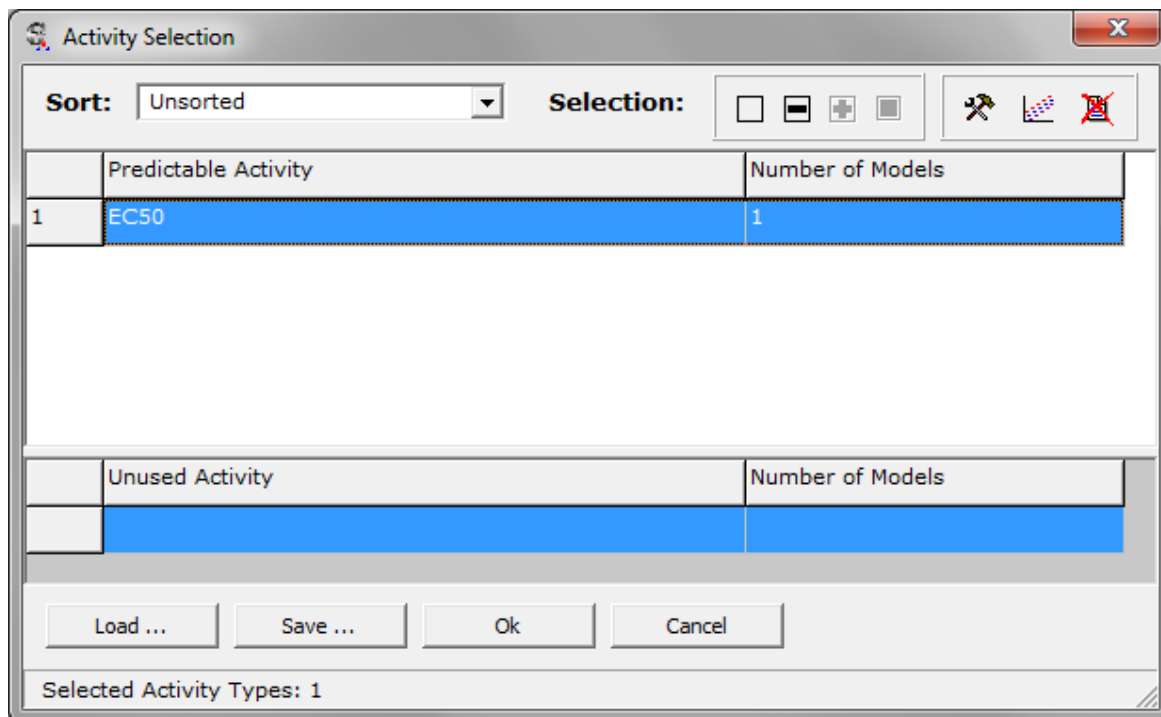
SAVE SAR BASE


Use the command "**Base|Save Base**" or the appropriate button  to save the current SAR Base with the created QSAR model(s). After application of this command you can load this SAR base at any time using the **OPENING SAR BASE** procedure.






ACTIVITY SELECTION WINDOW

Use **Base|Selection** menu command or press the button  to open the window for selection of desirable activities and QSAR models.



If you want to exclude some activity(s) from the list of predictable activities you should mark those activity(s) at the **Predictable Activity** column (use mouse or/and cursor control keys and **Shift**) and click  button. All these activities will be transferred to the **Unused Activity** table.

If you want to add some activity(s) to the list of predictable activities from the **Unused Activity** column you should mark those activity(s) at the **Unused Activity** column (use mouse or/and cursor control keys and **Shift**) and click  button. All these activities will be transferred to the **Predictable Activity** table.

You may also select all activity types by clicking  button, or cancel selection of all activity types by clicking  button.

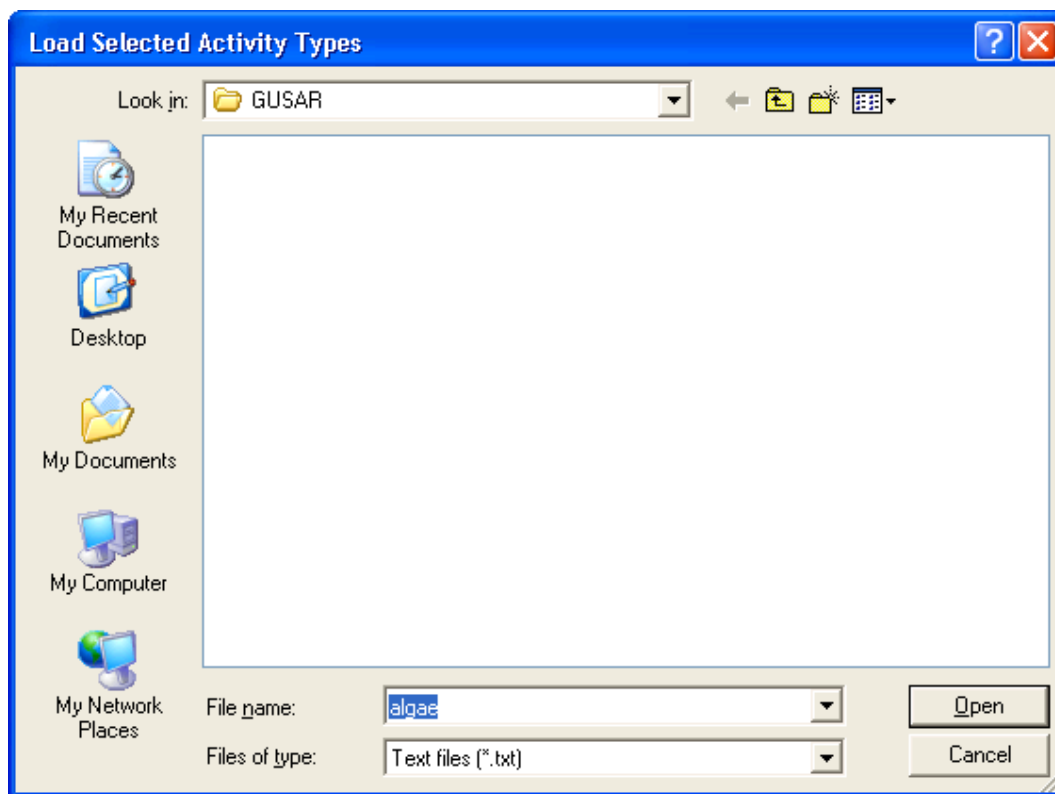
Note!

All predictable activities will be predicted for the analysed compounds.

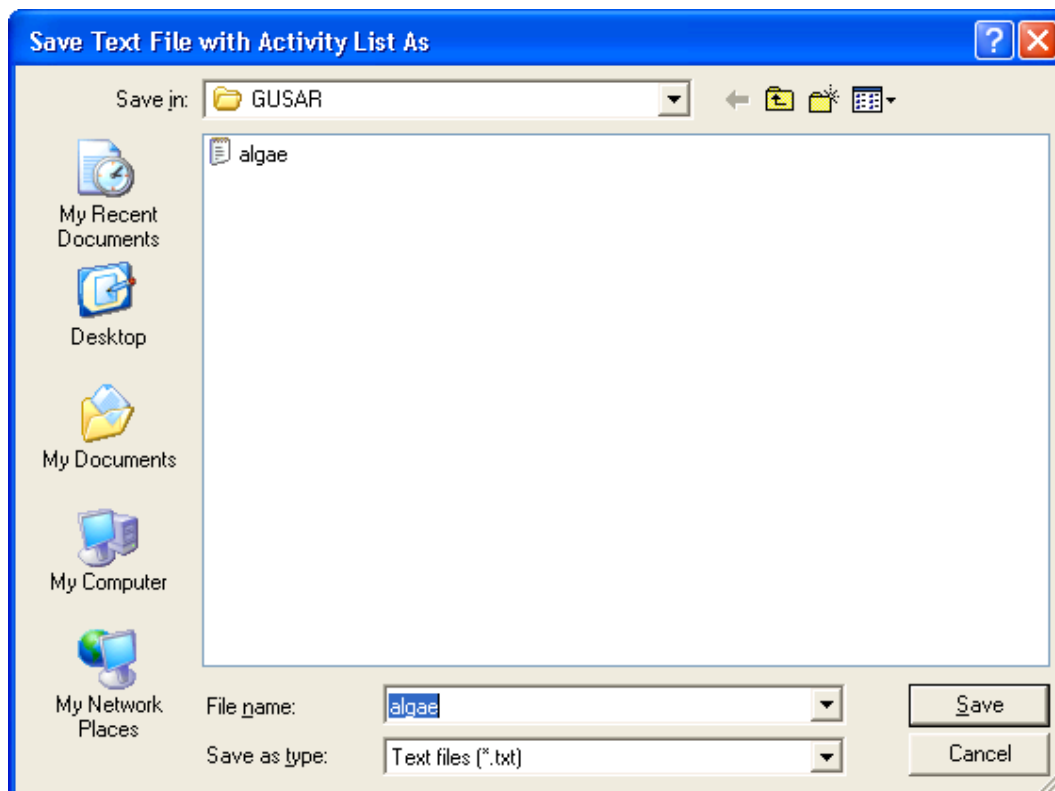
You may sort activity names for more convenience according to your criterion. It will help you to make up selection promptly. The names of activity types may be sorted in several ways by "**Sort**" drop-down menu command:

Sort condition	Description
Activities ascending	Alphabetically
Activities descending	Alphabetically (reverse order)
Number ascending	Ascending order of number of QSAR models for the activity.
Number descending	Descending order of number of QSAR models for the activity




- Click **Load...** button to load one of your text file (*.txt) with a particular version of the selected activity list. Only those activity types, which names are coincided with activity names in SAR Base, will be selected.



- Click **Save...** button to save the list of selected activity types in text file (*.txt).



