

# QSARINS v 2.1 Manual

## Index

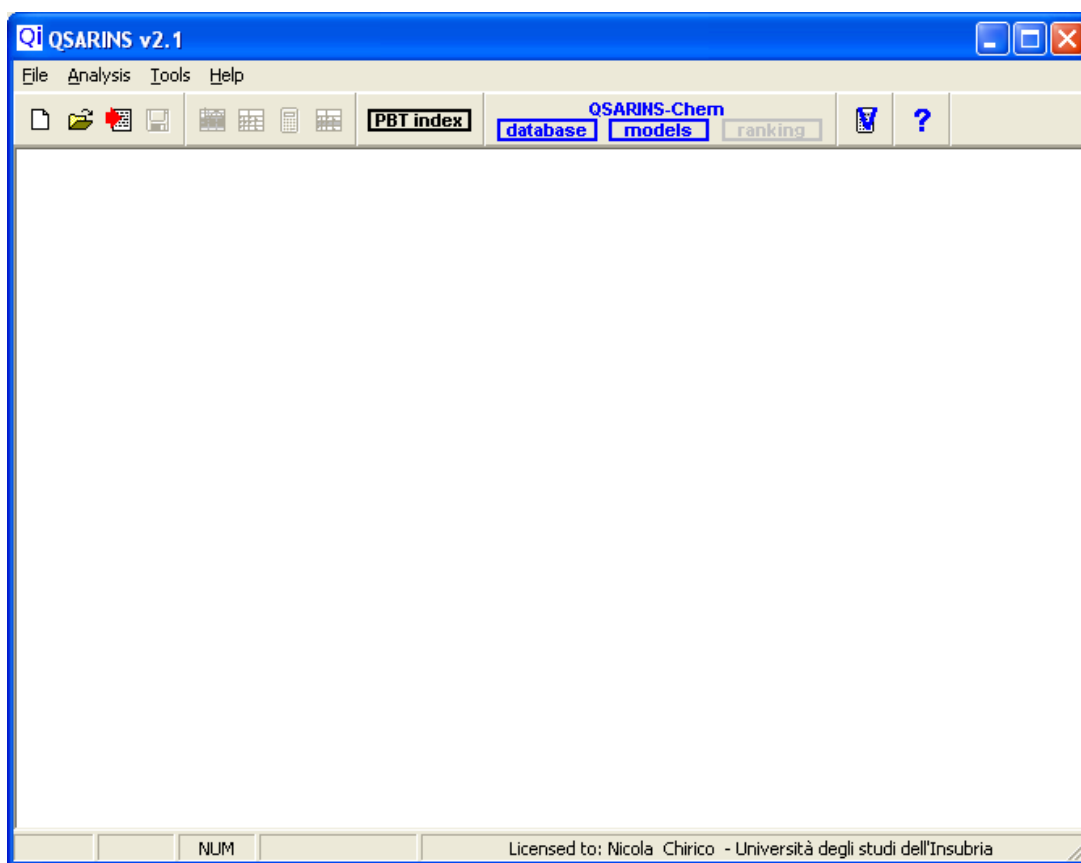
<b>How to use QSARINS</b>	<b>3</b>
<b>1. Import the dataset</b>	<b>5</b>
1.1 <i>Dataset format</i>	5
1.2 <i>Importing the dataset</i>	6
<b>2. View data</b>	<b>8</b>
<b>3. Data setup</b>	<b>14</b>
<b>4. Variable selection and Model calculation</b>	<b>18</b>
4.1 <i>Variable selection</i>	18
4.2 <i>Model Calculation</i>	21
<b>5. View and select models</b>	<b>23</b>
<b>6. Model validation</b>	<b>32</b>
<b>7. Analysis of single models</b>	<b>36</b>
<b>8. Consensus modeling</b>	<b>48</b>
<b>9. QSARINS-Chem module</b>	<b>52</b>
9.1. <i>Database</i>	52
9.1.1 <i>Querying the database</i>	53
9.1.2 <i>Visualize chemical structures</i>	59
9.1.3 <i>Prepare a user-defined dataset</i>	62

9.2 <i>Apply developed model</i>	65
9.2.1 <i>Apply developed model: the PBT Index example</i>	67
9.3 <i>Ranking</i>	70
9.4 <b>Validate experimental vs. predicted data</b>	73
<b>Additional Information</b>	76
<b>Contacts</b>	76
<b>References</b>	77
<b>Figure captions</b>	80

## HOW TO USE QSARINS (version 2.1, 2014)

QSARINS (QSAR-INSUBRIA) (presented in “QSARINS: A New Software for the Development, Analysis, and Validation of QSAR MLR Models”, *Journal of Computational Chemistry, Software News and Updates*, Gramatica et al., 2013, 34, 2121-2132, DOI:10.1002/jcc.23361, and in the new version in “QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS”, *Journal of Computational Chemistry, Software News and Updates*, Early View, DOI: 10.1002/jcc.23576, can be used after obtaining a license from the Insubria group (contacts: [paola.gramatica@uninsubria.it](mailto:paola.gramatica@uninsubria.it), [nicola.chirico@uninsubria.it](mailto:nicola.chirico@uninsubria.it), [stefano.cassani@uninsubria.it](mailto:stefano.cassani@uninsubria.it) ).

Once obtained a license, and agreed the terms of its use, the main screen (Figure 1) appears allowing the QSAR developer/user to begin working:



**Figure 1.** Main window of QSARINS

Here it follows a brief explanation of the options of the main screen, in the same order as they appear:



(or File->New project): Start a new QSARINS project (.qsi) (this option will erase all existing data)



(or File->Open project): load a QSARINS project (.qsi)



(or File->Import data): import data (descriptors, responses, splitting etc.) from external source



(or File->Save project): save QSARINS project (.qsi) (this option is disabled until data are imported or a project is loaded)



(or Analysis->View data): view dataset both in tabulated and in graphical forms (this option is disabled until data are imported or a project is loaded)



(or Analysis->Data setup): organize data for model calculation (this option is disabled until data are imported or a project is loaded)



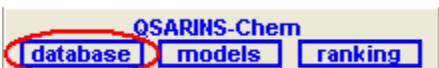
(or Analysis->Variable selection and models calculation): set the preferred technique and parameters for model calculation (this option is disabled until the data setup has been performed)



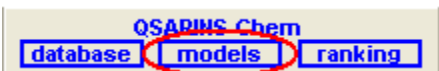
(or Analysis->View and select models): shows the calculated models in tabulated form and allows further analysis (this option is disabled until models are calculated)



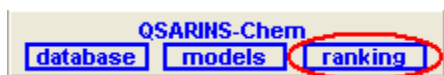
(or Tools->Apply PBT index model): apply the Insubria PBT index QSAR model



(or Tools->Query database): view and query molecules from the database (QSARINS-Chem module)



(or Tools->Apply developed models): apply stored models to new chemicals (QSARINS-Chem module)



(or Tools->Calculate ranking): once imported a dataset, calculate ranking from data columns (QSARINS-Chem module)



(or Validate experimental vs. predicted data): calculates various validation criteria from experimental vs. predicted data.



(or Help->QSARINS manual): view QSARINS manual.

In the Help menu the user can view the general manual and additional information about QSARINS, and can import an example dataset for training purpose.

## 1. Import the dataset

### 1.1 Dataset format

In order to be imported (as .txt or .csv files), a dataset must be formatted according to some rules, as in the following example:

ID	CAS	nC	ATSc5	ATSm1	Resp	split
1	000061-82-5	2	0	7.43992121	1.92	1
2	000094-97-3	6	-0.02065243	18.79289081	-999	2
3	000095-14-7	6	-0.00050502	10.0799409	3.78	2
4	000130-34-7	16	-0.02408049	49.89117957	-999	2

The first data row contains the labels of the corresponding columns, as the ID and CAS of the molecules, the descriptors, the response and the splitting status (training/prediction set). All the corresponding columns must contain the same number of rows and all data columns (including the labels) must be separated by the same character, called “separator” (in this example a tabulation character, but it can be different, as a semicolon, comma, or whatever character, provided it is not used by the data themselves). A column of chemicals name (CAS in the example above) or ID must precede all the other columns (the order of descriptors, responses and splitting columns is unimportant). If both ID and CAS are present, only the second column can be selected by QSARINS and treated as chemical names, keeping the same order of input data.


The decimal separator of numerical data must be the dot (hence, dot cannot be used as the separator of the columns).

The splitting columns can have only three values, i.e., 1 if it belongs to the training set, 2 if it belongs to the prediction set and 0 if the molecule is not included in any set.

(Note: PaDEL-Descriptor (Yap, 2011), DRAGON software (DRAGON, 2007) and QSPR-THESAURUS online platform (QSPR-THESAURUS, 2011), three packages used by the Insubria group for the calculation of the descriptors, generate output files that are compatible with the aforementioned dataset format.) The data file can be saved as .txt or .csv, the two formats recognized by QSARINS.

### *1.2 Importing the dataset*

The dataset, including the list of chemicals, molecular descriptors, response and splitting (if already set) can be imported in the following ways from the main window (Figure 1): (the number of chemicals and descriptors are basically limited to the machine resources, being the maximum allowable data matrix  $2^{31} \times 2^{31}$  elements)

File > Import, or click the icon 

The following dialog will appear:

Import data - C:\Documents and Settings\lab-qsar\Desktop\ProvaMANUALE.txt

Name column  Label row  ☒ Exclude if const. > %  Import

Missing value  Delimiter ☐ ☒ tab ☒ Exclude if corr. > %  Cancel

	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Col 11	Col 12	Col 13
Row 1	ID	Name	Log Koc Exp	Splitting	MW	TopoPSA	VAdjMat	nH	nC	nN	nO	nS	nF
Row 2	1	000050-00-0	0.56	2	30.01056466	17.07	2	2	1	0	1	0	0
Row 3	2	000050-29-3	5.33	1	351.9146887	0	5.247927513	9	14	0	0	0	0
Row 4	3	000050-32-8	5.83	2	252.0939004	0	5.321928095	12	20	0	0	0	0
Row 5	4	000051-66-1	1.4	2	165.0789786	38.33	4.584962501	11	9	1	2	0	0
Row 6	5	000052-68-6	1.6	2	255.9225784	65.57	4.584962501	8	4	0	4	0	1
Row 7	6	000053-70-3	6.25	2					22	0	0	0	0
Row 8	7	000054-11-5	2.01	1					10	2	0	0	0
Row 9	8	000055-21-0	1.46	2					7	1	1	0	0
Row 10	9	000055-38-9	3.18	1					10	0	3	2	1
Row 11	10	000056-23-5	1.85	2					1	0	0	0	0
Row 12	11	000056-38-2	3.2	2	291.0330302	112.73	5.169925001	14	10	1	5	1	1
Row 13	12	000056-49-5	6.1	2	268.1252005	0	5.392317423	16	21	0	0	0	0
Row 14	13	000056-53-1	4.14	2	268.1463295	40.46	5.321928095	20	18	0	2	0	0
Row 15	14	000056-55-3	5.4	2	228.0939004	0	5.169925001	12	18	0	0	0	0
Row 16	15	000057-13-6	1.03	2	60.03236275	69.11	3	4	1	2	1	0	0
Row 17	16	000057-55-6	0.36	2	76.0524295	40.46	3.321928095	8	3	0	2	0	0
Row 18	17	000057-97-6	5.36	2	256.1252005	0	5.321928095	16	20	0	0	0	0
Row 19	18	000058-89-9	3	2	287.8600663	0	4.584962501	6	6	0	0	0	0
Row 20	19	000058-90-2	3.35	2	229.8859754	20.23	4.459431615	2	6	0	1	0	0

Figure 2. Import data dialog

The first and the second up/down buttons ("Name column" and "Label row") are used to recognize the row of the labels (descriptor names, responses etc.) and the column of the chemical names: the row and the column are highlighted in light blue (Row 1 and Col 2 in Figure 2). The "Missing value" edit box allows modifying the arbitrary internal code -999, which corresponds to the lack of the experimental response value in the data matrix.

The edit box called "Delimiter" allows specifying the character used to separate the data to be imported. By default this character is the tabulation ("tab" checkbox in Figure 2).

The two checkboxes, "exclude if const. > %", "exclude if corr. > %", allows to pre-filter data before importing them for model calculation. The first option is used to exclude semi-constant descriptors, i.e. those having chemical compounds with a constant value for more than a certain percentage (suggested value: 80%). The second option serves to exclude descriptors that are too inter-correlated (suggested value: 95%).


Some columns must be excluded during this pre-filtering step, as the experimental responses and the splitting ones. This can be done by selecting the columns and right-clicking on “exclude from corr. and const. filtering”: the selected/excluded ones will be marked in yellow.

In order to avoid any bias while pre-filtering, it is possible to select a splitting column to force the exclusion of constant and correlated variables only using the training set. This can be done by selecting the splitting columns and right-clicking on “Select/Unselect splitting for filtering”: the selected/excluded splitting column will be highlighted in green, while the rows of the training chemicals in light green, as visual hints.

As shown in the popup menu (Figure 2), there is also the possibility to search any variable among the columns of data (clicking on “Search variable”).

## 2. View data

Once the data are imported, they can be further explored both in tabulated and in graphical forms. To perform this task, the user must select the following option from the main window:

Analysis > View data, or click the icon 

The following window will appear:






ID	Name	Status	Log Koc Exp	Splitting	MW	TopoPSA	VAdjMat	nH	nC	nN
1	000050-00-0	Training	0.56	2	30.01056468	17.07	2	2	1	0
2	000050-29-3	Training	5.33	1	254.0446003	0	5.247927513	9	14	0
3	000050-32-8	Training	5.83	2			5.321928095	12	20	0
4	000051-66-1	Training	1.4	2			4.584962501	11	9	1
5	000052-68-6	Training	1.6	2			4.584962501	8	4	0
6	000053-70-3	Training	6.25	2			5.459431619	14	22	0
7	000054-11-5	Training	2.01	1			4.584962501	14	10	2
8	000055-21-0	Training	1.46	2			4.169925001	7	7	1
9	000055-38-9	Training	3.18	1			5	15	10	0
10	000056-23-5	Training	1.85	2			3.321928095	0	1	0
11	000056-38-2	Training	3.2	2			5.169925001	14	10	1
12	000056-49-5	Training	6.1	2			5.392317423	16	21	0
13	000056-53-1	Training	4.14	2			5.321928095	20	18	0
14	000056-55-3	Training	5.4	2			5.169925001	12	18	0
15	000057-13-6	Training	1.03	2			3	4	1	2
16	000057-55-6	Training	0.36	2	76.0524295	40.46	3.321928095	8	3	0
17	000057-97-6	Training	5.36	2	256.1252005	0	5.321928095	16	20	0
18	000058-89-9	Training	3	2	287.8600663	0	4.584962501	6	6	0
19	000058-90-2	Training	3.35	2	229.8859754	20.23	4.459431619	2	6	0
20	000060-09-3	Training	2.79	2	197.0952974	50.74	4.906890596	11	12	3
21	000060-11-7	Training	3.87	2	225.1265975	27.96	5.087462841	15	14	3
22	000060-12-8	Training	1.5	2	122.0731649	20.23	4.169925001	10	8	0
23	000060-51-5	Training	1.08	2	228.9996219	114.76	4.584962501	12	5	1
24	000060-57-1	Training	4.44	2	377.870631	12.53	5.247927513	8	12	0



**Figure 3.** View data window

The main window (Figure 3) shows the tabulated data and some options. Data are arranged almost in the same way as in the “Importing the dataset” section.

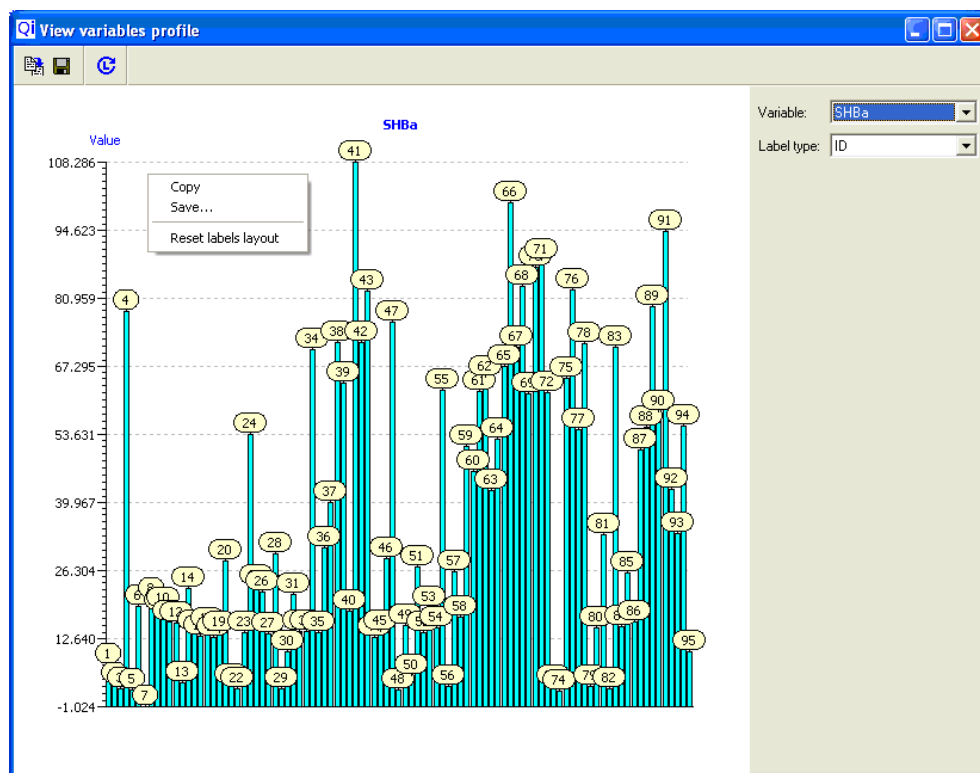
In QSARINS usually the options available as icons have a counterpart as a popup menu (when it is not exactly so, this is due to avoid “too crowded” interfaces, that can be difficult to manage by the user). Sometimes, as in the “View data” dialog, the popup menu has also some additional options. In this case the last two options (“Search variable” and “Search object”) allow searching, within the columns/row, the corresponding variable/object.