Activity Selection window contains three additional buttons with the following functions:

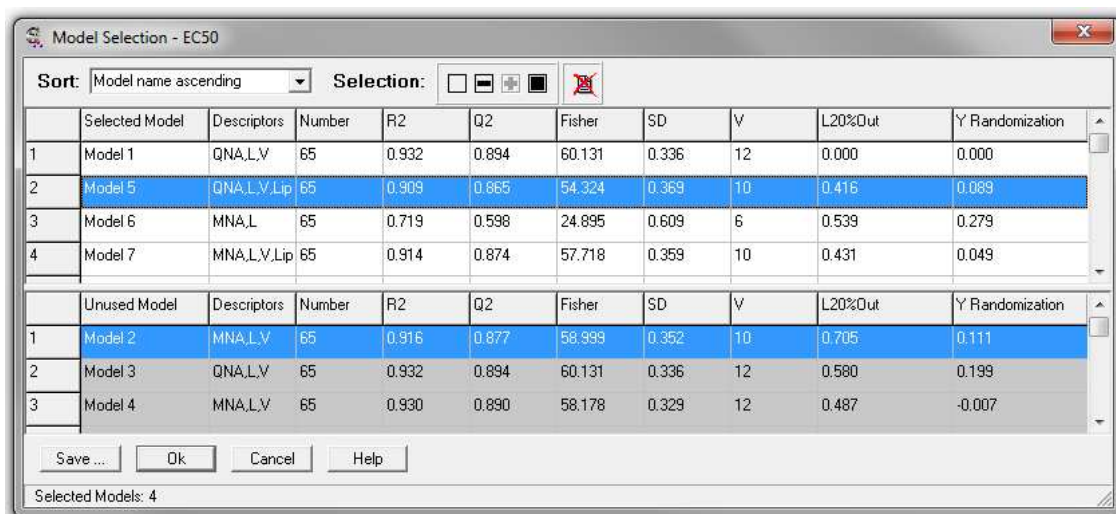
-  - press on this button displays the options of QSAR models for the selected activity by the **Options of Models** window. You may change parameters of applicability domain and the use of nearest neighbor's data during the prediction of activity for external test set. Each activity may be predicted using different options if **Predictable activity** table contains several activities. This function is not available for the unused activities.
-  - press on this button displays the results of consensus modeling in **Viewer** window for the first selected activity. This function is not available for the unused activities.
-  - press on this button deletes all unused activities with the appropriate QSAR model(s) from SAR base.

Note!

Double click on the selected activity opens **Model Selection** window. This window allows selecting the desirable QSAR model(s) that will be used for prediction of the selected activity.

MODEL SELECTION WINDOW


Double click on the selected activity in **Activity Selection** window, to activate **Model Selection** window for selection of the desirable QSAR model(s), which will be used for prediction of the selected activity.






The window contains two tables with the same columns. **Selected Model** table contains the model(s) that will be used for prediction of the selected activity. **Unused Model** table contains the model(s) that will not be used for prediction of the selected activity. Both tables contain the columns with the same titles:

- Descriptors** – a type of descriptors that were used for creation of the model (L – topological length of molecules, V – value of molecules and PhysChem – physico-chemical descriptors). The algorithms of calculation of these descriptors are described on page 57.
- Number** – number of structures in the training set that were used for creation of the model.
- R2** – the square of the regression coefficient.
- Q2** – the cross-validated R^2 .
- Fisher** – the value of Fisher's statistics.
- SD** – the standard deviation.

- V** – the number of variables in the final regression equation.
- LMany%Out** – the results of leave-many-out cross-validation procedure. It shows an average R^2 value for test sets predicted by QSAR models created on the basis of appropriate training sets.
- Y Randomization** – an average Q^2 value of QSAR models created after Y randomization.

If you want to exclude some model(s) from the list of selected models you should mark those model(s) at the **Selected Model** column (use mouse or/and cursor control keys and **Shift**) and click  button. All these models will be transferred to the **Unused Model** table.

If you want to add some model(s) to the list of unused models from the **Unused Model** column you should mark those model(s) at the **Unused Model** column (use mouse or/and cursor control keys and **Shift**) and click  button. All these models will be transferred to the **Selected Model** table.

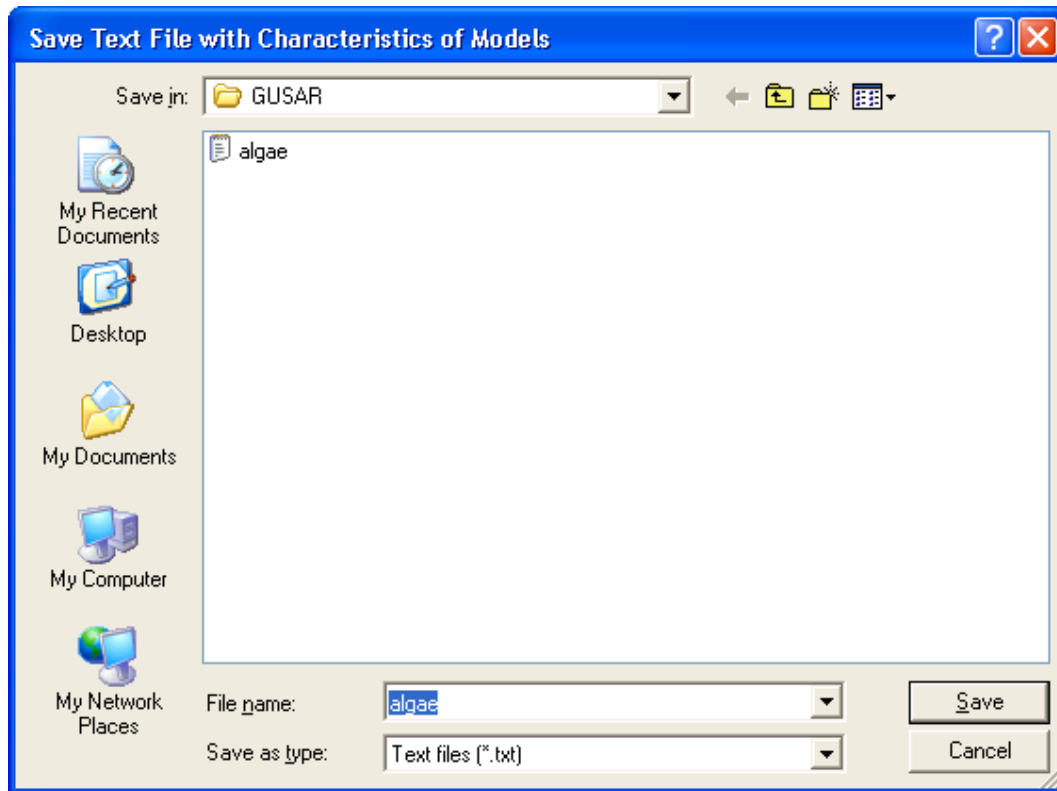
You may also select all models by clicking  button, or cancel selection of all models by clicking  button.

For more convenience you may sort models according to your criterion. It will help you to make up selection promptly. The models may be sorted in several ways by “**Sort**” drop-down menu command:

Sort condition	Description
Model name ascending	Alphabetically
Model name descending	Alphabetically (reverse order)
Number ascending	Ascending order of number of compounds in training set of the model.
Number descending	Descending order of number of compounds in training set of the model.
R2 ascending	Ascending order of R^2 of the model.
R2 descending	Descending order of R^2 of the model.
Q2 ascending	Ascending order of Q^2 of the model.
Q2 descending	Descending order of Q^2 of the model.

Fisher ascending	Ascending order of the value of Fisher's statistics of the model.
Fisher descending	Descending order of the value of Fisher's statistics of the model.
SD ascending	Ascending order of the standard deviation of the model.
SD descending	Descending order of the standard deviation of the model.
V ascending	Ascending order of the number of variables of the model.
V descending	Descending order of the number of variables of the model.
LMany%Out ascending	Ascending order of LMany%Out of the model.
LMany%Out descending	Descending order of LMany%Out of the model.
Y Randomization ascending	Ascending order of Y Randomization of the model.
Y Randomization descending	Descending order of Y Randomization of the model.

- Click **Save...** button to save the list with characteristics of selected models in the text file (*.txt).



Activity Selection window contains one additional button with the follow function:



- press on this button deletes all unused QSAR model(s) from SAR base.

CREATE QNA MODELS

Use the command "**Modeling|Create QNA Models**" to create QSAR model(s) on the basis of QNA descriptors generated during **ADD DATA TO SAR BASE** procedure. When creation of QSAR model(s) will be finished, the diagram with the correlation between the predicted and observed values will appear in the **Viewer** window.




CREATE MNA MODELS

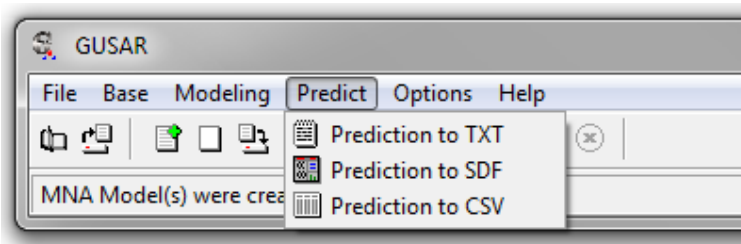
Use the command "**Modeling|Create MNA Models**" to create QSAR model(s) on the basis of MNA descriptors generated during **ADD DATA TO SAR BASE** procedure. When creation of QSAR model(s) will be finished, the diagram with the correlation between the predicted and observed values will appear in the **Viewer** window.

CREATE COMBINATORIAL MODELS

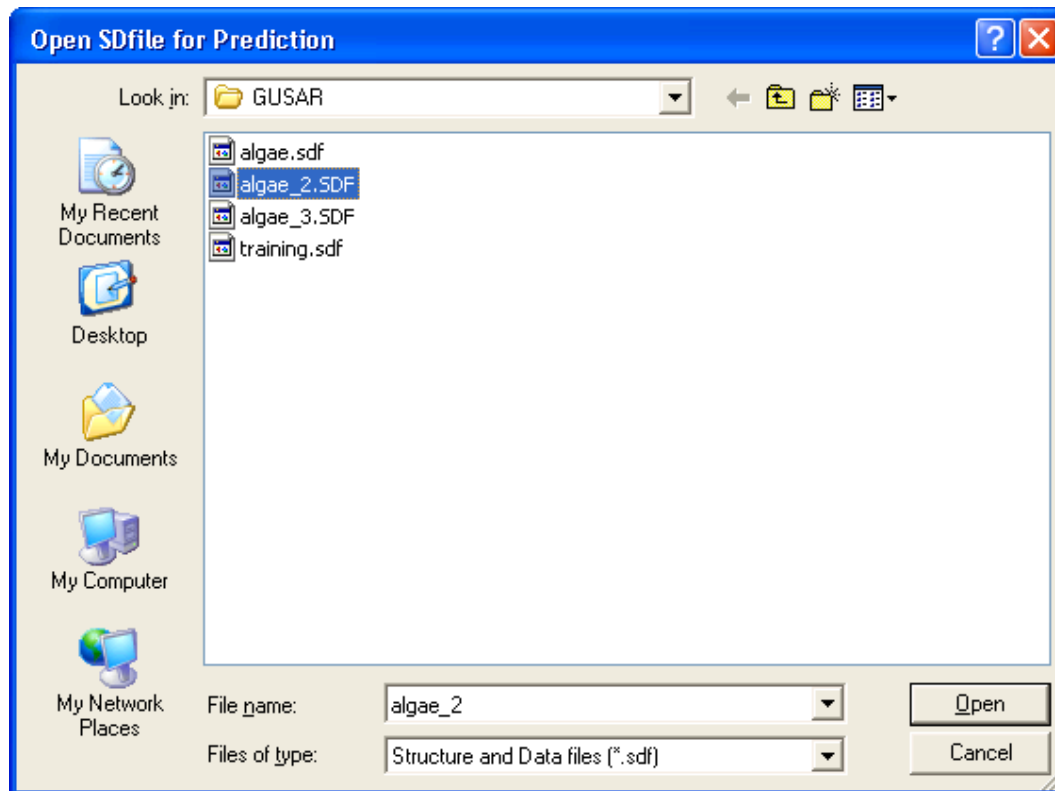
Use the command "**Modeling|Create Combinatorial Models**" to create QSAR models automatically on the basis of different combinations between QNA, MNA descriptors and additional variables (Length, Volume, PhysChem). When creation of QSAR models will be finished, the diagram with the correlation between the predicted and observed values will appear in the **Viewer** window.

PREDICTION BY GUSAR

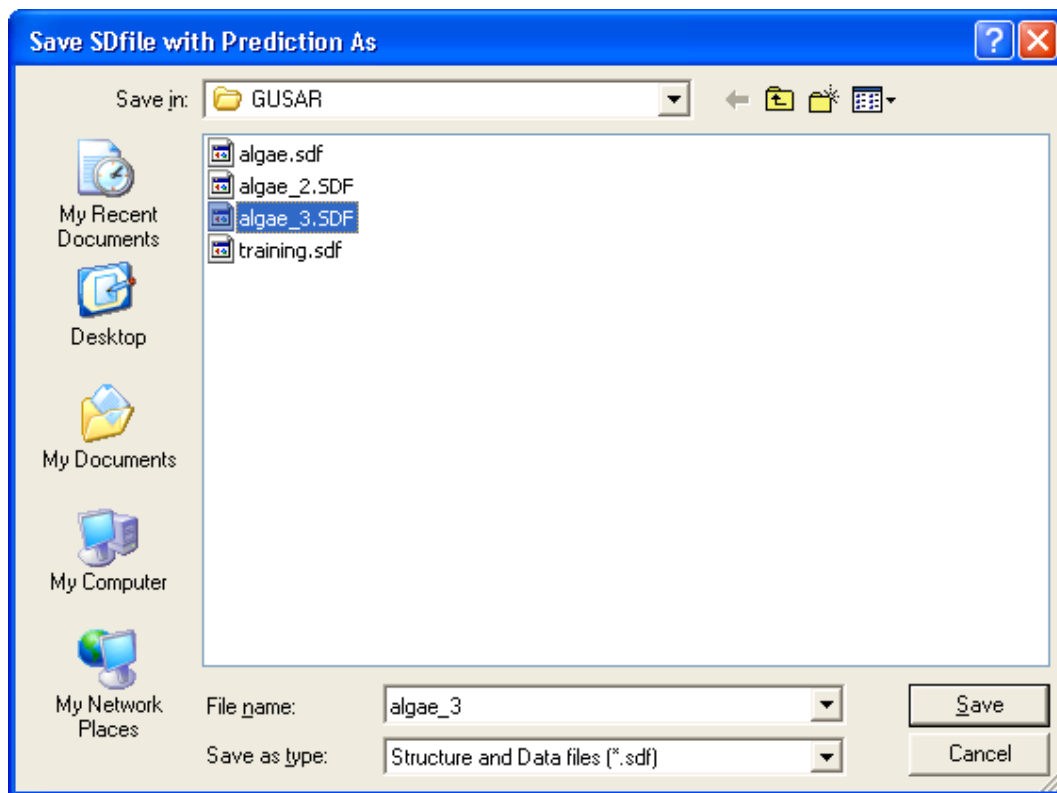
Use “**Predict|Prediction to TXT**”, “**Predict|Prediction to SDF**” and “**Predict|Prediction to CSV**” commands or appropriate buttons (, , ) to make prediction and save the prediction results in special formats (see **PREDICTION RESULTS**, page 12).



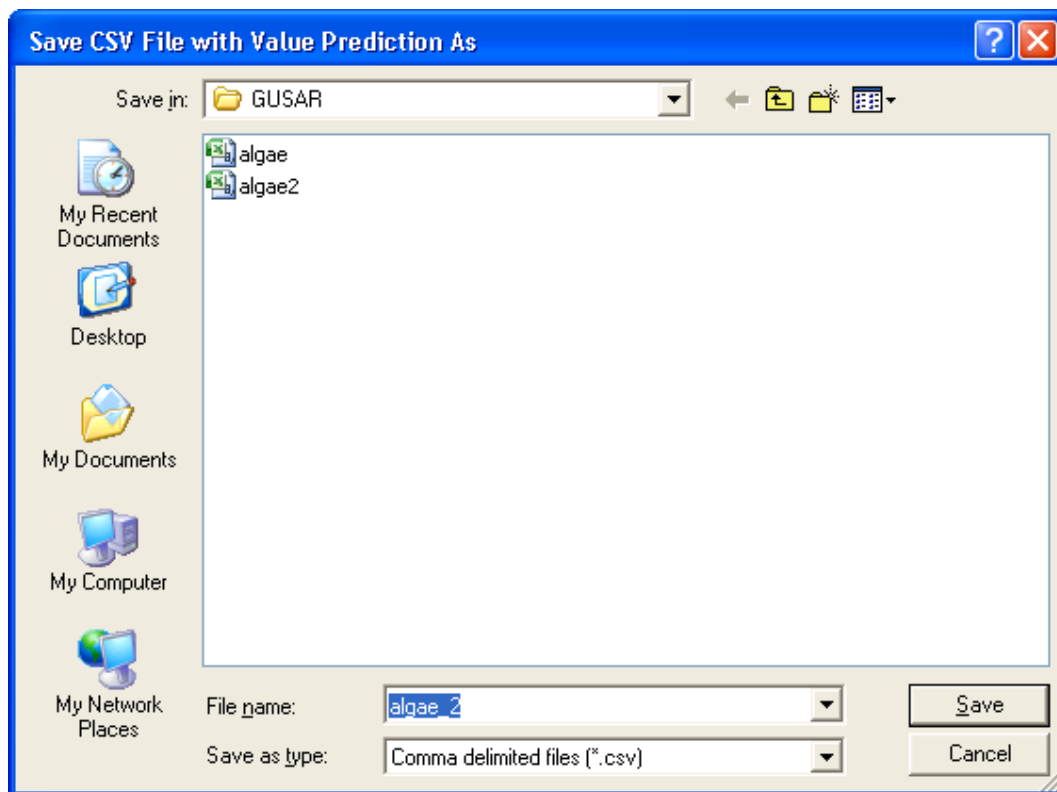
In the beginning of the procedure you should select SD file with structures for GUSAR prediction:



Then you should determine the name of SD or CSV file where prediction results will be saved:

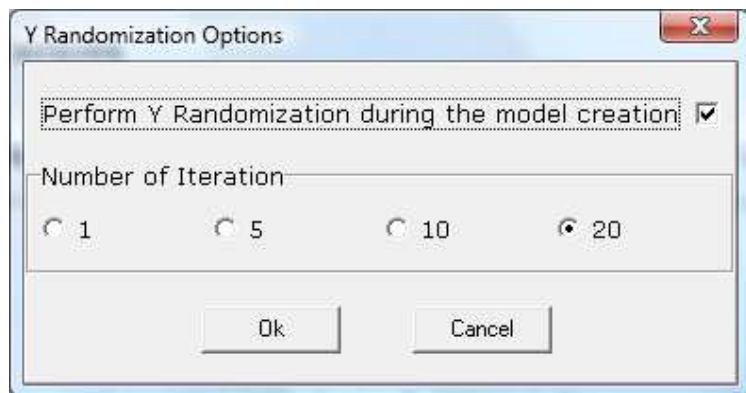


Or for CSV file:



Y RANDOMIZATION

Use “**Y Randomization**” command to determine the options of Y randomization. The algorithm of Y randomization procedure is described in Chapter 7, page 72.



The results of Y randomization procedure will appear in the appropriate column in **Model Selection** window (see the description of **Model Selection** window, page 40).