Concerning the options in Figure 3, the icon  (or “Copy” in the popup menu) copies in the clipboard the data, previously selected by one of the three subsequent icons where the icon  (or “Select Row” in the popup menu) selects the corresponding rows, the icon  (or “Select Column” in the popup



menu) the corresponding columns and the icon  (or “Select All” in the popup menu) selects the whole dataset.


The following four icons in Figure 3 are used to graphically display the data. The first one (, or “View variables profile” in the popup menu) shows the variable profile, i.e. after choosing a certain descriptor all the corresponding values for each chemical are displayed in a graph bar, while the second one (, “View objects profile” in the popup menu) shows the objects (the chemicals) profile, i.e. after choosing a certain chemical, the corresponding values of the descriptors are displayed.

As an example, here it follows a variable profile graph (Figure 4):




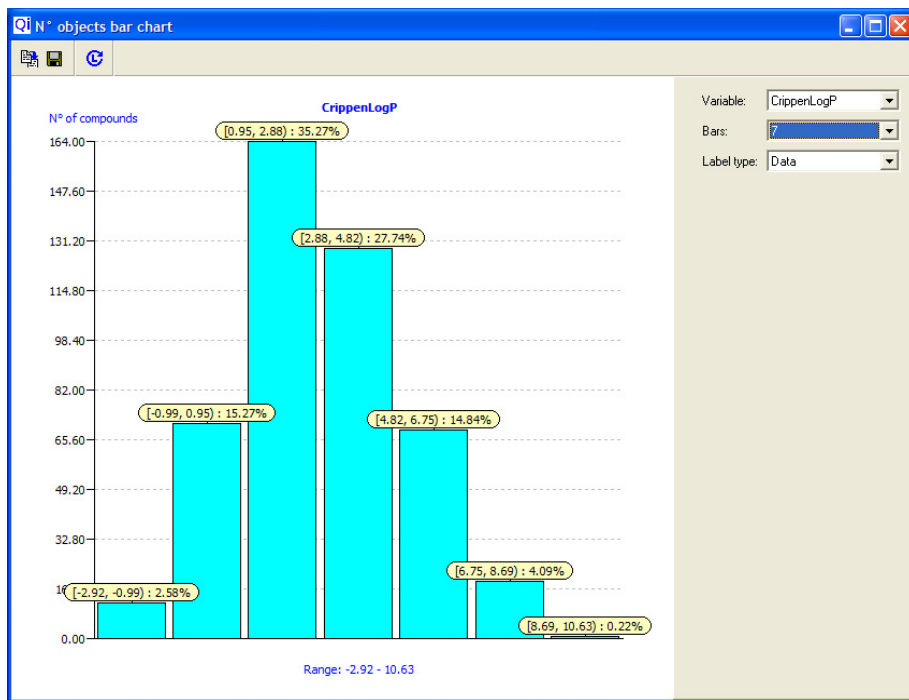
**Figure 4.** Graph of variable profile

The first icon from the left (, or “Copy” in the popup menu) copies the graph into the clipboard, while the second icon (, or “Save” in the popup menu) saves the image as a JPEG or BMP file. Since


the user can move the labels at will on the graph, the third icon (  , or “Reset labels layout” in the popup menu) resets their position to the default.

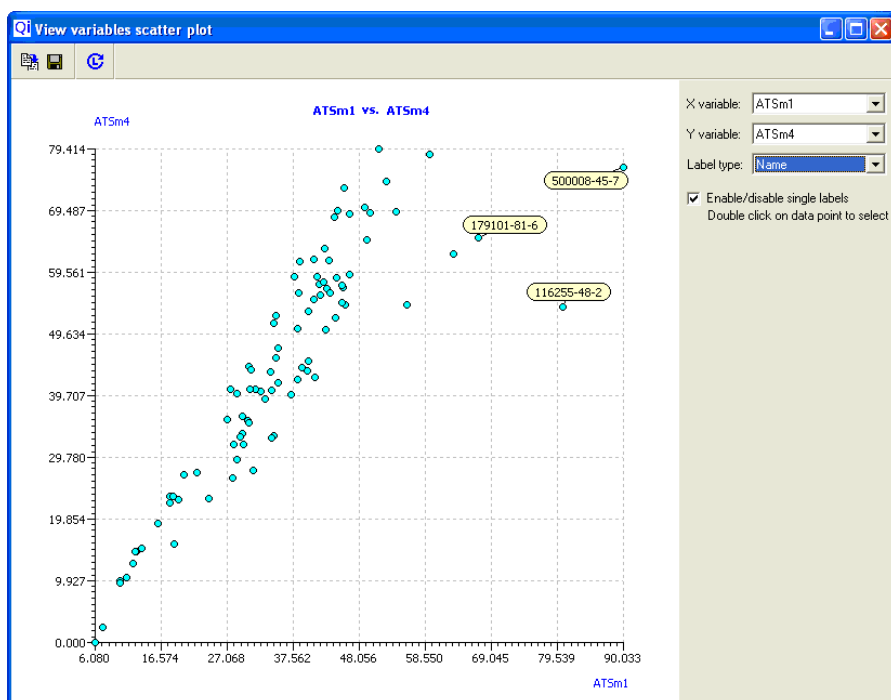
The disposition and function of the just mentioned three icons is the same for all graphs in QSARINS.

The third icon (  , or “View bar chart” in the popup menu) in Figure 3 displays a histogram, as in the following example, that shows how many objects fall in the range of values of the selected descriptor. The number of bars to be displayed is user-definable (Figure 5).




**Figure 5.** Histogram bar chart

The fourth icon (  , or “View scatter plot” in the popup menu) in Figure 3 displays a scatter plot of a couple of data columns chosen by the user, as in the following example (Figure 6):



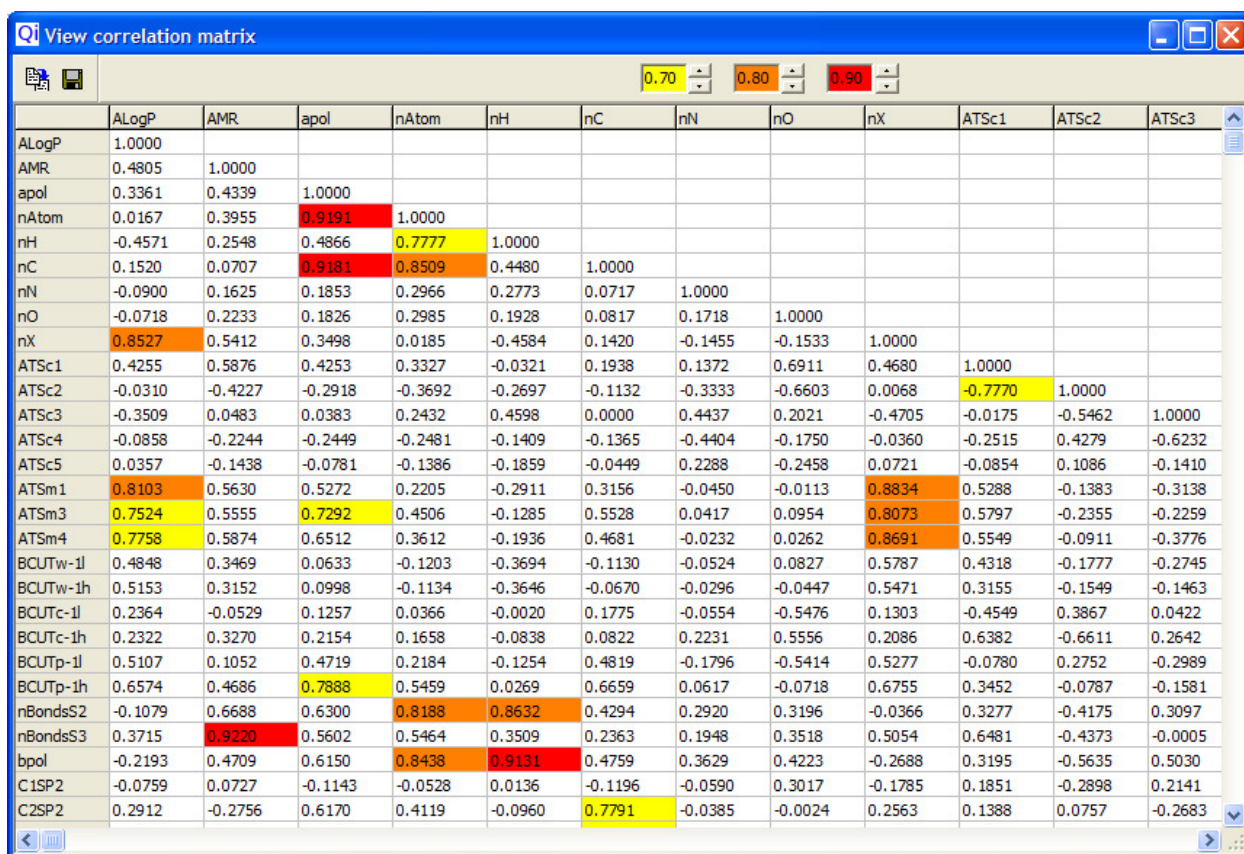
**Figure 6.** View variables scatter plot

The enable/disable single labels option allows the user to display the labels (ID or compound name) singularly for every data point. This option is available for every QSARINS scatter plot.



The last icon ( , or “View correlation matrix” in the popup menu) in Figure 3 shows a correlation matrix of all the imported data (Figure 7). The high correlations values are here highlighted with different colors:

- red, correlation greater than 0.90
- orange, correlation greater than 0.80
- yellow, correlation greater than 0.70

These thresholds can be changed by the user (see upper part of Figure 7), allowing to highlight any degree of correlation among the data.




**Figure 7.** Correlation matrix of the imported data

Pressing the first icon from the left (, or “Copy” in the popup menu) copies the data into the clipboard, while the second icon (, or “Save” in the popup menu) save them in a text file. It is here recalled that every data matrix displayed in QSARINS have Copy and Save options arranged in a similar way.

### 3. Data setup

Once data are loaded, the user selects the input descriptors, the response to be modeled, which molecules are in the training set, in the prediction set and so on. This step is essential to allow the subsequent variable selection for modeling (Section 4.1, "Variable selection").

The data setup can be accessed in the following manner from the main window (Figure 1):

Analysis > data setup, or click the icon 

The following screen (Figure 8) will appear:

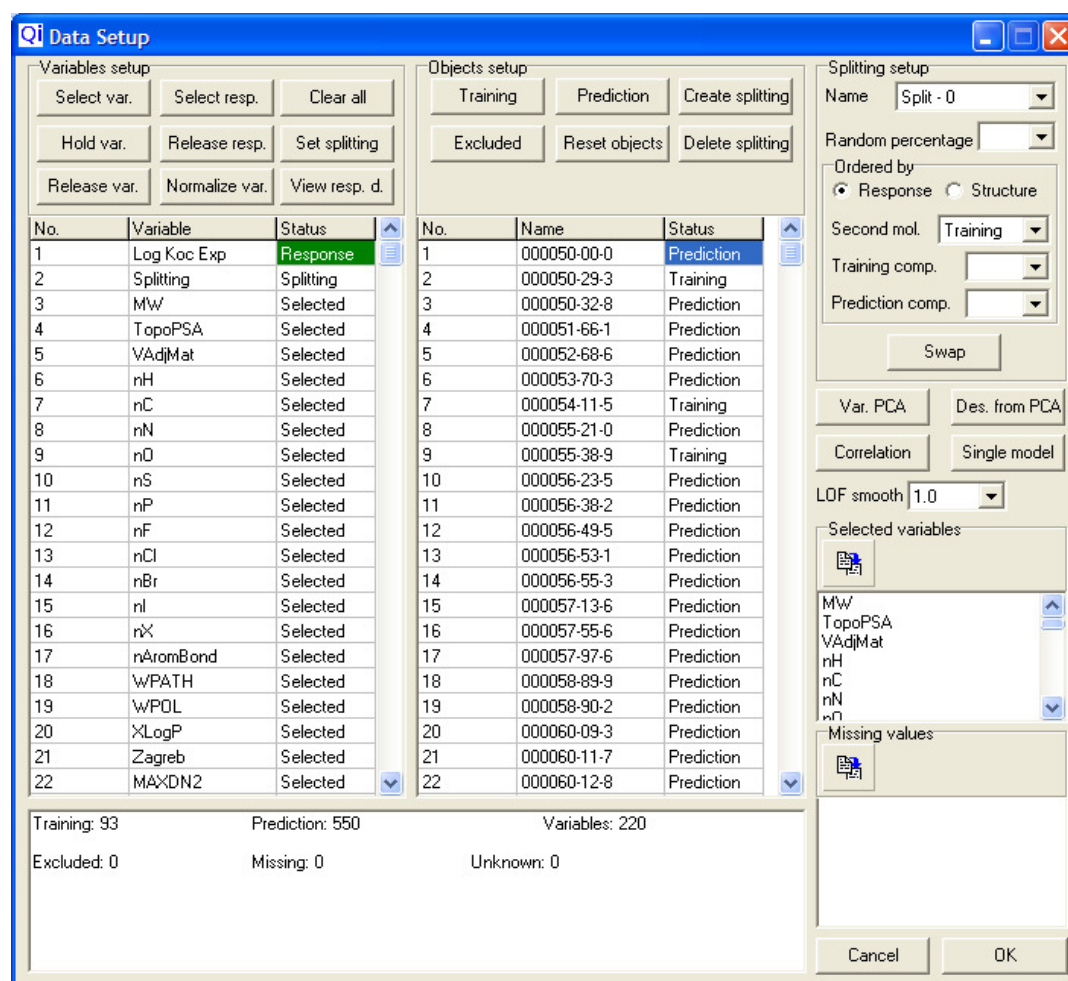


Figure 8. Data setup window

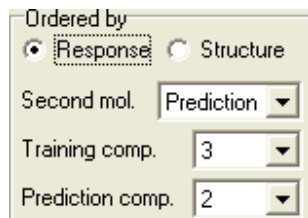
The box on the upper left of Figure 8 concerns the variable setup options. In the corresponding data grid is possible to select the descriptors, the response and (if already available) the pre-assigned splitting status (training/prediction) by means of the specific buttons (“Select/Release Var”, “Select/Release Resp” and “Set splitting”) or using the pop-up menu (mouse right click). Note that a descriptor with one or more undefined values (default: -999) for some chemicals cannot be used in MLR model calculation, thus it will subsequently automatically excluded in the variable selection (to know which are the missing descriptors in the respective chemicals, look at the “Missing values” box). The “Hold Var” option keeps fixed the selected descriptors during the following variable selection (see section 4.2: “Model calculation”, for further details). “Normalize Var” is used to normalize the selected descriptors. “Clear All” resets the selections while “Set Splitting” uses the corresponding imported column as the splitting for determining the chemicals status. The “View. Resp. d.” option plots the distribution of the responses both for training and prediction sets.

The subsequent data grid on the right of Figure 8 concerns the object (chemicals) setup. Once loaded, the chemicals are all set as Training, except the chemicals without the response, that are set as Unknown. The “Objects setup” allows chemical status (training, prediction, excluded) to be manually modified one by one, and saved by clicking the button “Create Splitting” (“Delete Splitting” does the opposite, erasing the selected splitting reported in Splitting setup box - in the figure is called “Split-0”). The “Reset objects” option set the object status to the initial status.

The box of options “Splitting setup” is related to the splitting applicable by the user at this step in QSARINS (if not already pre-defined in the above step). Note that the “Name” list box is automatically filled every time a splitting is saved by the user, both by selecting a pre-assigned splitting (if already available) or by creating a new splitting at this step.

Different kind of splitting can be made using the following options. The “Random percentage” option creates a splitting choosing at random molecules as the prediction set, whose percentage over the total of molecules is chosen by the corresponding list box. The molecules can be ordered (“Ordered by” box) for alternative splitting in two ways: by responses of the modeled end-point (“Response” button) or by structure (“Structure” button) ordering the molecules by the first axis of Principal Component Analysis (PCA)(Jackson, 1991) of the molecular descriptors first axis (PC1 score). The first and the last values of the ordered responses, or alternatively of the PC1 scores, are always set as training, to be sure that the prediction set values are within the model applicability domain. The list box, “Second mol”, identifies

the status of the second molecule (Training or Prediction set). Finally, the “Training comp.” and “Prediction comp.” list boxes are used to set how training and prediction set molecules are alternated within the range of the ordered responses or PC1 scores. It is here important to consider that the first and the last compounds are always set as training, and are not influenced by the “Training / Prediction comp.” and “Second mol.” options. It is here recalled that these options work from the second to the last but one molecule, *not* from the first to the last. Here it follows an example:



The screenshot shows a dialog box titled "Ordered by" with two radio buttons: "Response" (selected) and "Structure". Below this are three dropdown menus: "Second mol." set to "Prediction", "Training comp." set to "3", and "Prediction comp." set to "2".

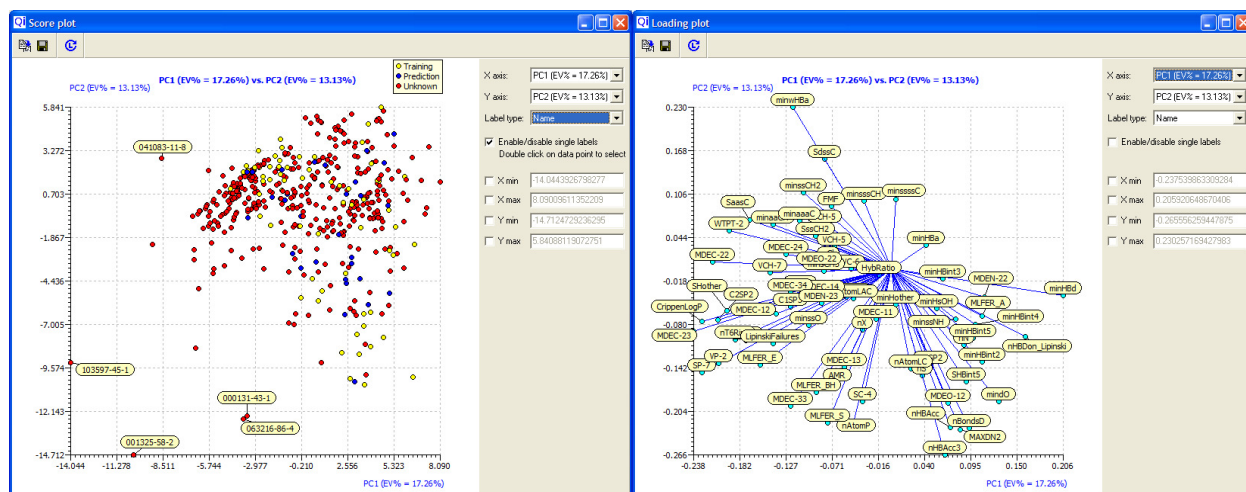
In this case, molecules are ordered according to the response values (ascending order). The second molecule is assigned to the prediction set, the third molecule is assigned to the prediction set (because “Prediction comp.” asks for 2 consecutive prediction set chemicals), and the subsequent three compounds are assigned to the training set (because “Training comp.” asks for 3 consecutive training set chemicals). The following responses are set alternatively as 2 predictions and 3 training, till the last but one molecule. The user can arbitrarily modify all these sequences.

The “Swap” option exchanges the training with the prediction set of the current splitting.


Once the splitting is set by the user, it can be saved using the “Create splitting” button of the “Object setup” section. The new splitting column can be viewed in the View Data option from the main menu (Analysis > View data).

The “Var PCA” button calculates a Principal Component Analysis (PCA) of the selected descriptors both by score plot, in order to check whether the molecules are clustered by structures and/or detect structural outliers, and by the loading plot, that shows which descriptors are influential for the aforementioned (if any) clusters/outliers (Figure 9). The chemicals of training and prediction sets, as well the unknown chemicals, are differently labeled to verify the distributions of the sets. Here it follows an example of the plotting:





**Figure 9.** PCA (Score plot and Loading plot) of the chemicals in the selected descriptors space

Other than saving (  ) the images as JPEG or BMP, the user can save the PCA tabulated data (Scores and Loadings) as text (.txt) file. This option is available for every PCA generated by QSARINS.

The “Des. from PCA” button creates new descriptor variables (also visible in View data, after the descriptors) deriving from the calculation of the PCA of the selected descriptors. Those new descriptors, called PCs, can then be used as modelling variables for the calculation of the corresponding models: this technique is called PC regression. This is useful when there is a need for orthogonal variables that condense more than one descriptor

The “Correlation” button shows the correlation matrix of the selected descriptors and/or responses (or any other imported data), as reported and commented above (Paragraph 2 “View Data”).

The “Single Model” button calculates the model using the selections made in the Variable and Object data grids, and display the results both in tabulated and in graphical form. This option can be applied here if a QSAR model, based on few descriptors, has been already developed (also by other software) or if a modeler applies a personal selection of molecular descriptors.

Regarding the analysis of a single model, since visualizing models is a shared feature with other options of QSARINS, it will be explained in a separate following section (7, “Analysis of Single Models”).

The “LOF smooth” list box value is the smoothness factor to be used for the Friedman Lack Of Fitting (LOF) calculation (Friedman 1991). The LOF penalizes the addition of descriptors in the model equation, and the smoothness factor is used to regulate the amount of penalty.

The remaining boxes of Figure 8 are informative ones concerning the selected variables, the variables containing missing values (that will be automatically excluded from model calculation) and the status of the selection made by the user (i.e. how many variables have been selected, how many chemicals are in the training set, how many in the prediction set and so on).

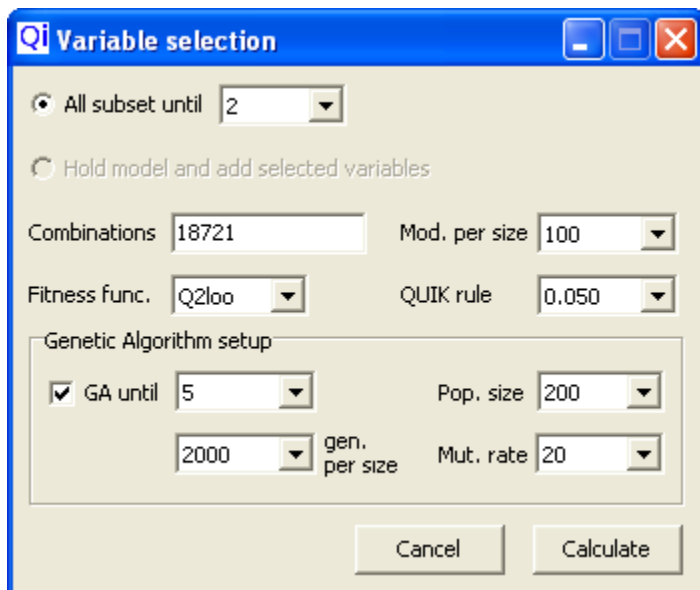
#### 4. Variable selection and Model calculation

##### 4.1. Variable selection

After the data setup has been completed, the next step is an automated procedure for choosing the variables to be used for models calculation. Pressing in the main screen (Figure 1):

Analysis -> Variable selection and models calculation, or click the icon 

the following dialog (Figure 10) will appear:



The dialog box titled "Variable selection" has a blue title bar with standard window controls. It contains two radio buttons: "All subset until" (selected) and "Hold model and add selected variables". The "All subset until" option has a dropdown menu showing the value "2". Below these are two rows of input fields: "Combinations" (text box with "18721") and "Mod. per size" (dropdown with "100"); "Fitness func." (dropdown with "Q2loo") and "QUIK rule" (dropdown with "0.050"). A section titled "Genetic Algorithm setup" contains a checked checkbox "GA until" followed by a dropdown with "5", a "Pop. size" dropdown with "200", a dropdown with "2000" followed by the text "gen. per size", and a "Mut. rate" dropdown with "20". At the bottom are "Cancel" and "Calculate" buttons.

**Figure 10.** Variable selection dialog

The user can apply the following approaches for variable selection (Figure 10):

- All subset: QSARINS explores all the possible combinations of the selected descriptors up to a user defined number of modeling variables (in the list box “All subset until”). The number of all the possible combinations is displayed by the software in the “Combinations” box. A warning message will be displayed if the number of combination seems too high. It is suggested to apply always this step till models based on 2-3 variables, in order to explore all the low-dimension combinations of descriptors.

- Hold model and add selected variables: in the data setup dialog (see section 3) it is possible to select the variables as “Hold”. Such variables are kept fixed in the variable selection procedure, while all the remaining selected variables are iteratively added one by one to this set. In short, such a procedure allows to calculate a set of models based on the preferred descriptors (the hold variables), but adding iteratively a new different descriptor.

- Genetic Algorithm (GA) setup: QSARINS executes the all subset variable selection and models calculation until the user selected model size is reached (“All subset until”). Usually the size of models that can be calculated by the all subset is limited by the number of combinations, which easily reaches a huge value. It is thus possible to continue variable selection by means of a genetic algorithm (GA) which aim is to find the best model based on the best combinations of variables without exploring them all. To apply this technique the user must select a model size in the checkbox “GA until”: this way QSARINS is enabled to continue with GA (Tournament Selection method, Haupt and Haupt, 2004).

Once selected, a second list box called “gen. per size” is activated allowing the user to choose how many GA iterations must be performed (the user can anyway stop the GA before the number of iterations has been reached, pressing the “Done” button as indicated in section 4.2, “Model calculation”).

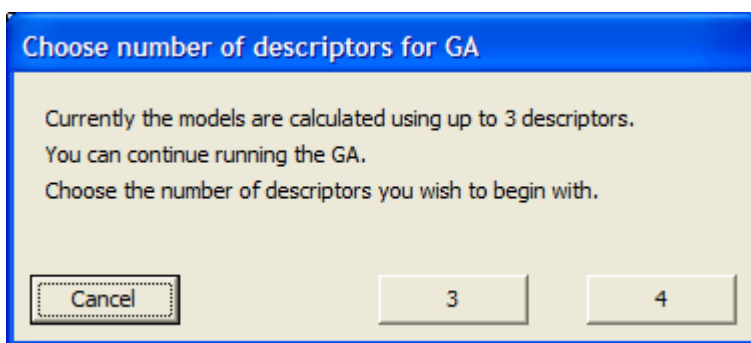
Alternatively, the software automates the model size increase. In this case, if we set a bigger size as the one indicated by default in the “GA until” list box, as for example 4 from a hypothetical base model size 2, and we press the “Calculate” button leaving the procedure running, the models of size from 2 to 4 will be calculated, iterating the GA for the number of times reported in the “gen. per size” list box. Once calculations have been performed, the user can continue the GA, calling again the Variable selection dialog.

In addition, for the whole variable selection procedure, the user can set:

- the fitness function ( $Q_{LOO}^2$  or  $R_{adj}^2$  to be maximized, LOF or  $RMSE_{CV}$  to be minimized) applied by QSARINS also to sort the models meanwhile they are generated.
- the QUIK rule (Todeschini et al., 1999) to automatically exclude the models, where the correlation between the block of the descriptors and the response ( $K_{xy}$ ) is lower than or too similar to the inter-correlation among the descriptors ( $K_{xx}$ ). This difference should be the highest as possible. (it is suggested to set a value  $\geq 0.05$ , i.e.  $K_{xy}-K_{xx} \geq 0.05$ )
- the number of models per size to be stored (only the best, according to the selected fitness function)
- the population size ("Pop. size"): number of models on which GA evolves (the bigger it is, the better the GA works because it can use more "chromosomes" thus exploring more combinations, but the slower are the GA iterations. It is up to the user to find a good tradeoff).
- Mutation rate (%) ("Mut. rate") applied during the GA iterations, to generate a pool of descriptors "variegated" by random mutations.

When choosing how many variables have to be used for calculations, if their number respect to the objects is too high, the user will be warned.

It is here recalled that, once the procedure is completed, i.e. the "Calculate" button has been pressed and the model calculation (see section 4.2) is finished, it is possible to call again the variable selection procedure and continue with the GA. In this case the following dialog will appear asking whether to continue the GA iterations using the same model size of the previous calculations or increasing the final model size:

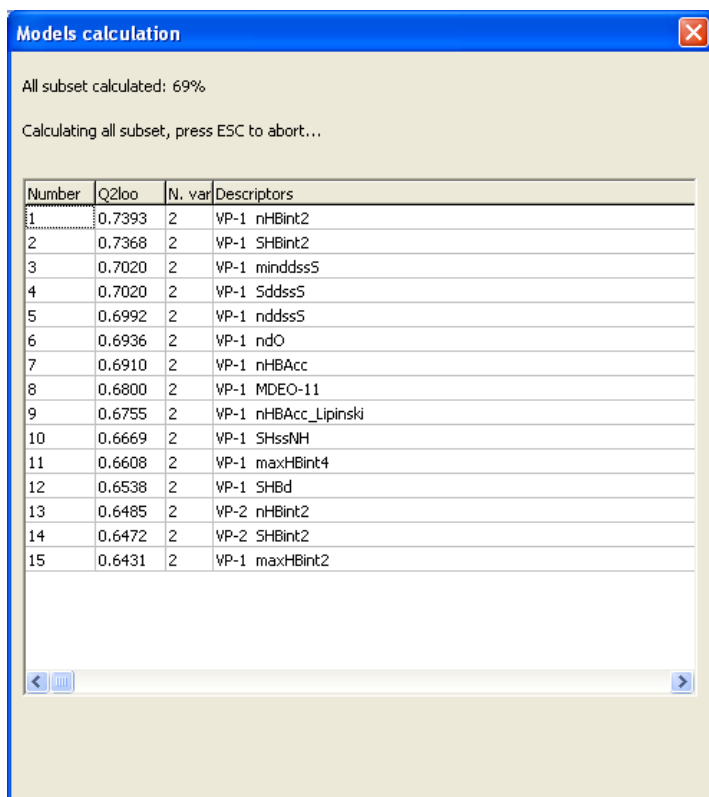


Once chosen the number of variables for restarting the GA, the dialog of variable selection (Figure 10) will appear and the user can set again the model size, number of iterations for variable, population size, number of model per size, fitness function, QUIK rule and mutation rate.

The GA procedure will restart adding to the previous population of models the new combinations of variables, according to the selected parameters (Figure 10).

#### 4.2. Model calculation

Once the variable selection setup has been completed and the model calculation begins, the following dialog box (Figure 11) appears:

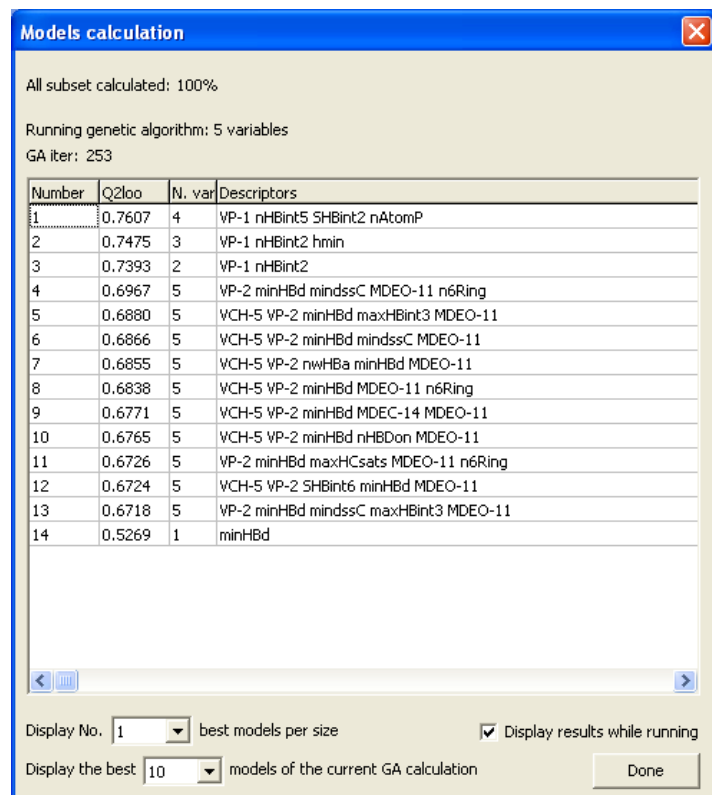


Number	Q2loo	N. var	Descriptors
1	0.7393	2	VP-1 nHBint2
2	0.7368	2	VP-1 SHBint2
3	0.7020	2	VP-1 minddss5
4	0.7020	2	VP-1 Sddss5
5	0.6992	2	VP-1 nddss5
6	0.6936	2	VP-1 ndO
7	0.6910	2	VP-1 nHBAcc
8	0.6800	2	VP-1 MDEO-11
9	0.6755	2	VP-1 nHBAcc_Lipinski
10	0.6669	2	VP-1 SHssNIH
11	0.6608	2	VP-1 maxHBint4
12	0.6538	2	VP-1 SHBd
13	0.6485	2	VP-2 nHBint2
14	0.6472	2	VP-2 SHBint2
15	0.6431	2	VP-1 maxHBint2

**Figure 11.** All Subset-Models calculation dialog

The best 15 models, according to the chosen fitness function, are displayed (Figure 11) to warn the user on the proceeding of the model calculation. To abort the model calculation the user can press the ESC key.