Number of Iteration parameter determines the number of QSAR models created on the basis of random values. The higher number of iteration leads to more significant result of Y randomisation. Y randomization procedure is executed before the creation of each QSAR model. Y randomization column in **Model Selection** window shows an average Q^2 value of QSAR models created after Y randomization.

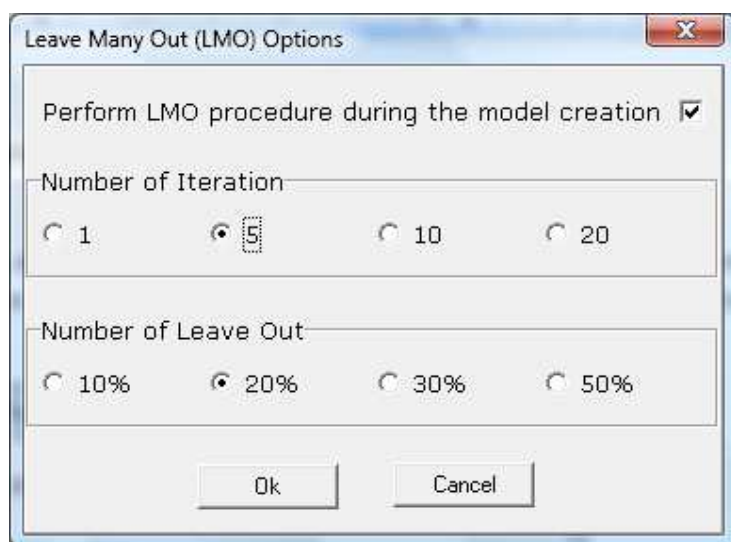
The close values of Y randomization and Q^2 of QSAR models shows that the models were overtrained. In this case, randomization of the real values of dependent variable led to creating QSAR model with similar parameters of QSAR model based on the real data.

Note!

We recommend to apply this procedure for the small training sets (less than 100 structures) because it requires a significant computational time.

LEAVE-MANY-OUT OPTIONS

Use “**Leave-Many-Out**” command to determine the options of LEAVE-MANY-OUT procedure. The algorithm of Leave-Many-Out (LMO) procedure is described in Chapter 7 (page 72).



The results of Leave-Many-Out procedure will appear in the appropriate column(s) in **Model Selection** window (see description of **Model Selection** window, page 40).

Number of Iteration determines the number of cases when the training and test sets are generated in accordance with proportion given in “Number of Leave Out” property. For example if we select 10% “Number of Leave Out” it means that 90% structures will be in the training set and 10% structures will be in the test set. LMO column in **Model Selection** window shows an average R^2 value for the test sets predicted by QSAR models created on the basis of the appropriate training sets.

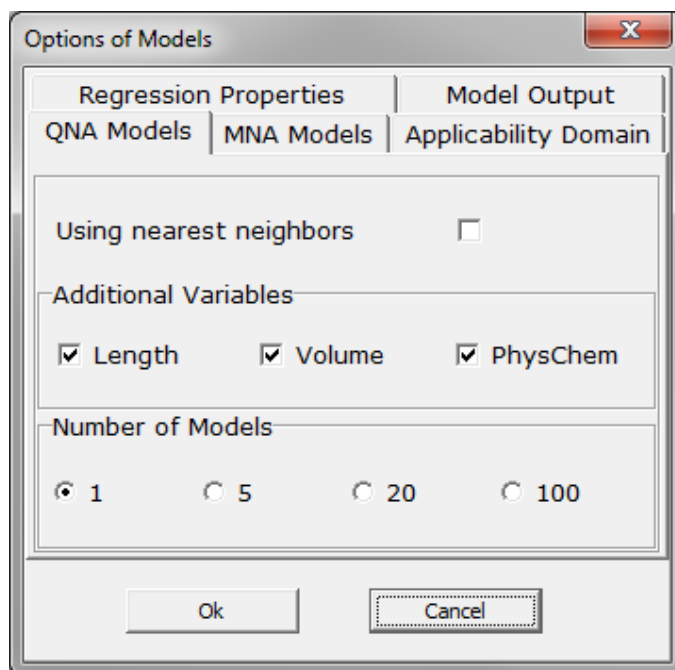
The higher number of iteration leads to more reliable result of LMO procedure. Twenty iterations mean that significance of the LMO results is 90%. The close values of LMO procedure and Q^2 of QSAR models show that the created models are robust.

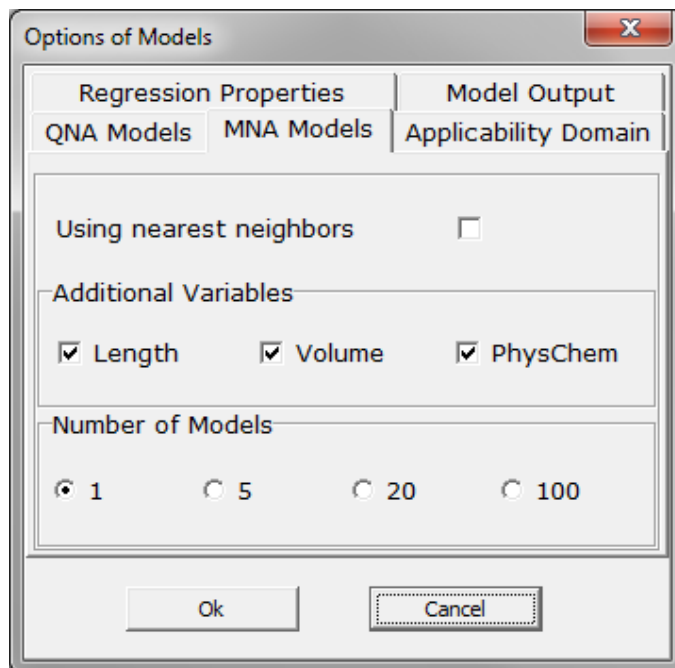
Note!

We recommend using 30% or 50% number of Leave Out in this procedure for large training sets (more than 2000 structures) because the procedure will be calculated faster.

OPTIONS OF MODELS

Use “**Options|Options of Models**” command to display the **Options of Models** window. The window determines options for creation of QSAR models by “**Modeling|Create QNA Models**” and/or “**Modeling|Create MNA Models**” commands. **Options of Models** window also determines parameters of Applicability Domain and application of correction of the prediction results on the basis of data for the nearest neighbours.



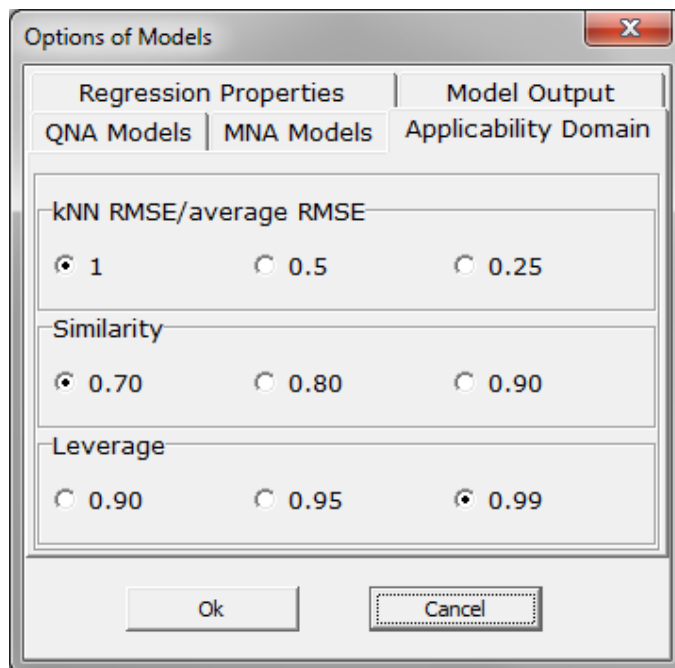


The parameter **“Using nearest neighbours”** means that during the prediction for the analyzed compound the final results of prediction will be corrected on the basis of experimental data of similar compounds from the training set (see Chapter 7 for details, page 67). This function usually increases the prediction accuracy for the external test sets for QSAR models created on the basis of training sets with 500 and more structures. Sometimes application of this function may lead to decrease of accuracy prediction, which depends on the quality of the training sets.

The parameter **“Additional Variables”** means that during the creation of models and the prediction of query structures the selected variables (Length, Volume and/or PhysChem) were used. The methods of these variables calculation are described on pages 62.

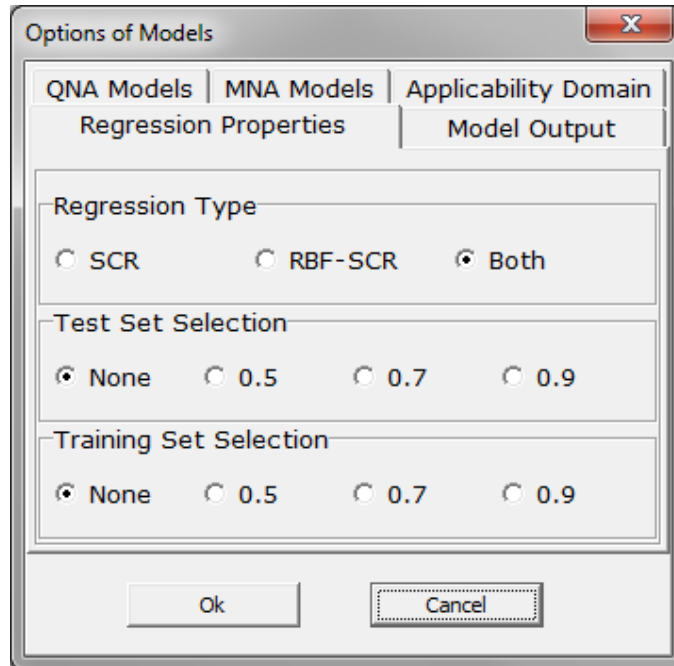
“Number of Models” parameter determines how many models will be created on the basis of the appropriate descriptors. One model for each type of descriptors is created by default. The generation of each model is made randomly, excluding the first model based on QNA descriptors. The highest number of models leads to increase accuracy and robustness by consensus prediction for compounds from the test sets but the time of calculation is considerably increased for large training sets (contained more than 1000 structures). For large training sets we recommend to create 5-6 models for each type of descriptors.