When the all subset calculations are completed, the variable selection by GA is executed (if previously selected in the variable setup), displaying the following dialog (Figure 12):




**Figure 12.** Genetic Algorithm-Models calculation dialog

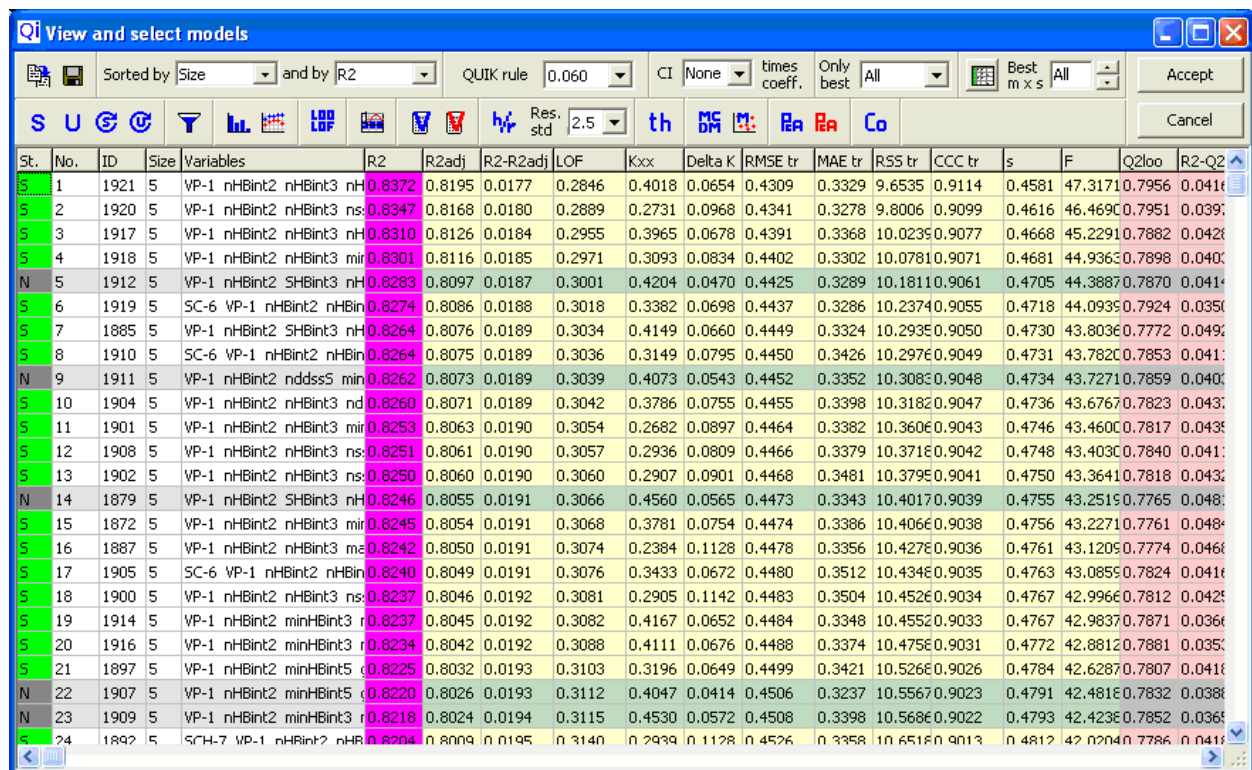
In this screen, the user can visualize the results of GA while it is running, selecting the “Display results while running” checkbox (Figure 12). When deactivated, it avoids QSARINS making calculations in organizing the data for the display, thus saving some time when there is no need to watch data (e.g. during overnight calculations). If activated, the “Display No.” and the “Display the best” list boxes become visible. The first list box is used to select the number of best model per size to be displayed (in the figure there is one model for every size till 4 variables). However, it is important to note that all the models for each size, according to the settings to “Mod. per size” in the “Variable selection” window, are saved and can be visualized. The second list box displays the number of best models to be visualized meanwhile they are calculated by the GA (in the figure there are 10 models of 5 variables size). If the calculations are automatically completed the user will be warned, otherwise it is possible to stop them at any time pressing the “Done” button.

## 5. View and select models

Once the models are calculated, they can be further explored selecting, from the main window (Figure 1), the following option:

Analysis -> View and select models, or click the icon 

The following window will appear (Figure 13a and 13b):





St.	No.	ID	Size	Variables	R2	R2adj	R2-R2adj	LOF	Kxx	Delta K	RMSE tr	MAE tr	RSS tr	CCC tr	s	F	Q2loo	R2-Q2
S	1	1921	5	VP-1 nHBint2 nHBint3 nH	0.8372	0.8195	0.0177	0.2846	0.4018	0.0654	0.4309	0.3329	9.6535	0.9114	0.4581	47.3171	0.7956	0.0411
S	2	1920	5	VP-1 nHBint2 nHBint3 ns	0.8347	0.8168	0.0180	0.2889	0.2731	0.0968	0.4341	0.3278	9.8006	0.9099	0.4616	46.4690	0.7951	0.0391
S	3	1917	5	VP-1 nHBint2 nHBint3 nH	0.8310	0.8126	0.0184	0.2955	0.3965	0.0678	0.4391	0.3368	10.0235	0.9077	0.4668	45.2291	0.7882	0.0421
S	4	1918	5	VP-1 nHBint2 nHBint3 mir	0.8301	0.8116	0.0185	0.2971	0.3093	0.0834	0.4402	0.3302	10.0781	0.9071	0.4681	44.9363	0.7898	0.0401
N	5	1912	5	VP-1 nHBint2 SHBint3 nH	0.8283	0.8097	0.0187	0.3001	0.4204	0.0470	0.4425	0.3289	10.1811	0.9061	0.4705	44.3887	0.7870	0.0411
S	6	1919	5	SC-6 VP-1 nHBint2 nHBin	0.8274	0.8086	0.0188	0.3018	0.3382	0.0698	0.4437	0.3286	10.2374	0.9055	0.4718	44.0935	0.7924	0.0351
S	7	1885	5	VP-1 nHBint2 SHBint3 nH	0.8264	0.8076	0.0189	0.3034	0.4149	0.0660	0.4449	0.3324	10.2935	0.9050	0.4730	43.8036	0.7772	0.0491
S	8	1910	5	SC-6 VP-1 nHBint2 nHBin	0.8264	0.8075	0.0189	0.3036	0.3149	0.0795	0.4450	0.3426	10.2976	0.9049	0.4731	43.7820	0.7853	0.0411
N	9	1911	5	VP-1 nHBint2 nddss5 min	0.8262	0.8073	0.0189	0.3039	0.4073	0.0543	0.4452	0.3352	10.3083	0.9048	0.4734	43.7271	0.7859	0.0401
S	10	1904	5	VP-1 nHBint2 nHBint3 nd	0.8260	0.8071	0.0189	0.3042	0.3786	0.0755	0.4455	0.3398	10.3182	0.9047	0.4736	43.6767	0.7823	0.0431
S	11	1901	5	VP-1 nHBint2 nHBint3 mir	0.8253	0.8063	0.0190	0.3054	0.2682	0.0897	0.4464	0.3382	10.3606	0.9043	0.4746	43.4600	0.7817	0.0431
S	12	1908	5	VP-1 nHBint2 nHBint3 ns	0.8251	0.8061	0.0190	0.3057	0.2936	0.0809	0.4466	0.3379	10.3716	0.9042	0.4748	43.4030	0.7840	0.0411
S	13	1902	5	VP-1 nHBint2 nHBint3 ns	0.8250	0.8060	0.0190	0.3060	0.2907	0.0901	0.4468	0.3481	10.3795	0.9041	0.4750	43.3641	0.7818	0.0431
N	14	1879	5	VP-1 nHBint2 SHBint3 nH	0.8246	0.8055	0.0191	0.3066	0.4560	0.0565	0.4473	0.3343	10.4017	0.9039	0.4755	43.2515	0.7765	0.0481
S	15	1872	5	VP-1 nHBint2 nHBint3 mir	0.8245	0.8054	0.0191	0.3068	0.3781	0.0754	0.4474	0.3386	10.4066	0.9038	0.4756	43.2271	0.7761	0.0481
S	16	1887	5	VP-1 nHBint2 nHBint3 m2	0.8242	0.8050	0.0191	0.3074	0.2384	0.1128	0.4478	0.3356	10.4276	0.9036	0.4761	43.1205	0.7774	0.0461
S	17	1905	5	SC-6 VP-1 nHBint2 nHBin	0.8240	0.8049	0.0191	0.3076	0.3433	0.0672	0.4480	0.3512	10.4346	0.9035	0.4763	43.0855	0.7824	0.0411
S	18	1900	5	VP-1 nHBint2 nHBint3 ns	0.8237	0.8046	0.0192	0.3081	0.2905	0.1142	0.4483	0.3504	10.4526	0.9034	0.4767	42.9966	0.7812	0.0421
S	19	1914	5	VP-1 nHBint2 minHBint3 r	0.8237	0.8045	0.0192	0.3082	0.4167	0.0652	0.4484	0.3348	10.4552	0.9033	0.4767	42.9837	0.7871	0.0361
S	20	1916	5	VP-1 nHBint2 minHBint3 r	0.8234	0.8042	0.0192	0.3088	0.4111	0.0676	0.4488	0.3374	10.4756	0.9031	0.4772	42.8812	0.7881	0.0351
S	21	1897	5	VP-1 nHBint2 minHBint5	0.8225	0.8032	0.0193	0.3103	0.3196	0.0649	0.4499	0.3421	10.5266	0.9026	0.4784	42.6287	0.7807	0.0411
N	22	1907	5	VP-1 nHBint2 minHBint5	0.8220	0.8026	0.0193	0.3112	0.4047	0.0414	0.4506	0.3237	10.5567	0.9023	0.4791	42.4816	0.7832	0.0381
N	23	1909	5	VP-1 nHBint2 minHBint3 r	0.8218	0.8024	0.0194	0.3115	0.4530	0.0572	0.4508	0.3398	10.5686	0.9022	0.4793	42.4236	0.7852	0.0361
S	24	1892	5	SC-7 VP-1 nHBint2 nHB	0.8204	0.8009	0.0195	0.3140	0.2999	0.1128	0.4526	0.3358	10.6516	0.9013	0.4812	42.0204	0.7786	0.0411

Figure 13a. Population of calculated models (with Fitting and some CV criteria)

RMSE cv	MAE cv	PRESS cv	CCC cv	Q2 LMO	R2 Yscr	Q2 Yscr	R2 Xmd	Q2 Xmd	R2 Ymd	Q2 Ymd	N. ext. OK	RMSE ext	MAE ext	PRESS ext	R2ext	Q2-F1	Q2-F2	Q2-F3	CCC ext
0.5433	0.4132	15.3489	0.8547	0.7961	0.0975	-0.1646	0.0963	-0.1578	0.0996	-0.1623	5	0.5061	0.3796	10.5026	0.7610	0.7600	0.7595	0.7754	0.8678
0.5448	0.4203	15.4326	0.8542	0.7198	0.1022	-0.2052	0.0977	-0.1550	0.0972	-0.2188	5	0.5147	0.3874	10.8611	0.7537	0.7518	0.7513	0.7677	0.8638
0.5830	0.4405	17.6746	0.8281	0.7453	0.0961	-0.2257	0.1015	-0.1500	0.0990	-0.2149	5	0.5222	0.3810	11.1800	0.7609	0.7446	0.7440	0.7609	0.8606
0.5050	0.3883	13.2607	0.8774	0.8064	0.0973	-0.1640	0.0970	-0.1560	0.0964	-0.1633	4	0.5598	0.4455	12.8482	0.7406	0.7064	0.7058	0.7252	0.8517
0.5140	0.3769	13.7378	0.8720	0.7991	0.1000	-0.1576	0.1006	-0.1512	0.0984	-0.1589	4	0.5010	0.3981	10.2915	0.7675	0.7649	0.7643	0.7799	0.8601
0.5429	0.4225	15.3292	0.8553	0.8038	0.0969	-0.1624	0.0967	-0.1567	0.1035	-0.1561	4	0.5317	0.4006	11.5917	0.7384	0.7352	0.7345	0.7521	0.8562
0.5442	0.4215	15.4025	0.8543	0.7985	0.0968	-0.1706	0.1013	-0.1510	0.0926	-0.1741	4	0.5246	0.3950	11.2849	0.7445	0.7422	0.7416	0.7587	0.8593
0.5379	0.4129	15.0481	0.8582	0.7920	0.0964	-0.1655	0.0987	-0.1522	0.0999	-0.1631	4	0.5180	0.4023	10.9999	0.7495	0.7487	0.7481	0.7648	0.8530
0.5380	0.4010	15.0497	0.8574	0.7963	0.0957	-0.1753	0.0962	-0.1586	0.0990	-0.1742	4	0.5183	0.3946	11.0124	0.7483	0.7484	0.7478	0.7645	0.8542
0.5408	0.4063	15.2060	0.8553	0.7566	0.0933	-0.2939	0.0959	-0.1581	0.0974	-0.2734	4	0.5302	0.4107	11.5242	0.7409	0.7367	0.7361	0.7535	0.8500
0.5455	0.4032	15.4734	0.8528	0.7920	0.0964	-0.1655	0.0987	-0.1522	0.0999	-0.1631	4	0.5314	0.4084	11.5769	0.7353	0.7355	0.7349	0.7524	0.8501
0.5830	0.4530	17.6748	0.8287	0.7862	0.0981	-0.1670	0.0995	-0.1540	0.0989	-0.1638	4	0.5442	0.4014	12.1425	0.7404	0.7226	0.7219	0.7403	0.8517
0.4942	0.3796	12.6992	0.8831	0.7915	0.0958	-0.1706	0.0984	-0.1550	0.0998	-0.1617	3	0.5546	0.4452	12.6099	0.7284	0.7119	0.7112	0.7303	0.8483
0.4977	0.3869	12.8819	0.8813	0.7900	0.0936	-0.1697	0.0983	-0.1536	0.0952	-0.1664	3	0.5600	0.4407	12.8596	0.7101	0.7062	0.7055	0.7250	0.8375
0.5051	0.3891	13.2646	0.8775	0.7862	0.0981	-0.1670	0.0995	-0.1540	0.0989	-0.1638	3	0.5548	0.4449	12.6192	0.7147	0.7117	0.7110	0.7301	0.8353
0.5022	0.3850	13.1161	0.8793	0.7905	0.0993	-0.1570	0.0949	-0.1594	0.0953	-0.1629	3	0.5498	0.4179	12.3934	0.7196	0.7168	0.7162	0.7349	0.8413
0.5086	0.3915	13.4512	0.8757	0.7954	0.0976	-0.1622	0.0915	-0.1632	0.0986	-0.1641	3	0.5422	0.4240	12.0516	0.7258	0.7246	0.7240	0.7423	0.8442
0.5134	0.3888	13.7063	0.8730	0.7854	0.1003	-0.1614	0.1034	-0.1479	0.0996	-0.1615	3	0.5621	0.4403	12.9553	0.7052	0.7040	0.7033	0.7229	0.8324
0.5078	0.3813	13.4093	0.8757	0.7954	0.0976	-0.1622	0.0915	-0.1632	0.0986	-0.1641	3	0.5614	0.4257	12.9240	0.7064	0.7047	0.7040	0.7236	0.8340
0.5176	0.3986	13.9303	0.8708	0.7959	0.0979	-0.1612	0.0939	-0.1610	0.0990	-0.1601	3	0.5624	0.4171	12.9677	0.7050	0.7037	0.7030	0.7227	0.8330
0.5140	0.3945	13.7368	0.8729	0.7916	0.0947	-0.1693	0.0975	-0.1564	0.1003	-0.1612	3	0.5431	0.4303	12.0940	0.7250	0.7237	0.7230	0.7414	0.8452
0.5168	0.3954	13.8871	0.8706								3	0.5447	0.4137	12.1651	0.7233	0.7221	0.7214	0.7398	0.8448
0.5178	0.3831	13.9441	0.8696								3	0.5222	0.4191	11.1805	0.7450	0.7446	0.7440	0.7609	0.8486
0.5170	0.3795	13.8998	0.8702	0.7896	0.0960	-0.1657	0.0992	-0.1532	0.0972	-0.1623	3	0.5621	0.4314	12.9541	0.7081	0.7040	0.7033	0.7230	0.8346


**Figure 13b.** Population of calculated models (with some CV and external criteria)

In these screens, QSARINS visualizes the population of all the calculated models, ordered by the selected criterion ( $R^2$ ,  $Q_{LOO}^2$  and so on,  $R^2$  in Figure 13a), using the “Sorted by” list box. The column of the selected criterion for sorting the models will be colored (magenta). The columns of the validation criteria are differently colored: yellow for fitting, pink for CV and cyan for external (Figure 13a and 13b). During the all subset and GA calculation, the models are labelled with a unique ID (see ID column in Figure 13a) in order to make easier their identification both on the tabulated data and in the graphs.

The user can select some or all the models in the list and their data can be copied in the clipboard (icon  or “Copy” from the popup menu) or saved to a file (icon , or “Save...” from the popup menu). In order to avoid repetitions, it is here recalled that all the icons have a corresponding voice in the popup menu).


When sorting per model size (“Size” in the list box) or for the number of external validation criteria, (calculated in Section 6 “Model validation”) that are fulfilled (“N. ext. OK” in the list box), the second list box on the right is activated (“and by” list box). This allows to sort the models according to the chosen



criteria (e.g.  $Q_{LOO}^2$  in the “and by” list box), for instance considering firstly the model size. The user can also filter the calculated models by applying the QUIK rule, if not selected before in the variable selection dialog, or a more restrictive QUIK rule if it has been previously selected (see section 4.1). The value of the QUIK rule can be changed by selecting the appropriate value from the list box. “Bad” models according to the QUIK rule are automatically unselected, and can be deleted from the list by clicking the filter option (icon , see the table below for further details).

Pressing the “CI” list box it is possible to filter the models according to the magnitude of the interval of confidence and significance of the coefficients of models. If the ratio between the interval of confidence and the coefficient of one of the model descriptors, or the intercept, is greater than the one selected from the list box (from 1.0 to 1.5), then the model will be automatically unselected. The same happens if the significance of one of the coefficient, or the intercept, is greater than 0.05. By double clicking on the model the user can know, in more details, which are and to what extend, the unreliable coefficients (See Section 7, “ Analysis of single models” for more details).

The “Only best” options allows to display only a certain number of models, according to the selected criteria in the “Sorted by” (or the “and by”) option, deleting the other models.

During the model selection (either manually and/or by the QUIK rule/CI filter), usually, a certain number of models are selected for further analysis, while others are not. To see the list of the only selected ones the user must press the  icon (pressing again the icon will restore the standard model view). If needed, it is possible to display only the best “m” number of models per model size, selecting the desired number in the “Best m x s” box (“m x s” means models per size). This will be applied only to the selected model, the unselected ones will not be considered.


The subsequent row of icons in Figures 13 allows further operations on the models.


The first four icons are used for manual selection of the models:



**S** : selects the models to be used for filtering, further analysis, validation, PCA, consensus, etc.



 : unselects models.

 : selects all models.


 : unselects all models.

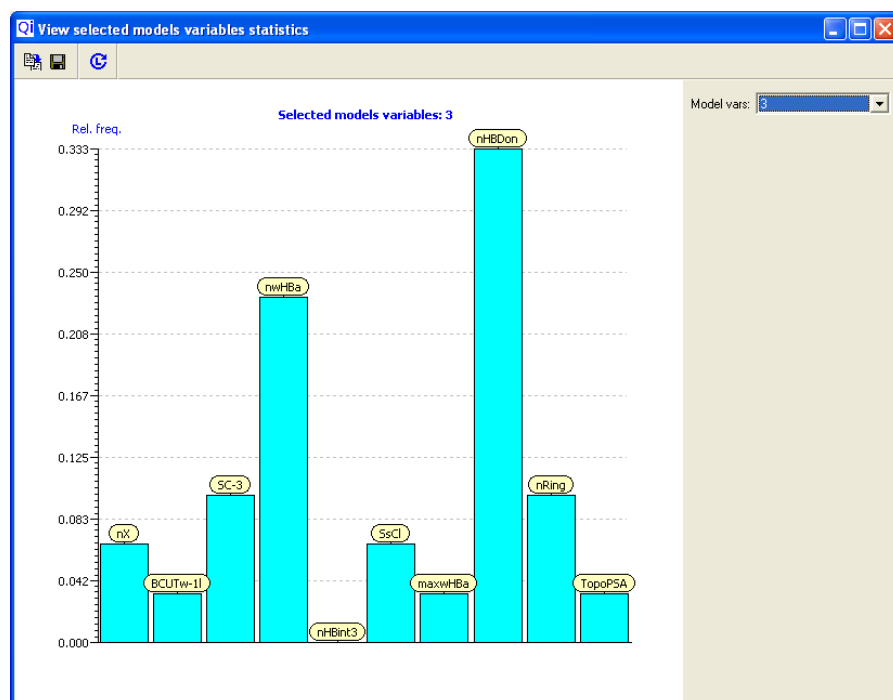
Then other icons are for various tools.



 : deletes unselected models (either unselected (U status) by the QUIK rule, the CI filter or manually by the user pressing the icon  ).




 : shows the relative frequency of descriptors in the population of calculated models, as in the following example (Figure 14), where the bars are the relative frequency of the descriptors used by all the models having a size of 3 variables:

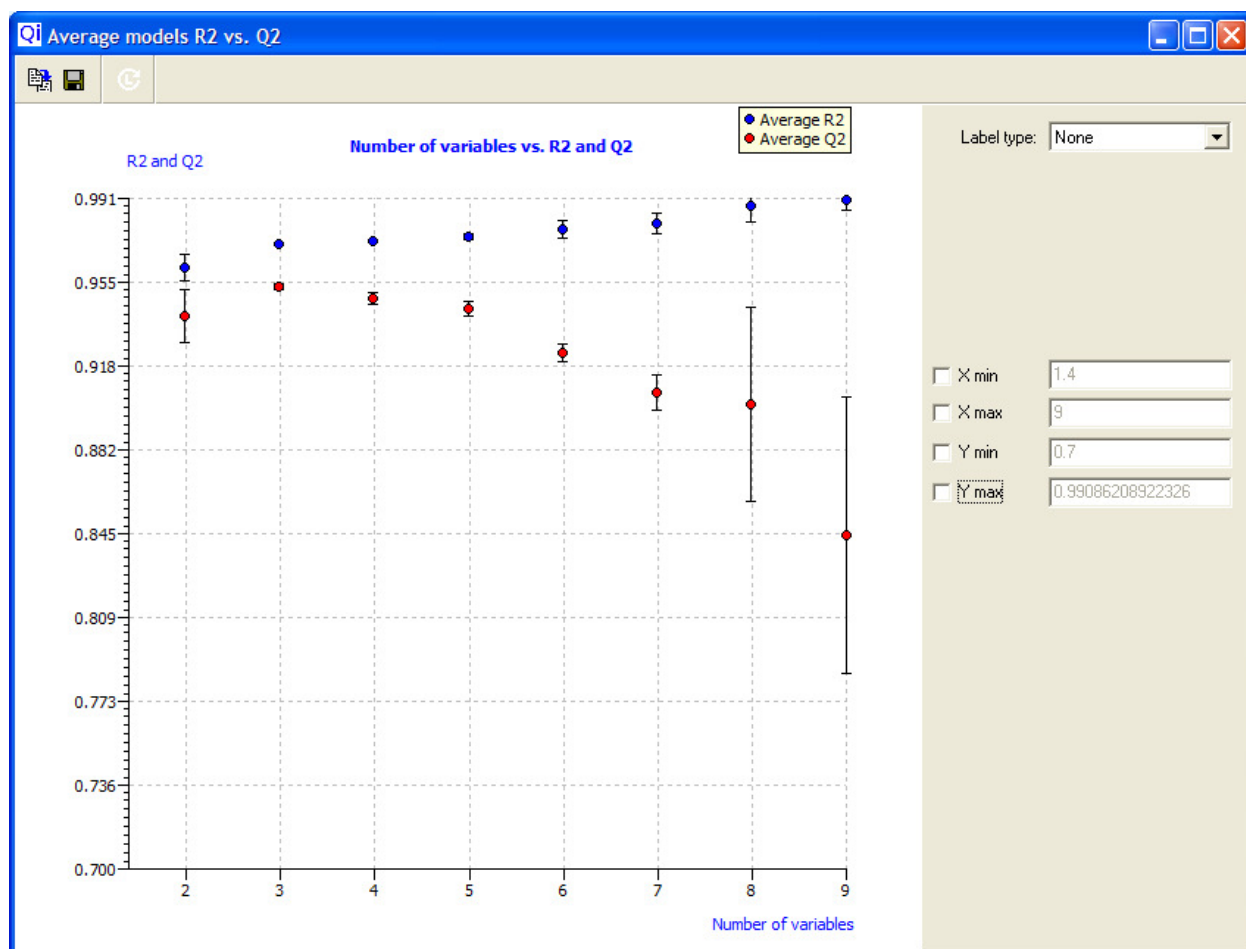


**Figure 14.** Relative frequency of descriptors in the population of calculated models

This tool is informative on the most modeling descriptors in the resulting model population.



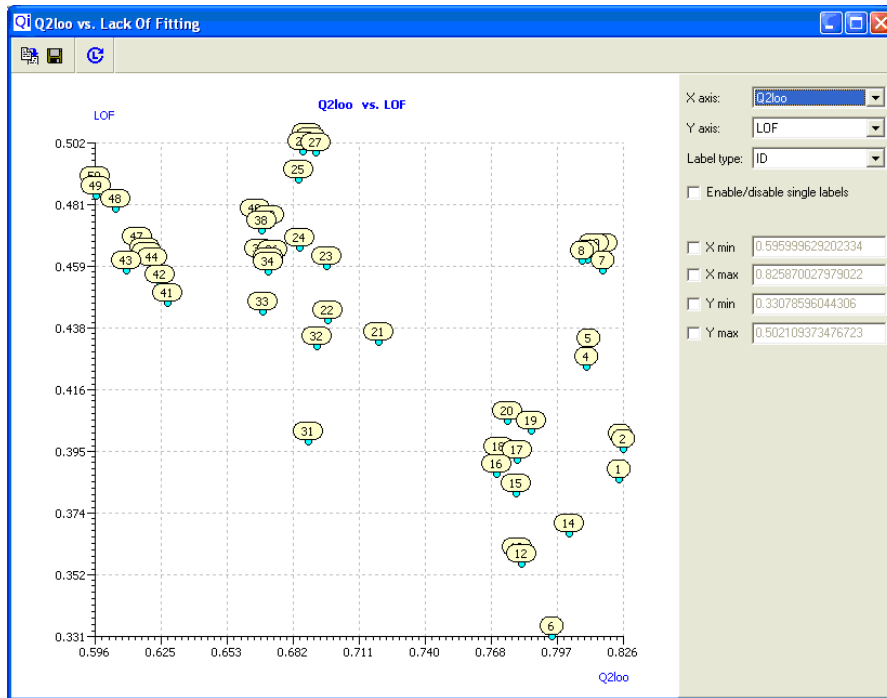
 : plots the average values (with their standard deviation) of  $R^2$  and  $Q^2_{LOO}$  in relation to the number of modeling variables. This is used to evaluate the model performances vs. the size of the developed models, and whether exists any overfitting in the resulting populations of models. In fact, it is known that  $R^2$  values increase when one descriptor is added to the previous one, while  $Q^2_{LOO}$  values only increase until useful descriptors are added. Thus, if “noisy” descriptors are added to the model (from 4 variables in Figure 15),  $Q^2_{LOO}$  will decrease, showing that the predictive power of the model could be deceptive.



**Figure 15.** Plot of  $R^2$  and  $Q^2_{LOO}$  average values vs. number of modeling variables




**LOF**: plots the values of  $Q^2_{LOO}$  vs.  $LOF$ . This is used to evaluate the models performances in predictions ( $Q^2_{LOO}$ ) vs. the performances in fitting ( $LOF$ , which accounts also for the number of descriptors used in the models).

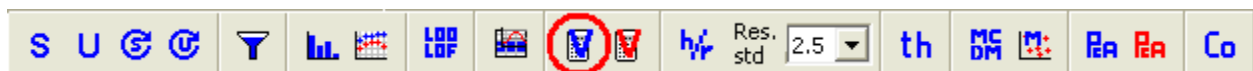



**Figure 16.** Plot of  $Q^2_{LOO}$  vs LOF

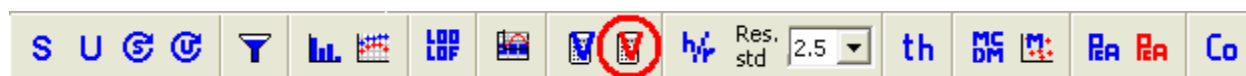
This graph (Figure 16) can help in the selection of the models with the best compromise between high predictivity (high  $Q^2_{LOO}$ ) and small dimension (low LOF) (as, for example, the model 6 in the figure above).




: visualizes the performances of the selected model, both in tabulated and in graphical form, as explained in Section 7 ("Analysis of Single Models").




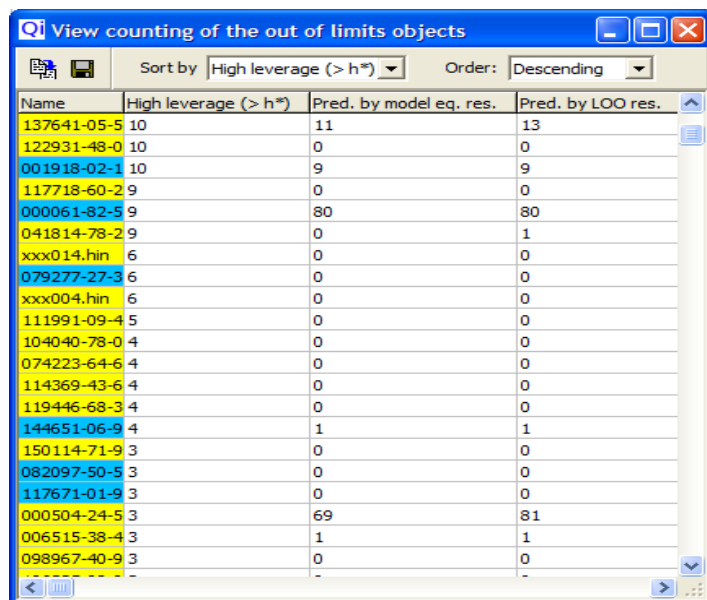
: applies the internal validation tools (LMO, Y-scrambling and random responses/descriptors techniques) on the selected models: see section 6 (“Models validation”) for further details.



: applies the external validation tools ( $R_{EXT}^2$ ,  $Q_{F1}^2$ ,  $Q_{F2}^2$ ,  $Q_{F3}^2$ ,  $CCC$ ,  $\overline{r_m^2}$  and  $\Delta r_m^2$ ) on the selected models: see section 6 (“Models validation”) for further details.



: calculates and count the structural outliers (X-outliers, i.e. those with leverage values above the  $h^*$  value (see Section 7, “Analysis of single Models”) and the residual outliers (Y-outliers, i.e. those with predictions above a user defined standardized residual threshold, in this case set using the “Res. std” list box). When the icon is pressed the following window appears (Figure 17):



Name	High leverage (> $h^*$ )	Pred. by model eq. res.	Pred. by LOO res.
137641-05-5	10	11	13
122931-48-0	10	0	0
001918-02-1	10	9	9
117718-60-2	9	0	0
000061-82-5	9	80	80
041814-78-2	9	0	1
xxx014.hin	6	0	0
079277-27-3	6	0	0
xxx004.hin	6	0	0
111991-09-4	5	0	0
104040-78-0	4	0	0
074223-64-6	4	0	0
114369-43-6	4	0	0
119446-68-3	4	0	0
144651-06-9	4	1	1
150114-71-9	3	0	0
082097-50-5	3	0	0
117671-01-9	3	0	0
000504-24-5	3	69	81
006515-38-4	3	1	1
098967-40-9	3	0	0

**Figure 17.** View counting of the out of limits objects.

The first column in Figure 17 reports the name of the chemicals (CAS number in this example). The other columns report the number of models in which each chemical is classified as X-outlier (column “High leverage ( $>h^*$ )”) and Y-outlier for responses predicted by model equation and by LOO (columns “Pred. by model eq. res.” and “Pred. by LOO res.”, respectively). All chemicals are colored according to the status: yellow for training set, blue for prediction set and red if the experimental response is unknown.

In the above example: the first row means that for the chemical “137641-05-5” there are 10 models having it as an X-outlier, 11 models predicting it Y-outlier by applying the model equation and 13 models predicting it as Y-outlier by LOO. This tool is useful in highlighting possible common outliers that the GA population of descriptors selected by GA is not able to model and that could be deleted from the dataset, because they are out of the applicability domain of the population of models.

Once this window is closed, the X- and Y- outliers will be counted for every model in the list of the main window, as in the following example (Figure 18):

N. Y-outl. und. mod. eq.	N. Y-outl. over. mod. eq.	N. Y-outl. und. LOO	N. Y-outl. over. LOO	N. X-outl.	N. out AD
2	0	2	0	1	1
3	0	3	0	4	5
3	0	3	0	4	5
2	0	2	0	1	1
2	0	2	0	1	1
3	2	3	2	7	7
3	2	3	2	7	7
3	0	3	0	6	7
3	0	3	0	6	7
4	0	4	0	2	3
4	0	4	0	2	3
3	2	3	2	5	5
3	2	3	2	5	5
3	0	3	0	2	2
3	0	3	0	2	2
2	0	3	0	0	0

**Figure 18.** Tabulated values of structural and response outliers for each model.

Each row is a different model while the columns have the following meaning:

“N. Y-outl. und. mod. eq.” is the number of Y-underestimated outliers (i.e. those with the standardized residual smaller than the threshold value set by the user) calculated using the model equation.