Applicability Domain options contain three different parameters related to determination of reliability of prediction for the analysed compound:



- **kNN RMSE/average RMSE:** limitation by accuracy of prediction of nearest neighbours of the studied compound. 1 means that RMSE of nearest neighbours is equal or less that RMSE of a model. 0.5 - RMSE of nearest neighbours is equal or less that 0.5 RMSE of a model. 0.25 - RMSE of nearest neighbours is equal or less that 0.25 RMSE of a model. The less value of kNN RMSE/average RMSE leads to increase in accuracy of prediction for external test set but it decreases the number of compounds falling into AD (decreases coverage of a test set);
- **Similarity:** limitation by similarity of studied compound with the structures of nearest neighbours. The higher value of similarity leads to increase in accuracy of prediction for external test set but it decreases the number of compounds falling into AD (decreases coverage of a test set);
- **Leverage:** limitation by similarity of studied compound with structures of the whole training set. The highest value of leverage leads to increase in accuracy of prediction for the external test set but it decreases the number of compounds falling into AD (decreases coverage of a test set).

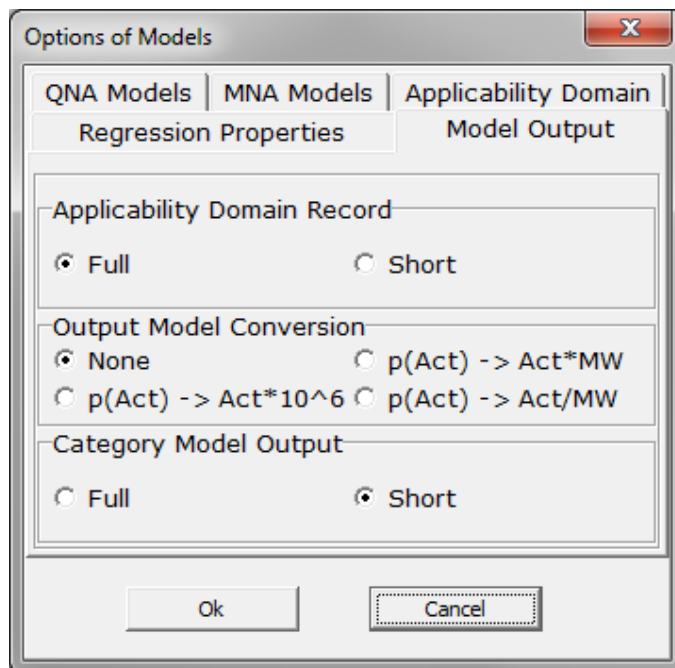
The detailed algorithms of calculation of these Applicability Domain parameters are represented in Chapter 7 (page 70).



“**Regression Type**” parameter determines what type of regression analysis will be used for model developing. Self-consistent regression (SCR) is used by default. RBF-SCR is a non-linear modification of the self-consistent regression, which is based on linear radial basis functions (more details are presented in Chapter 7). This type of regression provides more accurate prediction results, but it is time consuming. Selection of Both types of regression methods allows achieving predictions from both methods as consensus result.

The parameter “**Test Set Selection**” means that during leave-many-out cross validation procedure only models that exceed the selected value will be kept.

The parameter “**Training Set Selection**” means that during leave-one-out cross validation procedure only models that exceed the selected value will be kept.



“**Applicability Domain Record**” parameter determines type of output prediction results for CSV file format. Output data can be organized in two formats: Full and Short. Description of these formats is presented in Chapter 3.

“**Output Model Conversion**” parameter allows to convert prediction results into appropriate values according to described equations. $P(\text{Act})$ means $\text{Log}_{10}(\text{activity value})$, Act means activity value, MW means molecular weight.

“**Category Model Output**” parameter determines type of output prediction results for category models. Short type selection provides binary output prediction results: active or inactive. Full type selection provides output prediction results in digital format.

CHAPTER 6

TROUBLESHOOTING

The best practice is to create regular backup from all newly created or modified data files:

- SAR Base files,
- generates SDF files

If needed, one can reinstall GUSAR from the distribution media and use the restored data files generated earlier.

Please contact us by e-mail: support@way2drug.com for GUSAR support.

CHAPTER 7

TERMS AND DEFINITIONS

Basis of QSAR

Quantitative structure-activity relationships (QSARs) have been employed in numerous areas from drug design to the assessment of chemical toxicity. From a general point of view, the estimate y_{pred} of activity for an organic molecule can be represented as:

$$y_{pred} = a_0 + \sum_i a_i f_i(S), \quad (1)$$

where a_0, a_1, \dots are the variable coefficients, $f_1(S), f_2(S), \dots, f_i(S), \dots$ are independent from the coefficients a_0, a_1, \dots different functions of an organic molecule's structure S . This equation represents the QSAR model.

The obtained QSAR model for the studied activity may be used to screen unstudied compounds in order to establish priorities for expensive and time-consuming traditional bioassays. When conditions do not permit traditional bioassays, QSARs are an alternative to bioassays for estimating required activity.

GUSAR software was developed to create power QSAR models for different biological activities or physical-chemical properties. The algorithm of GUSAR is based on Neighborhoods of Atoms descriptors and self-consistent regression.

MOLECULAR DESCRIPTORS

Multilevel Neighborhoods of Atoms (MNA) Descriptors

MNA descriptors are based on the molecular structure representation, which includes hydrogens according to the valences and partial charges of the atoms and does not specify the types of bonds. MNA descriptors are generated as recursively defined sequence:

- zero-level MNA descriptor for each atom is the mark A of the atom itself;
- any next-level MNA descriptor for the atom is the sub-structure notation $A(D_1 D_2 \dots D_i \dots)$,

where D_i is the previous-level MNA descriptor for i -th immediate neighbour's of the atom A .

The mark of atom may include not only the atomic type but also any additional information about the atom. In particular, if the atom is not included into the ring, it is marked by "-". The neighbour descriptors $D_1D_2...D_i...$ are arranged in unique manner, e.g., in lexicographical order. Iterative process of MNA descriptors generation can be continued covering first, second, etc. neighbourhoods of each atom.

For the regression analysis this molecule structure representation was transformed using original PASS algorithm. This algorithm estimates the biological activity profiles for chemical compound using MNA descriptors as input parameters. Therefore, we used the results of PASS prediction as independent variables for regression analysis. The results of PASS prediction are given as a list of biological activities, for which the difference between probabilities to be active (P_a) and to be inactive (P_i) was calculated. The activities from the list of predicted biological activities were randomly selected as input independent variables for regression analysis. This allows to obtain the different QSAR models. PASS version 12.1 predicts 6400 kinds of biological activity with the mean prediction accuracy of about 95%. The list of predicted biological activities includes pharmacotherapeutic effects, mechanisms of action, adverse and toxic effects, metabolic terms, susceptibility to transporter proteins and activities related to gene expression. The results of the PASS procedure are output as a list of the difference between the probabilities, for each biological activity, that the compound is active (P_a) or inactive (P_i). For building the different QSAR models in GUSAR, subsets of these P_a - P_i values were randomly selected from the total list of predicted biological activities as input independent variables for the regression analysis.

MNA descriptors and their transformation do not provide information about the shape, and volume of a molecule although this information may be important for determination of the structure-activity relationships. Therefore, these parameters were additionally added to the obtained variables. It was appeared that the use some physic-chemical descriptors as an additional variables sometimes improved the accuracy of the created models. Therefore we provide the possibility use these variables optionally during the creation of the models. Description of physic-chemical descriptors is presented on page 62.

Topological length (L) of a molecule is calculated as the maximal distance between any two atoms.

Volume (V) of a molecule – as the sum of each atom's volume, $\frac{4}{3}\pi R^3$, where R is the atomic radius (see Table 1).

Physico-chemical descriptors are described on page 62.

The number of initial variables depends on the number of compounds in the training set. If the number of compounds in the training set less than 25, then the number of the initial variables is 24. If the number compounds in the training set varied from 25 to 100, then the number of the initial variables is computed by the following equation:

$$A = (\ln(B) \times 18.755) - 36.37,$$

where A – initial variables and B – the number of compounds in the training set.

If the number of compounds in the training set varied from 100 to 2000, then the number of the initial variables is computed by the following equation:

$$A = \frac{1}{2} \times B,$$

where A – initial variables and B – the number of compounds in the training set.

If the number of compounds in the training set exceeds 2000, then the number of initial variables is 1000.

GUSAR uses self-consistent regression for models building. The number of the final variables in QSAR equation selected after self-consistent regression procedure is significantly less comparing to the number of the initial variables (see SCR description on the page 65).

Quantitative Neighborhoods of Atoms (QNA) Descriptors

QNA descriptors are calculated based on the connectivity matrix (\mathbf{C}) and the standard values of ionization potential (IP) and electron affinity (EA) of atoms in a molecule. For any given atom i QNA descriptors are calculated as following:

$$P_i = B_i \sum_k (\text{Exp}(-\frac{1}{2} \mathbf{C}))_{ik} B_k \quad (2)$$

$$Q_i = B_i \sum_k (\text{Exp}(-\frac{1}{2} \mathbf{C}))_{ik} B_k A_k, \quad (3)$$

where $A_k = \frac{1}{2}(IP_k + EA_k)$, $B_k = (IP_k - EA_k)^{-\frac{1}{2}}$. The values of EA and IP collected from many different sources and used in this work are represented in Appendix (Table 1). Though the value $\mu P - Q$ can be considered by convention as the partial atomic charge, where μ is the chemical potential, in general, the P and Q values are not the estimates of partial atomic charges, hardness, etc.