“N. Y-outl. over. mod. eq.” is the number of Y-overestimated outliers calculated using the model equation.

“N. Y-outl. und. LOO” and “N. Y-outl. over. LOO” have the same meaning as the two above, except that responses are calculated using LOO predictions instead of using the model equation.

For a precautionary approach, in cases of toxicity models, it is suggested to select, among models with similar performances, those with the least number of chemicals underestimated as less toxic.

“N. X-outl.” is the number of structural outliers using training and prediction chemicals while “N. out AD” is the same, but adding the prediction for the compounds without the experimental response (Unknown).

For a more reliable modeling, it is suggested to select, among models with similar performances, those with the least number of chemicals out of AD (thus models with highest number of interpolated predicted data and lowest number of extrapolated predictions).



The icons **th**, **MC DM** and **M** are related also to the external validation parameters and will be explained in the Section 6 (“Model validation”).




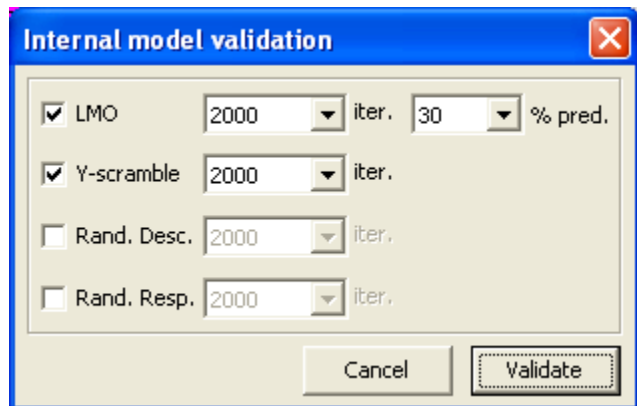
The icons **RA** and **RA** are used for the selection of the most diverse models to be averaged in the Consensus modeling **Co**, see section 8 (“Consensus modeling”) for further details.

## 6. Model validation

During model calculation, only the cross validated and external validation criteria that require few calculations (e.g. RMSE, MAE and PRESS) are automatically computed and included into the output. If additional criteria, more demanding in terms of calculations, are required, the following dialog (Figure 19) for internal validation can be visualized (automatically) by the single model option from the Data



setup (see section 3) or (manually, pressing the  icon, or clicking the corresponding popup menu item) from the “View and select models” dialog (see section 5). The first box contains four techniques for internal validations that require a high number of iterations (this is why they are optional).




**Figure 19.** Dialog for internal validation

The first one is the Leave More (or Many) Out (LMO) technique (Wehrens et al., 2000), that iteratively excludes at random, from the training set, a certain percentage of chemicals to be used for predictions. The number of iterations can be selected using the “iter” list box, while the percentage of prediction elements (over the entire training set) can be selected using the “% pred.” list box (set as 30% by default). The average results of the procedure will be found as  $Q_{LMO}^2$  [Q2LMO] (squared brackets are here used for the criteria names used in QSARINS). Those values must be significantly near to  $Q_{LOO}^2$  of the model, and the randomized values of  $Q_{LMO}^2$ , of all the iterations, must be not too disperse.

The “Y-scramble” list box calculates iteratively a certain number of models (see corresponding “iter.” list box) shuffling at random the experimental responses (Rücker et al., 2007). The results ( $R_{Y-SCRAMBLE}^2$  and  $Q_{Y-SCRAMBLE}^2$ ) will be stored as [R2Yscr] and [Q2Yscr]. The  $R^2$  and  $Q_{LOO}^2$  of the model under test must be reasonably higher than the scrambled ones, to exclude chance correlation in the original model.


A similar concept is applied to the “Rand. Desc.” and “Rand. Resp.” options. In the first case the responses are iteratively generated at random within the range of the true responses, while in the

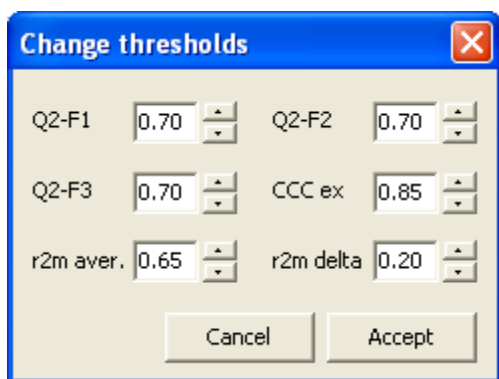
second case the descriptors are iteratively generated at random within a reasonable range of true descriptors (Rücker et al., 2007). As before, in both cases the  $R^2$  and  $Q_{LOO}^2$  of the model under test must be reasonably higher than the ones of the random models ( $R_{RND-RESP}^2$  [R2Yrnd],  $Q_{RND-RESP}^2$  [Q2Yrnd] and  $R_{RND-DESCR}^2$  [R2Xrnd],  $Q_{RND-DESCR}^2$  [Q2Xrnd]). The second case will exclude that the original modeling descriptors are just numbers selected by chance, thus without any structural meaning.

Pressing the  button (or clicking the corresponding popup menu item), it is possible to perform the external model validation (that will be asked automatically if we are calculating a single model during the data setup, see section 3) of the following validation criteria:  $R_{EXT}^2$  [R2ext](external determination coefficient),  $Q_{F1}^2$  (Shi et al., 2001)[Q2-F1],  $Q_{F2}^2$  (Schüürmann et. al, 2008)[Q2-F2],  $Q_{F3}^2$  (Consonni et al. 2009, 2010) [Q2-F3], CCC (Lin, 1989)(Chirico and Gramatica, 2011, 2012) [CCC ext],  $\overline{r_m^2}$  (Ojha et al., 2011) [r2m aver.] and  $\Delta r_m^2$  (Ojha et al., 2011)[r2m delta].

In relation to the external validation criteria, in the “View and select models dialog:



the icon  visualizes the following dialog (Figure 20) concerning the external validation criteria thresholds.



**Figure 20.** Dialog for setting the external validation thresholds


The user can then set the thresholds of the external validation criteria ( $Q_{F1}^2$  [Q2-F1],  $Q_{F2}^2$  [Q2-F2],  $Q_{F3}^2$  [Q2-F3], CCC [CCC ext],  $\overline{r_m^2}$  [r2m aver.] and  $\Delta r_m^2$  [r2m delta]) here proposed and applied, as in a recent paper, where comparable thresholds were defined (Chirico and Gramatica, 2012).

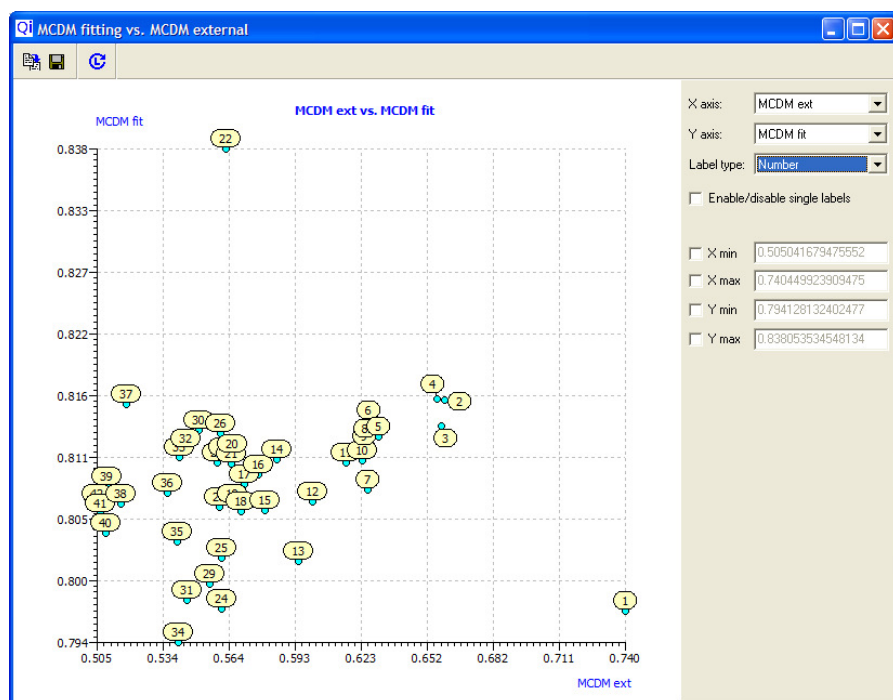
These values for the external parameters are used in counting how many external validation parameters are fulfilled by the model, this number is found in the “N. ext. OK” column of the models list in the “View and select models” main window



: pressing this icon the MultiCriteria Decision Making score will be calculated.

The Multi-Criteria Decision Making (MCDM)(Keller et al., 1991) is a technique that summarizes the performances of a certain number of criteria simultaneously, as a single number (score) between 0 and 1. This is done associating to every validation criteria a desirability function which values range from 0 to 1 (where 0 represents the worst validation criteria value and 1 the best). The geometric average of all the values obtained from the desirability functions gives the MCDM value. By default, pressing the Calculate button, the MCDM of fitting (maximizing  $R^2$ ,  $R_{adj}^2$  and  $CCC_{TR}$ , while minimizing  $R^2 - R_{adj}^2$ ), cross validation (maximizing  $Q_{LOO}^2$ ,  $Q_{LMO}^2$  and  $CCC_{cv}$ , while minimizing  $R_{Y-SCRAMBLE}^2$ ) and external validation (maximizing  $Q_{F1}^2$ ,  $Q_{F2}^2$ ,  $Q_{F3}^2$  and  $CCC_{EXT}$ ), are automatically calculated using all the corresponding criteria. If one or more of the criteria are not available (the user will be warned of this fact), the corresponding MCDM will not be calculated. The scores will be displayed in the model list respectively as: MCDM fit (fitting), MCDM cv (cv = cross validation = internal validation), MCDM ext (external validation) and MCDM all (calculated using all the criteria used for MCDM : fitting, cross validated and external). The model with the best MCDM compromise among the selected validation criteria will be sorted as the best, in the “View and selects models” window using the “Sorted by” option.

: this option plots the values of MCDM for external prediction vs. MCDM on fitting, as in the following example (Figure 21):



**Figure 21.** Plot of MCDM ext vs. MCDM fit

In Figure 21, the model 22 has the best fitting performances, but it is not externally predictive, while for instance model 1 has the better compromise between fitting and predictivity.


## 7. Analysis of single models

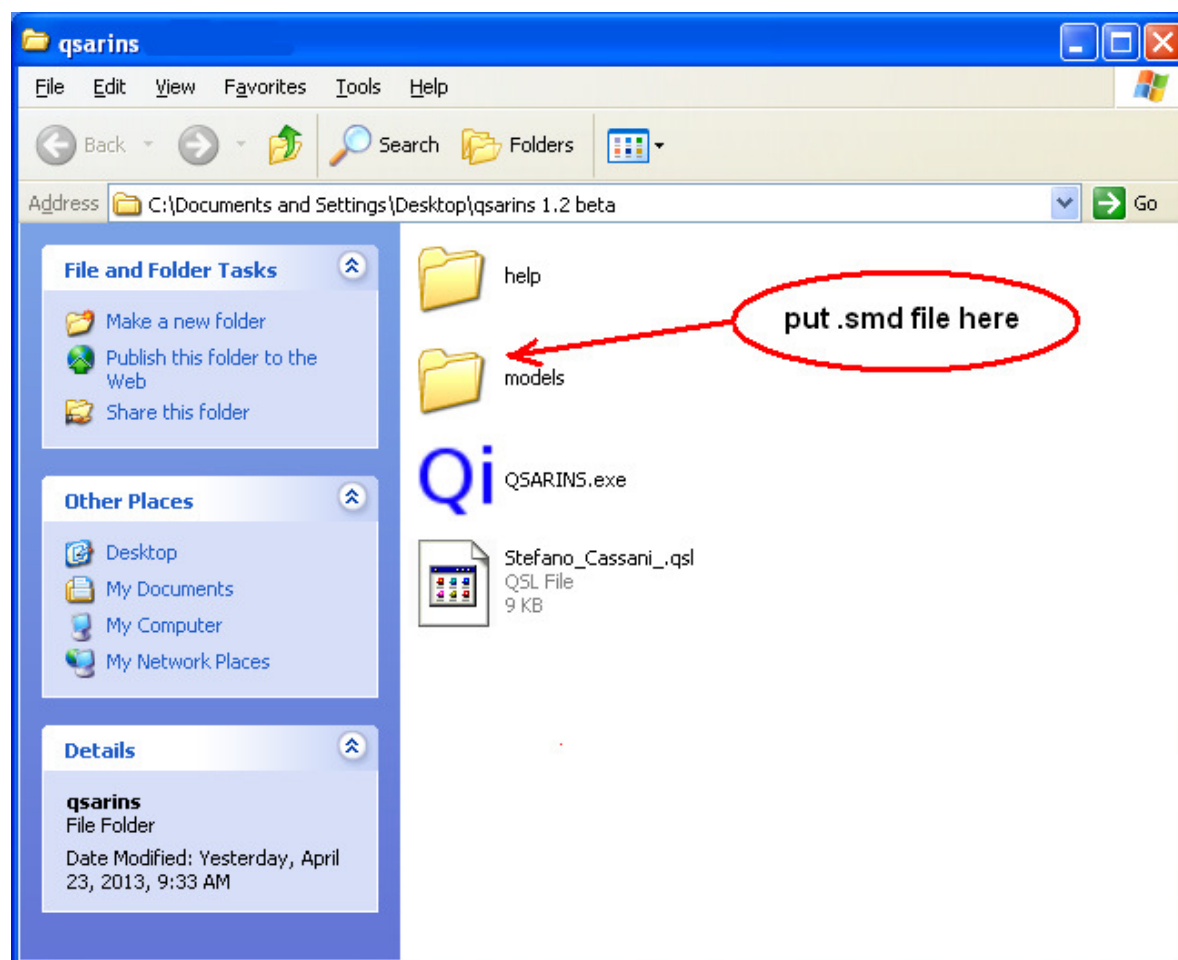
This option is used to explore a QSAR model in full details (Gramatica et al., 2012), both in tabulated and in graphical form, to assess its performances. This function can be accessed in two following alternative ways:

1. Click the button “Single model” in Data setup dialog (Section 3) to analyze a model based on the selected descriptors in that step.
2. Select one model from the list of calculated ones, as reported in section 5, “View and select models”



Here it follows an example of the single model dialog box (Figure 22):



The save model option (  ) proposes two files type: \*.txt and \*.smd. The first one saves the model data as displayed in the dialog, using a text file format, while the second one saves the model itself, as a custom file (\*.smd), to be applied later using new chemicals (see also Section 9.1, “Apply developed model”). In order to be available to QSARINS, these files *must* be located in the “models” subfolder of the main QSARINS folder, as exemplified in the following figure (Figure 23):



**Figure 23.** Single model file (.smd) folder location

If the user wish to add its personal QMRF as a .pdf file, is suffices it has the same name of the model and is put in the same folder (e.g. if the model is PBT.smd, the pdf file must be PBT.pdf). This also applies for the .sdf files. The  and  icons will be activated at the next launch of QSARINS when the single model data are requested.

The upper data grid in Figure 22 displays the model equation coefficients, the standardized coefficients, standard error, the confidence intervals (Co. int. 95%) of the regression coefficients and p-value. If the ratio of the interval of confidence and the descriptor coefficient (or the intercept) is greater than 1, the corresponding cell will be highlighted in red (as in figure 22). This is because when such ratio is considered (arbitrarily) too high, the coefficient/intercept, thus as a consequence the model, should be considered suspect. The same happens for the significance (p-value) of the coefficients/intercept (that are related to the corresponding interval of confidence). If their value is too high, in this case a p-value greater than 0.10, the corresponding cells are highlighted in red. If the value is considered borderline, i.e. between 0.05 and 0.10, the corresponding cell is highlighted in yellow, and the models should be considered with caution.

The middle data grid contains the statistics concerning the model performances, organized in the following manner (statistics name in QSARINS are reported in squared brackets):

**Fitting criteria:**  $R^2$  [R2],  $R_{adj}^2$  [R2adj],  $R^2 - R_{adj}^2$  [R2-R2adj],  $LOF$  (Friedman lack of fit criteria) [LOF],  $K_x$  [Kxx] (inter-correlation among descriptors),  $\Delta K$  (difference of the correlation among the descriptors ( $K_x$ ) and the descriptors plus the responses ( $K_{xy}$ )) [Delta K], RMSE [RMSE tr], MAE [MAE tr], RSS [RSS tr], CCC [CCC tr], s [s] and F [F].

**Internal validation criteria:**  $Q_{LOO}^2$  [Q2loo],  $R^2 - Q_{LOO}^2$  [R2-Q2loo], RMSE [RMSE cv], MAE [MAE cv], PRESS [PRESS cv], CCC [CCC cv],  $Q_{LMO}^2$  [Q2LMO],  $R_{Y-SCRAMBLE}^2$  [R2Yscr],  $Q_{Y-SCRAMBLE}^2$  [Q2Yscr],  $R_{RND-DESCR}^2$  [R2Xrnd],  $Q_{RND-DESCR}^2$  [Q2Xrnd],  $R_{RND-RESP}^2$  [R2Yrnd],  $Q_{RND-RESP}^2$  [Q2Yrnd].