Table 1. Electron affinity (EA) and first ionization potential (IP), electron volts, and atomic radius (AR), angstroms.

Atom	EA	IP	AR	Atom	EA	IP	AR	Atom	EA	IP	AR
H	0.75	13.60	0.46	Kr	-0.42	14.00	1.98	Lu	0.20	6.15	1.75
He	0.08	24.59	1.22	Rb	0.49	4.18	2.48	Hf	0.33	7.50	1.59
Li	0.62	5.39	1.55	Sr	-0.15	5.69	2.15	Ta	0.40	7.89	1.46
Be	-0.20	9.32	1.13	Y	0.31	6.22	1.81	W	0.67	7.98	1.40
B	0.28	8.30	0.91	Zr	0.33	6.84	1.60	Re	0.23	7.88	1.37
C	1.26	11.26	0.77	Nb	0.51	6.88	1.45	Os	1.44	8.73	1.35
N	0.44	14.53	0.71	Mo	0.68	7.09	1.39	Ir	1.57	9.10	1.35
O	1.46	13.62	0.73	Tc	0.54	7.23	1.36	Pt	1.10	8.96	1.38
F	3.45	17.42	0.71	Ru	1.10	7.37	1.34	Au	1.25	9.23	1.44
Ne	0.00	21.57	1.60	Rh	1.14	7.46	1.34	Hg	-0.19	10.44	1.57
Na	0.55	5.14	1.87	Pd	1.11	8.34	1.37	Tl	0.31	6.11	1.71
Mg	-0.31	7.64	1.60	Ag	1.22	7.58	1.44	Pb	1.39	7.42	1.75
Al	0.30	5.99	1.43	Cd	-0.43	8.99	1.56	Bi	0.97	7.29	1.82
Si	1.39	8.15	1.34	In	0.31	5.79	1.66	Po	1.97	8.42	1.56
P	0.75	10.49	1.30	Sn	1.39	7.34	1.58	At	2.90	9.20	1.48
S	2.00	10.36	1.04	Sb	0.90	8.64	1.61	Rn	-0.15	10.75	2.27
Cl	3.61	12.97	0.99	Te	1.97	9.01	1.70	Fr	0.48	3.98	2.80
Ar	-0.37	15.76	1.92	I	3.23	10.45	1.53	Ra	-0.15	5.28	2.35
K	0.50	4.34	2.36	Xe	-0.25	12.13	2.18	Ac	0.80	5.20	2.03
Ca	-0.19	6.11	1.97	Cs	0.47	3.89	2.66	Th	0.80	6.10	1.80
Sc	0.19	6.56	1.64	Ba	-0.15	5.21	2.23	Pa	0.84	6.00	1.62
Ti	0.33	6.82	1.46	La	0.30	5.59	1.87	U	0.82	6.19	1.53
V	0.53	6.74	1.34	Ce	0.25	5.54	1.83	Np	0.82	6.20	1.50
Cr	0.67	6.77	1.27	Pr	0.20	5.47	1.83	Pu	0.84	6.06	1.62
Mn	-0.17	7.43	1.30	Nd	0.20	5.53	1.82	Am	0.85	6.00	1.70
Fe	0.50	7.90	1.26	Pm	0.20	5.58	1.81	Cm	0.85	6.09	1.55
Co	0.66	7.86	1.25	Sm	0.20	5.64	1.80	Bk	0.82	6.23	1.49
Ni	1.16	7.64	1.24	Eu	0.20	5.67	2.04	Cf	0.84	6.27	1.42
Cu	1.23	7.72	1.28	Gd	0.20	6.15	1.80	Es	0.86	6.47	1.43
Zn	-0.44	9.39	1.39	Tb	0.20	5.86	1.78	Fm	0.86	6.60	1.38
Ga	0.30	6.00	1.39	Dy	0.20	5.94	1.77	Md	0.83	6.68	1.38
Ge	1.39	7.90	1.39	Ho	0.20	6.02	1.78	No	0.79	6.58	1.47
As	0.80	9.79	1.48	Er	0.20	6.11	1.76	Lr	0.85	6.69	1.30
Se	2.02	9.75	1.17	Tm	0.20	6.18	1.75	Db	0.46	6.43	1.14
Br	3.45	11.81	1.14	Yb	0.20	6.25	1.94	Jl	0.50	6.78	1.01

QNA describes each of the atoms in a molecule, and, at the same time, each of the P and Q values depends on the whole composition and structure of a molecule (Figure 1).

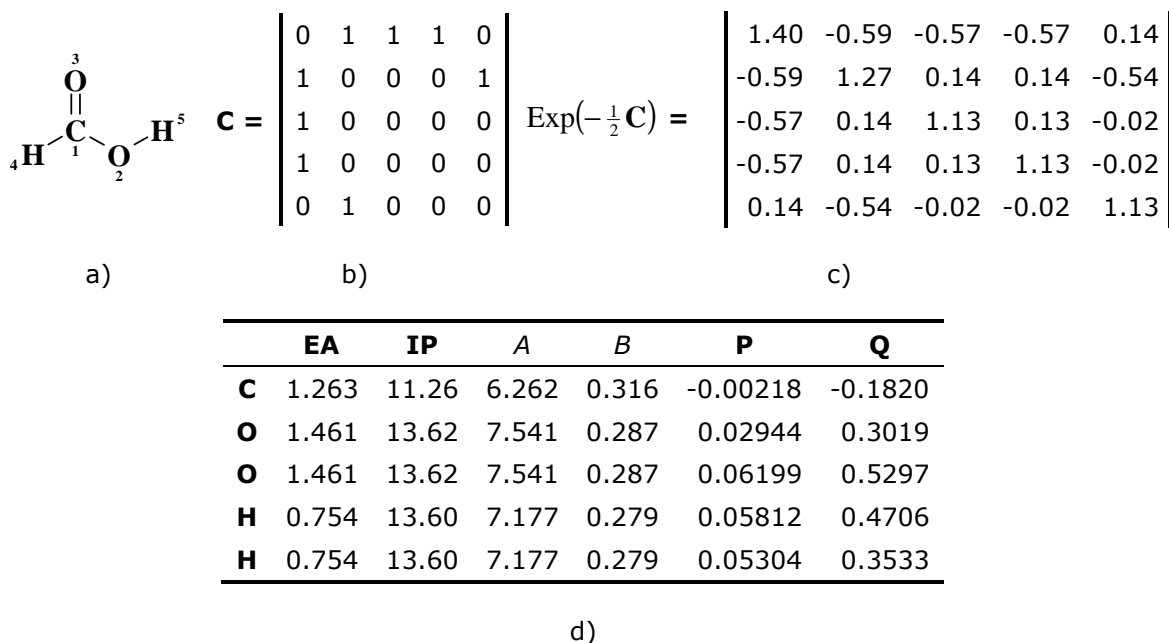


Figure 1. Example of QNA description of *formic acid*: (a) structure diagram; (b) connectivity matrix; (c) exponent of the connectivity matrix; (d) electron affinities, ionization potentials, variables of equations (2) and (3), P and Q values for each atom of *formic acid* molecule.

From Figure 1c it is clear that any atom influences the others, though the influence is decreased with the increase of the distance between them; e.g. components of matrix $\text{Exp}\left(-\frac{1}{2}\mathbf{C}\right)$ for atom 1 (C) are: 1.40 for atom 1 itself, -0.59 for its immediate neighbor atom 2 (O), -0.57 for atoms 3 (O) and 4 (H), and 0.14 for atom 5 (H).

QNA describe each particular atom of a molecule, at the same time, each P or Q value depends on the total molecule composition and structure. To use this molecule structure representation in equation (1) we have proposed to calculate each $f_i(S)$ function of the structure of a molecule as the average value of $g_i(P, Q)$ function of P and Q variables for those m molecule atoms, which have two or more immediate neighbors:

$$f_i(S) = \frac{1}{m} \sum_k g_i(P_k, Q_k) \quad (4)$$

After substitution of expression (4) into the equation (1) and interchange of summations we found:

$$y_{pred} = a_0 + \sum_i a_i \frac{1}{m} \sum_k g_i(P_k, Q_k) = \frac{1}{m} \sum_k \left(a_0 + \sum_i a_i g_i(P_k, Q_k) \right) \quad (5)$$

According to the equation (5) the estimate y_{pred} for molecule can be interpreted as an average of values predicted for particular atoms in a molecule. Formally, QNA descriptors represent a molecule structure by two descriptors only (P and Q), in contrast to numerous traditional descriptors used in QSAR.

Since P and Q values have different ranges (standard deviations are 0.023 and 0.208, respectively), to optimize a family of functions $g_i(P, Q)$ we made the normalization.

Normalization has been performed by calculation of the average values (E_P and E_Q), standard deviations (D_P and D_Q) and correlation between P and Q values (R_{PQ}):

$$P' = \frac{P - E_P}{D_P}, \quad Q' = \frac{Q - E_Q}{D_Q} \quad (6)$$

$$U = \frac{P' + Q'}{\sqrt{2(1 + R_{PQ})}}, \quad V = \frac{P' - Q'}{\sqrt{2(1 - R_{PQ})}} \quad (7)$$

The orthonormal U and V have zero mean, unit variance and they are uncorrelated.

Chebyshev polynomials were chosen as the family of functions $g_i(P, Q)$, and the orthonormal U and V values have been additionally transformed by using hyperbolic tangent, so, the "normalized QNA" vary from -1 to 1. After this, the functions $g_i(P, Q)$ in equation (5) are represented using Chebyshev polynomials as:

$$g_i(P, Q) = T_{uv}(P, Q) = \text{Cos}(u * \text{ArcCos}(\text{TanH}(U))) * \text{Cos}(v * \text{ArcCos}(\text{TanH}(V))), \quad (8)$$

where the integer numbers $u, v = 0, 1, 2, \dots$ define 2-dimensional Chebyshev polynomial degree. The final equation for estimate y_{pred} using QNA descriptors is:

$$y_{pred} = \frac{1}{m} \sum_k (a_0 + \sum_{uv} a_{uv} T_{uv}(P_k, Q_k)) = a_0 + \sum_{uv} a_{uv} T_{uv}, \quad T_{uv} = \frac{1}{m} \sum_k T_{uv}(P_k, Q_k). \quad (9)$$

QNA descriptors and their polynomial transformations (6)-(8) do not provide information on the shape and volume of a molecule although this information may be important for determination of the structure-activity relationships. Therefore, these parameters were added to QNA descriptors. It was appeared that the use some physico-chemical descriptors as an additional variables sometimes improved the accuracy of the created models. Therefore we provide the possibility use these variables optionally during the creation of the models.

Physico-Chemical Descriptors

1. **Topological length** of a molecule was calculated as the maximal distance between any two atoms.
2. **Volume of a molecule** – as the sum of each atom's volume, $\frac{4}{3}\pi R^3$, where R is the atomic radius (see Table 1 (AR)).
3. Number of positive charges.

4. Number of negative charges.
5. Number of hydrogen bond acceptors.
6. Number of hydrogen bond acceptors.
7. Number of aromatic rings.
8. The molecular weight of the molecule.
9. Number of halogen atoms.

10. **Lipophilicity** is calculated on the basis of the set of compounds with structures and known experimental data of LogP collected from different literature sources. The set consists from 3358 compounds with the data of LogP values ranged in the interval from -7 to 8. Special QSAR models have been created on the basis of PASS prediction results and Chebyshev polynomials of QNA descriptors for consensus prediction of LogP values. The accuracy of created models was evaluated by leave 50% out cross-validation procedure repeated by 20 times ($R^2_{L20\%OCV}$). The models have the following characteristics:

LogP (MNA) – R^2 0.91, Q^2 0.88, $R^2_{L20\%OCV}$ 0.85.

LogP (QNA) – R^2 0.84, Q^2 0.79, $R^2_{L20\%OCV}$ 0.75.

The first, second and third power of topological length, volume of a molecule and lipophilicity were used. The use of these variables is determined in “**Options of Models**” (see page 51).

The Chebyshev polynomials are arranged in ascending order of their degrees $u + v$. For $u + v = 1$ they are $T_{1,0}$, $T_{0,1}$; for $u + v = 2$ they are $T_{2,0}$, $T_{1,1}$, $T_{0,2}$; for $u + v = 3$ they are $T_{3,0}$, $T_{2,1}$, $T_{1,2}$, $T_{0,3}$, etc. The first, second and third power of topological length and volume of a molecule were used. The number of initial variables equals to the number of Chebyshev polynomials plus the number of the first, second and third power of topological length, volume of a molecule and lipophilicity.

The number of initial variables depends on the number of compounds in the training set. If the number of compounds in the training set less than 25, then the number of the initial variables is 24. If the number of compounds in the training set varies from 25 to 100, then the number of the initial variables depends on the following equation:

$$A = (\ln(B) \times 18.755) - 36.37,$$

where A – the number of the initial variables and B – the number of compounds in the training set.

If the number of compounds in the training set varied from 100 to 2000, then the number of the initial variables depends on the following equation:

$$A = \frac{1}{2} \times B,$$

where A – the number of the initial variables and B – the number of compounds in the training set.

If the number of compounds in the training set exceeds 2000, then the number of the initial variables is 1000.

The number of the final variables in QSAR equation selected after self-consistent regression procedure is significantly less comparing to the number of the initial variables (see SCR description, page 65).

GUSAR algorithm uses three procedures described below for generation of different QSAR models based on QNA descriptors.

- 1) U and V values from (8) are multiplied by value, which are randomly chosen from 0.98 to 1.09.
- 2) Coefficient before matrix C is randomly chosen in range from -0.1 to -0.9.
- 3) Calculation of QNA descriptors is randomly chosen from two cases. First case is when P and Q values are calculated for all atom types. Second case is when P and Q values are calculated for each atom, which has connection with more than one carbon.

All these procedures are used randomly for creation of each QSAR model based on QNA descriptors. These procedures allow obtaining the different QNA models.

METHODS

Self-Consistent Regression (SCR)

The classical MLR has a number of limitations. Particularly, the number of objects in the training set should significantly exceed the number of independent variables, and it is important to use the non-collinear variables only. To overcome these limitations, we have employed the approach based on statistical regularization of ill-posed problems. It has resulted in the regularized least-squares method:

$$\mathbf{a} = \text{ArgMin} \left[\sum_{i=1}^n (y_i - \sum_{k=0}^m x_{ik} a_k)^2 + \sum_{k=1}^m v_k a_k^2 \right] \quad (10)$$

where \mathbf{a} is the regression coefficient, n is the number of objects, y_i is the response value of i^{th} object, m is, here and below, the number of the independent variables, x_{ik} is the value of k^{th} independent variable of i^{th} object, a_k is the k^{th} value of the regression coefficients, v_k is the k^{th} value of the regularization parameters. Equation (10) has the following solution:

$$\mathbf{a} = \mathbf{T}\mathbf{X}^T \mathbf{y}, \quad \mathbf{T} = (\mathbf{X}^T \mathbf{X} + \mathbf{V})^{-1}$$

where \mathbf{X}^T is the transposed regression matrix \mathbf{X} , \mathbf{V} is a diagonal matrix of the regularization parameters. The best regularization \mathbf{V} satisfies to the equations:

$$v_k (a_k^2 + s^2 t_k) = s^2, \quad k = 1, \dots, m$$

where s is the standard deviation of residuals, t_k is the k^{th} diagonal element of matrix \mathbf{T} .

We called this method "self-consistent regression" (SCR) because the same data samples (\mathbf{X} and \mathbf{y}) are used to estimate both the regression coefficients and the regularization parameters. Unlike the stepwise regression and other methods of combinatorial search, the initial SCR model includes all regressors. Nevertheless, the final model may contain a few variables only, correctly representing the existing relationship.

Radial basis function (RBF)

Radial basis function (RBF) neural networks form a class of ANNs, which has certain advantages over other types of ANNs, including better approximation capabilities, simple network architecture, and faster learning algorithms. The main idea of RBF neural networks is to create the proper number of hidden nodes, which is represented by radial basis function, and determine the weight of each node. Often for selection of centers of the nodes the different clustering methods are used. Some of them require setting up the initial number of nodes (centroid-based clustering), another methods calculate the optimal number of nodes (distribution-based clustering). After selection of nodes it is

necessary to calculate the weight of each node. For this purpose the simple least square method can be used.

Although RBF network is very powerful approach, one disadvantage should be mentioned. As the most of ANN, RBF neural network need to select the hidden nodes, which is both tricky and sometimes poor reproducible. To avoid that problem the radial basis function interpolation approach can be applied. The difference between radial basis function interpolation approach and RBF network is that the first one has the number of hidden nodes equal to the number of input variables (training examples). Thus, the learning procedure of RBF interpolation approach is performed through all elements in the training set:

$$y(x) = \sum_{i=1}^N w_i \phi(\|x - x_i\|) = \Phi w \quad (11)$$

where the approximating function $y(x)$ is represented as a sum of N radial basis functions, each associated with a different center x_i , and weighted by an appropriate coefficient w_i .

If the points x_i are distinct, than the interpolation matrix Φ in the above equation is non-singular. Thus, the weights w can be solved by simple way:

$$w = \Phi^{-1}y \quad (12)$$

However, RBF interpolation approach can be sensitive to noise created by both a huge number of descriptors and quality data.

The basic purpose of the SCR method is to remove the variables, which poorly describe the appropriate target value. The final number of variables in the QSAR equation, selected after the SCR procedure, is significantly less compared to the initial number of variables. Also, SCR is robust against noise in the data. The regression coefficients, obtained from SCR, reflect the contribution of each particular descriptor (variable) in the final equation. The higher absolute value of coefficient is, the greater contribution of descriptor is. Thus, regression coefficients received after SCR can be used for weighting of descriptors (variables) in accordance with their importance. This advantage was used for combining the self-consistent regression and the radial basis function interpolation methods.

Typically, the radial basis function is calculated using the distance or similarity between descriptors and centroid. In the case of RBF interpolation the same similarity is calculated between input variables or descriptors. Taking into account contribution of each descriptor the more accurate similarity value can be achieved. For this purpose, during

calculation of radial basis functions, the descriptors are weighted by coefficients obtained after SCR. Thus, RBF-SCR method is calculated as:

$$y(x) = \sum_{i=1}^N w_i \phi(\|ax - a_i x_i\|) \quad (13)$$

where a is taken from equation 10.

Therefore, RBF-SCR can be presented as 3 steps algorithm:

- 1) First step, self-consistent regression determines coefficients and selects descriptors.
- 2) Second step, radial basis functions are calculated using similarity, which is weighted by SCR coefficients.
- 3) RBF weights are determined by least square method.

The linear radial basis function was applied for RBF interpolation. In comparison to Gaussian function the linear function has opposite meanings: the more input variables (compounds) are dissimilar, the more contribution they provide. Thus, linear radial basis function can be used for diverse training sets with high level of dissimilarity between the objects.

Use of Nearest Neighbours

Algorithm of GUSAR allows using nearest neighbours for improvement of prediction value obtained from regression model. The corrected value is estimated by taking an average of the three chemicals values in the training set that are the most similar to the chemical under study. Similarity of the any pair of chemical compounds is estimated as Pirson's coefficient calculated in the space of independent variables obtained after SCR. If the using nearest neighbours in "**Options of Models**" window is chosen, then the predicted value is estimated as average between the corrected value and the value obtained from regression equation. If this option is not chosen, then the predicted value is estimated only by the regression equation.