**External validation criteria:** RMSE [RMSE ext], MAE [MAE ext], PRESS [PRESS ext],  $R_{EXT}^2$  [R2ext],  $Q_{F1}^2$  [Q2-F1],  $Q_{F2}^2$  [Q2-F2], and  $Q_{F3}^2$  [Q2-F3], CCC [CCC ext],  $\overline{r_m^2}$  [r2m aver.] and  $\Delta r_m^2$  [r2m delta].

After the external validation criteria, the angle of the regression line of the external data points respect the diagonal (perfect agreement between experimental and predicted data) is reported. All statistics for evaluating the Golbraikh and Tropsha method (Golbraikh and Tropsha, 2002), and details of the  $r_m^2$  statistics, are reported, both for predictions by LOO and from predictions on the external data set (if available) as in the following example:

Predictions by LOO:


Exp(x) vs. Pred(y): R2: 0.7928 R'2o: 0.7611 k': 0.9894 Clos': 0.0400 r'2m: 0.6516

Pred(x) vs. Exp(y): R2: 0.7928 R2o: 0.7912 k: 0.9991 Clos: 0.0020 r2m: 0.7610

Obs(x) vs. Pred(y) means that the experimental data are on the abscissa (x) and the predicted data on the ordinate (y), the opposite is applied for Pred(x) vs. Exp(y). R2 is the value of  $R^2$  calculated for the experimental vs. predicted regression line, R'2o is calculated forcing the regression line to pass through the origin ( $R_0'^2$ ), k' is the slope of the regression line, Clos. is the closeness between  $R_0'^2$  and  $R^2$ . Those statistics, excepts  $r_m^2$ , can be used for the Golbraikh and Tropsha method for model evaluation. All statistics without the ' (apostrophe) symbol are calculated exchanging the axis.

The "Std. res. thresh." list box is the threshold value, in standardized residuals units, used for determining the response-outliers (Y-outliers).

The last data grid in Figure 22 shows the details of the chemicals used in model calculation, providing the following information: ID, Name, Status (training, prediction, excluded), experimental endpoint value ("Exp. endpoint"), predicted value using the model equation ("Pred. by model eq."), residual calculated using the model equation ("Pred.Mod.Eq.Res."), predicted value calculated using cross validation ("Pred. LOO"), residual calculated using cross validation ("Pred. LOO Res."), leverage values (diagonal elements of the HAT matrix "HAT i/i" highlighting those with h value >  $h^*$ , defined as  $3p'/n$ , where  $p'$  is the number of the model variables + 1 and n is the number of training compounds) (Atkinson, 1985, Gramatica, 2007), the standardized residuals both predicted from the model ("Std.Pred.Mod.Eq. Res.") and predicted by cross validation (Std.Pred.LOO Res.), highlighting those above or below the standardized residual threshold ("Std. res. thresh." list box).

Note that after pressing the icon  and visualizing the Score and Loading plot related to the PCA analysis of the model descriptors, new columns containing the PC Scores for every compound became available at the right of the "Std.Pred.LOO Res." column (Figure 24).



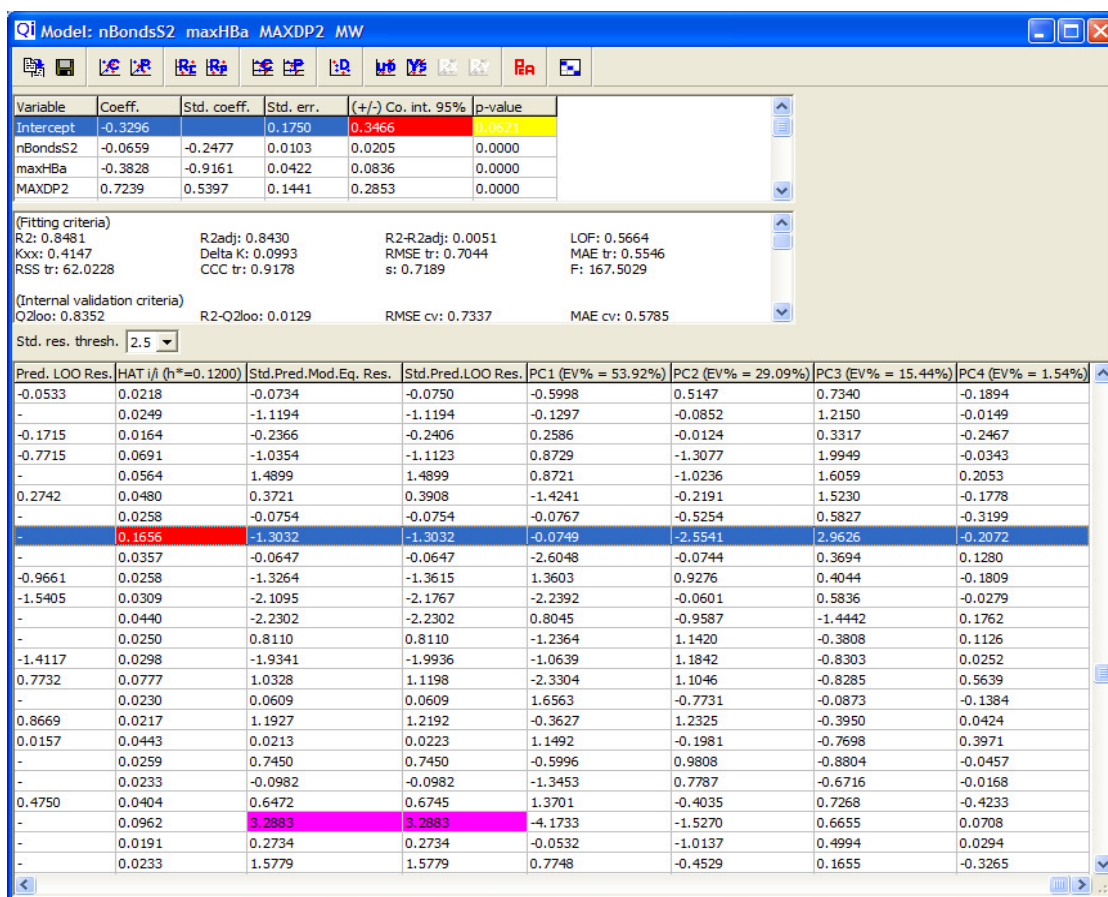


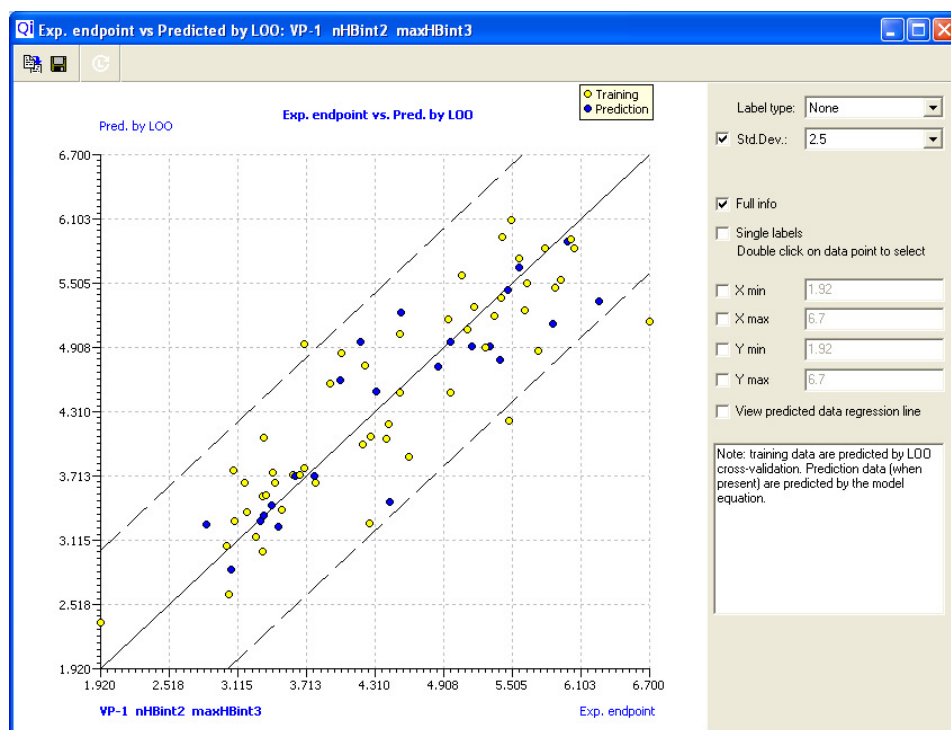




Figure 24. Single model dialog box, where PC columns became available



In addition to the aforementioned tabulated data, various graphs can be visualized, as explained below.

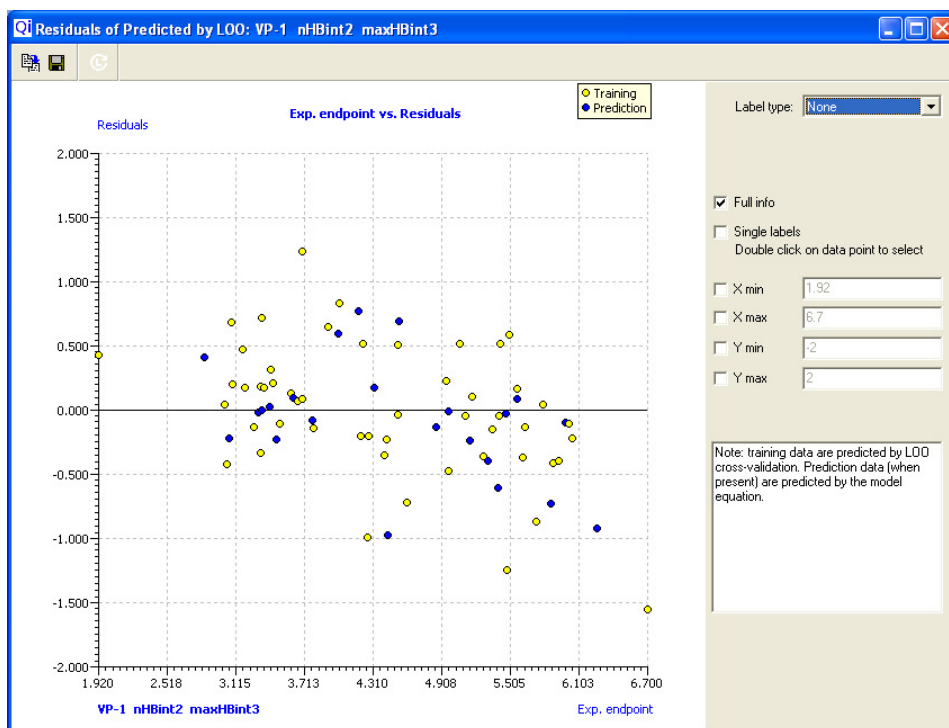
	Scatter plot of Experimental endpoint vs. Predictions by model equation
	Scatter plot of Experimental endpoint vs. LOO predictions (Figure 25)




**Figure 25.** Scatter plot of experimental vs. predicted data by LOO (press  icon)



On the abscissa the experimental endpoint data are used, while on the ordinate the prediction by LOO are used. On the ordinate axis, the yellow points (training set) are calculated using the LOO predictions while the blue points (prediction set) are calculated using the model equation. If the user press the  instead, the values for training set (yellow points) are different because are calculated using the model equation instead of the LOO predictions, while those for the prediction set (blue points) are the same. “Std. Dev.” option allows to visualize and manage the Standard Deviations in this graph.

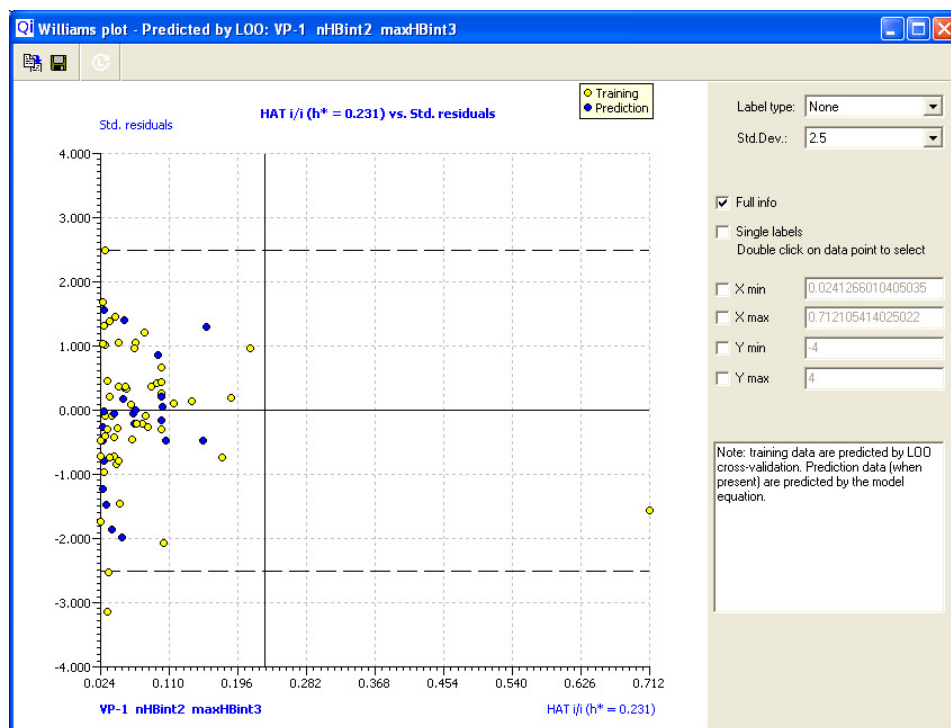
	Residual plot of Experimental endpoint vs. residuals from the predictions by model equation
	Residual plot of Experimental endpoint vs. residuals from the LOO predictions (Figure 26)



**Figure 26.** Scatter plot of the residuals of the predicted data by LOO (press  icon)

On the abscissa axes the values of the experimental endpoints are reported, while on the ordinate the values of the residuals of the predictions are reported. As in the previous figure the yellow data points (training set) are the values predicted by LOO, while the blue ones (prediction set) are calculated using the model equation.


	Williams plot of diagonal hat elements vs. standardized residual predictions
	Williams plot of diagonal hat elements vs. standardized residual predictions by LOO (Figure 27)

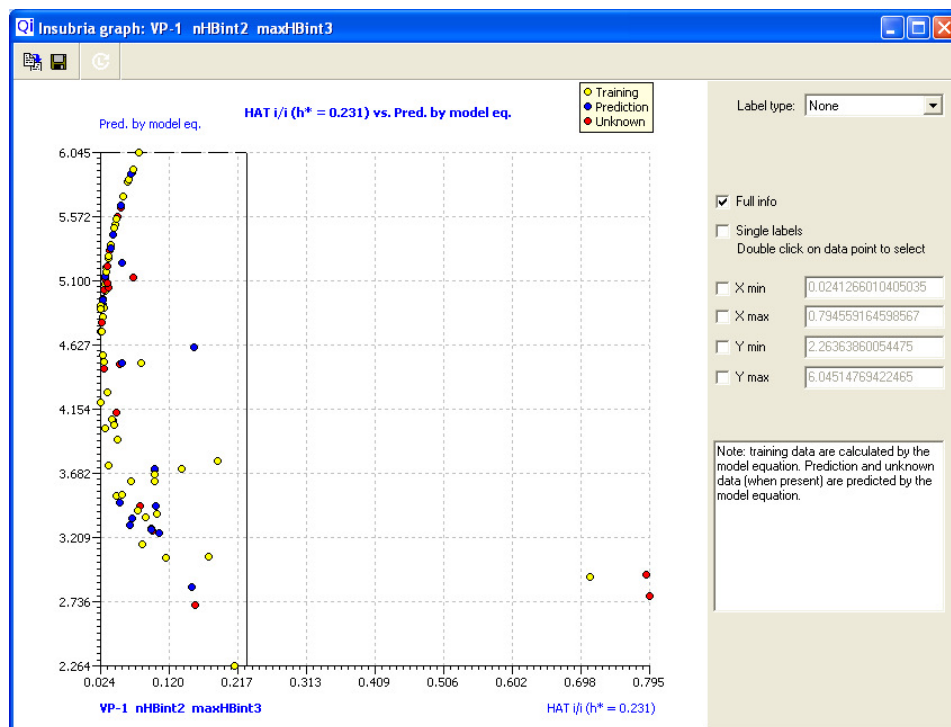



**Figure 27.** Williams plot using data predicted by LOO (press  icon)

When residuals cannot be calculated (this sometimes happens for molecules having a very high HAT value) QSARINS warns the user that these data points cannot be displayed.





On the abscissa the HAT values of the diagonal elements are reported, while the standardized residuals of the predictions (yellow data points, training set or for the LOO and blue, prediction set, for the ones calculated using the model equation) are on the ordinate. The vertical line corresponds to the HAT threshold value of the structural domain ( $h^*$ ), while the dashed horizontal ones are the user defined threshold (“ResLine” list box in the figure) for Y-outliers. “Std. Dev.” option allows managing the Standard Deviations in this graph.