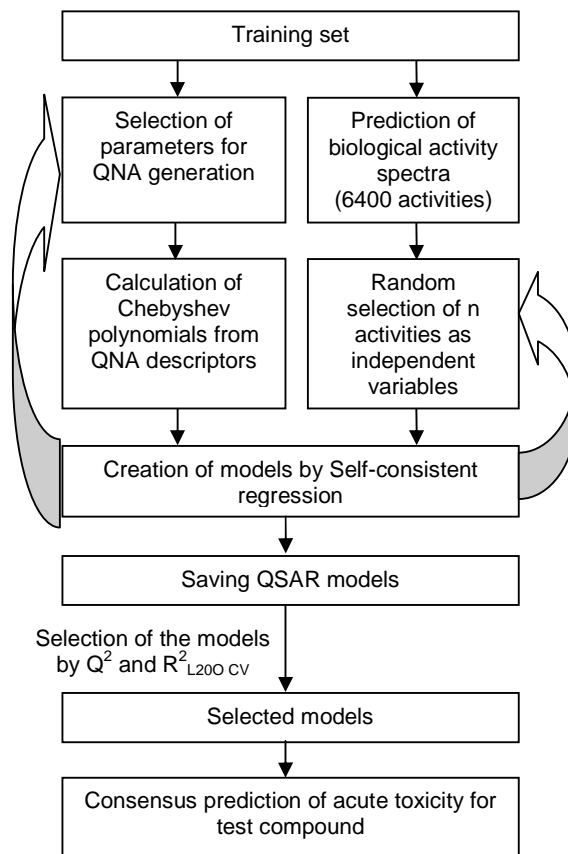
This function usually increases the prediction accuracy for the external test sets for QSAR models created on the basis of training sets with 500 and more structures. Sometimes the use of this function may lead to decrease of accuracy prediction, which depends on the quality of the training sets.

Consensus method

The predicted value is estimated by taking into account a weighted average of the predicted values from the obtained QSAR models (QSAR models provide the predictions that are within the respective applicability domains). Prediction obtained from each developed model is weighted on its applicability domain values (AD). GUSAR algorithm calculates the similarity value; leverage value and estimation of three nearest neighbours' accuracy of prediction for the applicability domain assessment (see AD description, page 70). Weighting of the predicted value is performed using similarity and leverage values only. For each types of AD the end-user could chose a threshold in the **"Options of Models"** window.

If all predicted values from the obtained QSAR models exceed the similarity threshold and leverage threshold, then the consensus prediction value is calculated from similarity weighted predicted values. If all predicted values from the obtained QSAR models are less than the leverage and similarity thresholds, then the consensus prediction value is calculated from leverage weighted predicted values. If the predicted values from the obtained QSAR models exceed similarity and is less than leverage thresholds, then the consensus prediction value is calculated as average from leverage weighted predicted values and similarity weighted predicted values. If the predicted values from the obtained QSAR models are less than similarity and exceed the leverage thresholds, then the consensus prediction value is calculated as average from all weighted predicted values.

The algorithm combining the results of QSAR modelling on the basis of QNA descriptors and PASS predicted biological activity profiles, is represented in the following Scheme:



APPLICABILITY DOMAIN ASSESSMENT

Similarity

Three nearest neighbours from the training set are calculated for each chemical compound under study using similarity estimation. Similarity of the pair of chemical compounds is estimated as Pearson's coefficient calculated in the space of independent variables obtained after SCR. The average similarity of three nearest neighbours is used for assessment of the applicability domain (AD) of the model. If the average similarity exceeds the threshold selected in the "**Options of Models**" window (see page 53), then the studied chemical compound falls in AD of the model and vice-versa. The higher value of the threshold was selected, the more similar compounds fell in AD of the model.

Leverage

Hat value of leverage was used for domain applicability assessment. Hat values from the leverage matrix representing the "distance" of the molecule to the model structural space were calculated as:

$$Leverage = x^T (X^T X)^{-1} x ,$$

where x is a vector of descriptors of a query compound, and X is a matrix formed with rows corresponding to the descriptors of molecules from the training set. Hat value of leverage was calculated for each compound of the training set and then a distribution of the obtained values was estimated.

A warning level of hat value was considered as percentile, which defines in "**Options of Models**" window. The end-user could select 99th, 95th and 90th percentile. Therefore, if a chemical compound from the external test set has hat value exceeded this warning level, then this compound is considered as being out of the applicability domain.

Accuracy of three nearest neighbours' predictions

For the assessment of this type of the applicability domain the following equation is used:

$$AD_{value} = RMSE_{3NN} / RMSE_{train} ,$$

where AD_{value} is the applicability domain value, $RMSE_{3NN}$ is root mean square error of prediction of three most similar compounds from the training set, $RMSE_{train}$ is root mean square error of predictions for the training set.

Three nearest neighbours from the training set are calculated for each studied chemical compound using similarity value. Similarity of the two chemical compounds is estimated as Pearson's coefficient calculated in the space of independent variables obtained by SCR.

If the AD_{value} is less than the threshold determined in the "**Options of Models**" window (see page 53), then the studied chemical compound falls into AD of the model and vice-versa. Therefore, any chemical compound, whose three nearest neighbours are predicted worse than the prediction accuracy of the whole training set, is considered as being out of the applicability domain.

VALIDATION METHODS

Leave-Many-Out procedure

This procedure is used for assessment of external prediction accuracy. The initial data is randomly divided onto the training and the external test sets. The proportions of this splitting are selected in the "**LEAVE-MANY-OUT OPTIONS**" window (see page 50). End-user could select the following proportions:

- 1) 10% for the external test set and 90% for the training set;
- 2) 20% for the external test set and 80% for the training set;
- 3) 30% for the external test set and 70% for the training set;
- 4) 50% for the external test set and 50% for the training set.

For obtaining the objective assessment of predictive accuracy and robustness of the developed models each splitting should be repeated many times. End-user could select the following number of repeats: 1, 5, 10 and 20.

Y-Randomization procedure

This procedure allows to be ensuring that the developed models do not have the overfitting. In this procedure the dependent-variable vector, Y vector, is randomly shuffled and new QSAR model is developed using the original independent-variable matrix. It is expected that the resulting models should generally have low Q^2 values. This procedure could be repeated many times for each model, and then the average Q^2 value is calculated. The number of iteration could be selected in the "**Y RANDOMIZATION**" window (see page 49).

REFERENCES

1. Lagunin A.A., Glorizova T.A., Dmitriev A.V., Volgina N.E., Poroikov V.V. Computer Evaluation of Drug Interactions with P-Glycoprotein. *Bulletin of Experimental Biology and Medicine*, 2013, **154(4)**, 521-524.
2. Masand V.H., Mahajan D.T., Patil K.N., Hadda T.B., Youssoufi M.H., Jawarkar R.D., Shibi I.G. Optimization of antimalarial activity of synthetic prodiginines: QSAR, GUSAR, and CoMFA analyses. *Chem Biol Drug Des.* 2013, **81(4)**, 527-536.
3. Zakharov A.V., Peach M.L., Sitzmann M., Filippov I.V., McCartney H.J., Smith L.H., Pugliese A., Nicklaus M.C. Computational tools and resources for metabolism-related property predictions. 2. Application to prediction of half-life time in human liver microsomes. *Future Med Chem.* 2012, **4(15)**, 1933-1944.
4. Zakharov A.V., Lagunin A.A., Filimonov D.A., Poroikov V.V. Quantitative prediction of antitarget interaction profiles for chemical compounds. *Chemical Research in Toxicology*, 2012, **25(11)**, 2378-2385.
5. Lagunin A., Zakharov A., Filimonov D., Poroikov V. QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction. *Molecular Informatics*, 2011, **30(2-3)**, 241-250.
6. Filimonov D.A., Zakharov A.V., Lagunin A.A., Poroikov V.V. QNA based 'Star Track' QSAR approach. *SAR and QSAR Environ. Res.*, 2009, **20 (7-8)**, 679-709.
7. Lagunin A., Zakharov A., Filimonov D., Poroikov V. In silico assessment of acute toxicity in rodents. *Toxicology Letters*, 2009, **189** (S1), S254.
8. Filimonov D.A., Poroikov V.V. Probabilistic approach in activity prediction. In: *Cheminformatics Approaches to Virtual Screening*. Eds. Alexandre Varnek and Alexander Tropsha. Cambridge (UK): RSC Publishing, 2008, 182-216.
9. Filimonov D.A., Poroikov V.V. Prediction of biological activity spectrum for organic compounds. *Rus. Chem. J.*, 2006, **50**, 66-75.
10. Filimonov D.A., Lagunin A.A., Poroikov V.V. Prediction of activity spectra for substances using new local integrative descriptors. *QSAR and Molecular Modelling in Rational Design of Bioactive Molecules*. Eds. Esin Aki Sener, Ismail Yalcin, Ankara (Turkey), CADD & D Society, 2005, 98-99.
11. Poroikov V. V., Filimonov D. A., Borodina Yu. V., Lagunin A. A., Kos A. Robustness of Biological Activity Spectra Predicting by Computer Program PASS for Noncongeneric Sets of Chemical Compounds. *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1349-1355.
12. Filimonov D., Poroikov V., Borodina Yu., Glorizova T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 666-670.
13. Poroikov V. V., Filimonov D. A., Borodina Yu. V., Glorizova T. A., Sitnikov V. B., Sadovnikov S. V., Sosnov A. V. Quantitative relationships between structure and delayed neurotoxicity of chemicals studied by the Self-Consistent Regression method using the PASS program. *Pharmaceutical Chemistry Journal*, 2004, **38** (4), 188-190.

14. Filimonov D. A., Akimov D. V., Poroikov V. V. The method of Self-Consistent Regression for the quantitative analysis of relationships between structure and properties of chemicals. *Pharmaceutical Chemistry Journal*, 2004, **38** (1), 21-24.