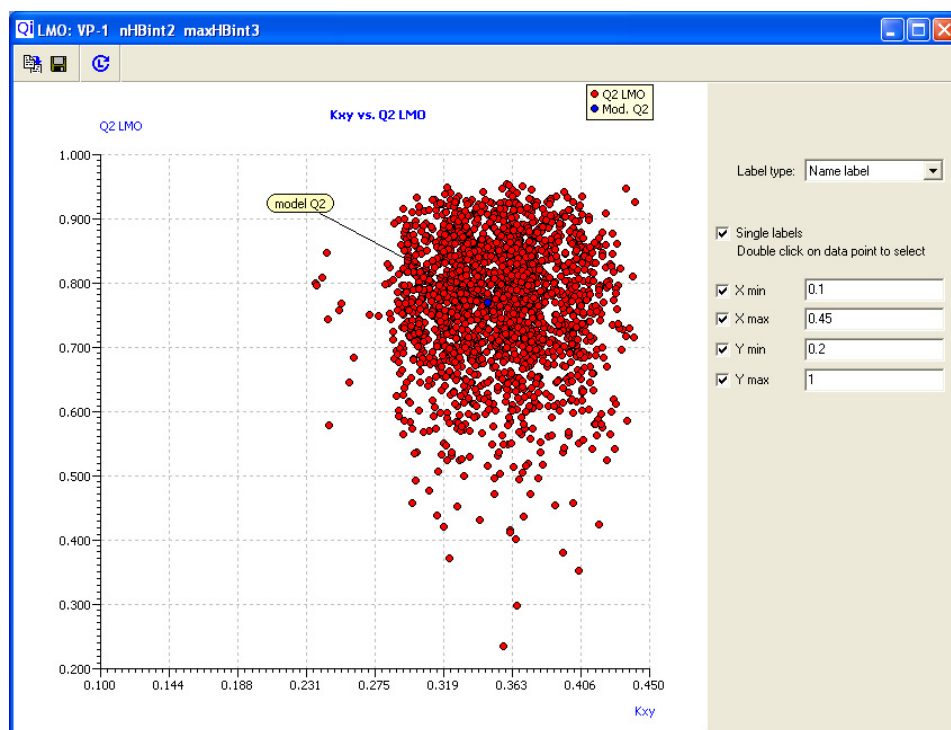
	Insubria graph: diagonal hat elements vs. predictions by model equation (Figure 28)
---	---



**Figure 28.** Example of Insubria graph (press  icon)

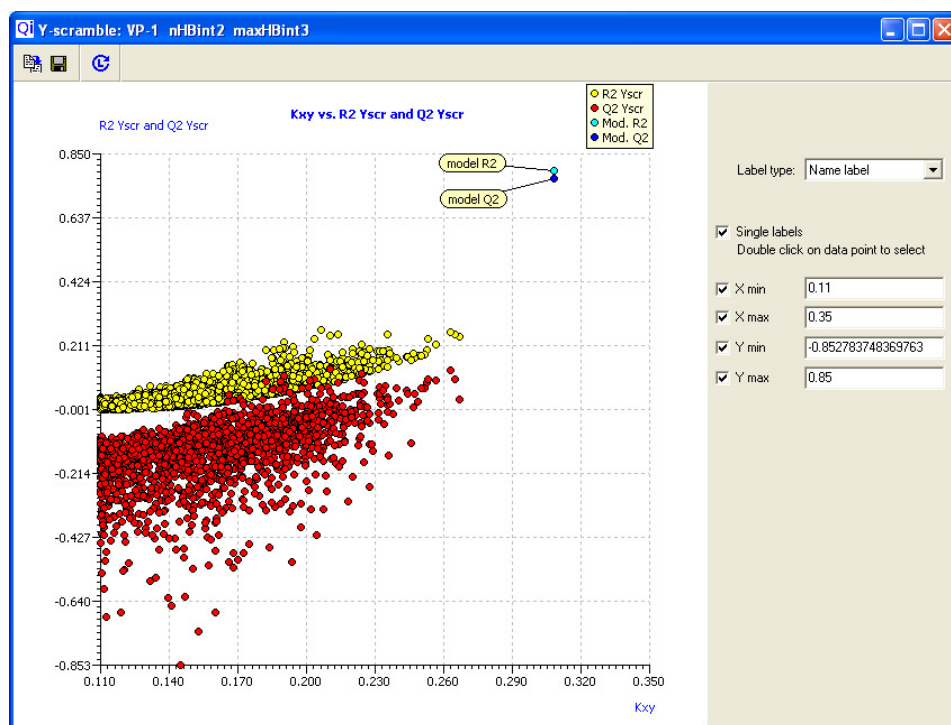
HAT diagonal values are reported on the abscissa axes while the predicted data are reported on the ordinate one. The vertical line corresponds to the HAT threshold value ( $h^*$ ) of the structural domain. Yellow and, blue data points (training and prediction set) are data predicted by model equation if the experimental endpoint is known, the data points labeled in red are the chemicals without experimental data (unknown, predicted by the model equation).

	Leave many out (LMO) plot (Kxy vs. $Q^2$ ) (Figure 29)
	Y-Scramble plot (Kxy vs. $R^2$ and $Q^2$ ) (Figure 30)
	Random descriptors plot (Kxy vs. $R^2$ and $Q^2$ )
	Random responses plot (Kxy vs. $R^2$ and $Q^2$ )





**Figure 29.** Scatter plot of LMO models compared to the QSAR model (press  icon)

On the abscissa the correlations among the block of the descriptors and the experimental data ( $K_{xy}$ ) are reported, while the QSAR model performance ( $Q^2$  as blue point, labeled as “model Q2”) and the LMO models performances ( $Q^2$  as red points) are reported on the ordinate axes. For robust and internally predictive models the performances of LMO models must be as similar as possible to the original model performances.






**Figure 30.** Scatter plot of Y-scrambled models compared to the QSAR model (press  icon)

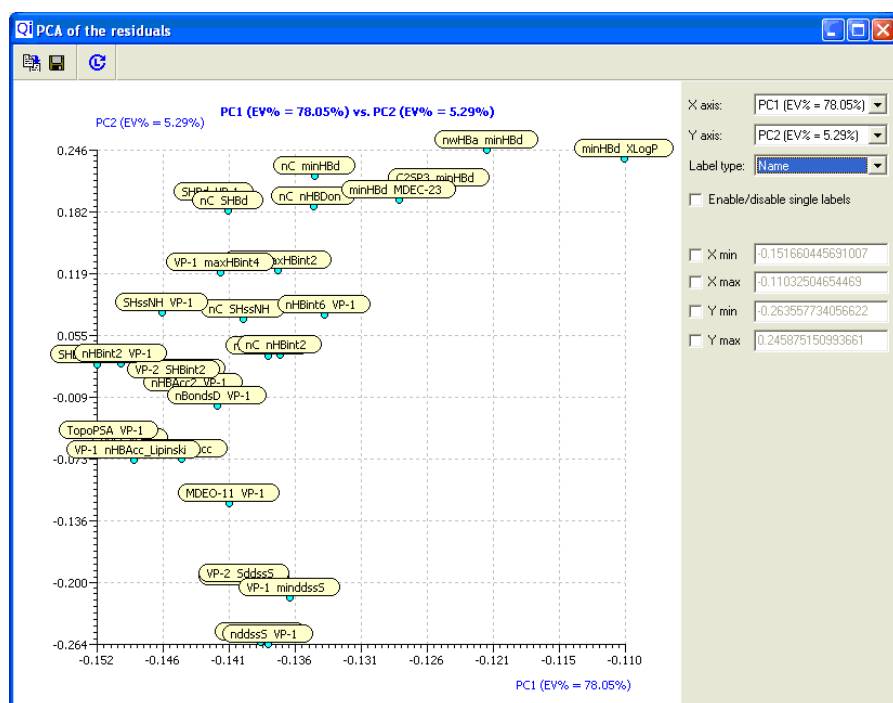
On the abscissa the correlations among the block of the descriptors and the experimental data are reported, while the original QSAR model performances ( $R^2$  and  $Q^2$  as cyan and blue points, labeled as “model R2” and “model Q2”) and the Y-scrambled models performances ( $R^2$  and  $Q^2$  as yellow and red points) are reported on the ordinate axes. In this case, the lost of the correlation between descriptors and response and the “disappearance” of the model must be evident both in the decrease of the Kxy values and  $R^2/Q^2$  in comparison to the original model.

The graphs obtained from the Random descriptors () and random responses () options are qualitatively similar to the one of the Y-scramble (in fact the concept behind is similar, i.e. to destroy the correlation among descriptors and responses) and can be evaluated in a similar manner.


## 8. Consensus modeling

In QSARINS it is possible to apply the consensus modeling (Gramatica et al., 2004), by selecting the most diverse models (among the best) in the GA population by PCA of the residuals (icon ) or of the modeling molecular descriptors of the selected models (icon )

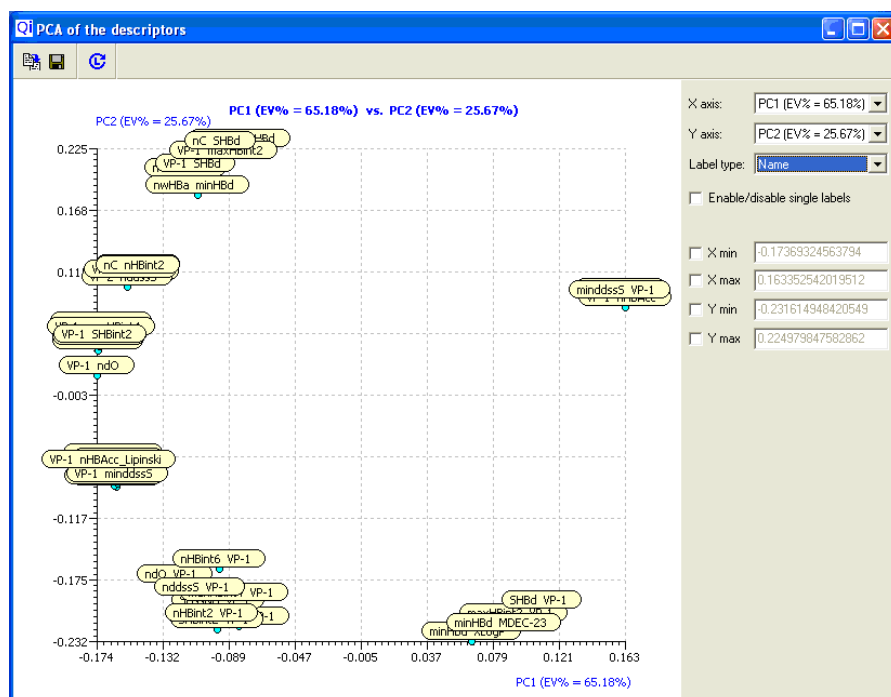
In the “View and select models” dialog (Section 5),  calculates the PCA based on residuals of the selected models, as in the following example (Figure 31):



**Figure 31.** PCA based on residuals of the selected models

 calculates the PCA based on the descriptors of the selected models. This is used to check whether some models are clustered, i.e. they are based on similar structural information, as in the following example. (Note: the models must have the same number of descriptors)







**Figure 32.** PCA based on the descriptors of the selected models

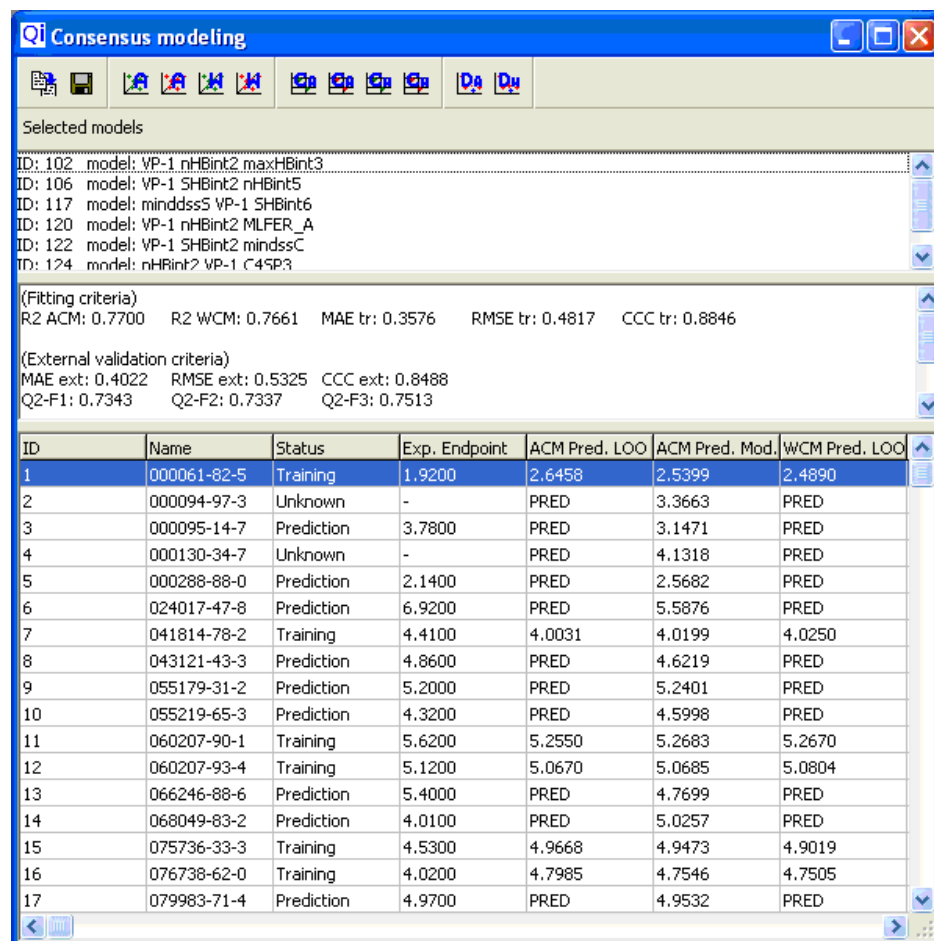
The most diverse models are those more distant in the above two plots (one considering the predicted responses, the other considering the structural diversity) and should be selected for the following consensus modeling.



When, using the PCA as a basis, a list of models has been selected, it is possible to calculate the average of their performances by means of a consensus modeling, clicking on the icon  or selecting the corresponding item “Consensus modeling” from the pop-up menu of the “View and selects models” (section 5) window.

When the user clicks the  icon, all the selected models (status “S”) will be used to calculate the consensus model. Consensus predictions are calculated both by a simple arithmetic average (ACM) and by a weighted average (WCM, Li *et al* 2008) of the individual predictions.

For the Consensus model, QSARINS provides the following information (Figure 33):



**Figure 33.** Consensus model dialog box

The first box in Figure 33 reports the list of selected models used for Consensus (Model ID and molecular descriptors)

The second box reports the statistical parameters for the Consensus, described below (the names in squared brackets are the ones reported by QSARINS).



**Fitting criteria:**  $R_{ACM}^2$  [R2 ACM],  $R_{WCM}^2$  [R2 WCM], MAE tr, RMSE tr, CCC tr.

**External validation criteria:** MAE ext, RMSE ext, CCC ext,  $Q_{F1}^2$  [Q2-F1],  $Q_{F2}^2$  [Q2-F2],  $Q_{F3}^2$  [Q2-F3].

The angle of regression calculated on external data from diagonal (scale shift), is used to evaluate the bias in the external predictions (Chirico and Gramatica, 2012).

The data matrix, for every chemical, lists the experimental data (Exp. Endpoint), the average consensus prediction ("ACM Pred.", both predicted by the model equation and by LOO), weighted consensus prediction ("WCM Pred." both predicted by the model equation and by LOO, calculated averaging the predictions of the molecules separately from every model), standard deviation of the individual fitting and LOO predictions ("Sd.Pred. fit." and "Sd.Pred. LOO"), minimum and maximum prediction values among the individual models ("Min Pred. fit", "Min Pred. LOO", "Max Pred. fit.", "Max Pred. LOO"), and average leverage value ("Av.HAT i/i").

The icons under the dialog title in Figure 33 have the following meaning.

The first icon from the left () copies the results into the clipboard while the second icon () saves the data in a text file.

The remaining icons (It is here recalled that a popup menu can be used for selecting options instead of the icons) can be used to plot various graphs on the consensus model performances, as exemplified in the table:



Scatter plot of experimental endpoint vs. ACM predicted by model equation



Scatter plot of experimental endpoint vs. ACM predicted by LOO



Scatter plot of experimental endpoint vs. WCM predicted by model equation



Scatter plot of experimental endpoint vs. WCM predicted by LOO



Williams plot of average hat diagonal elements vs. ACM. predicted by model equation residuals



Williams plot of average hat diagonal elements vs. ACM. predicted by LOO residuals



Williams plot of average hat diagonal elements vs. WCM. predicted by model equation residuals



Williams plot of average hat diagonal elements vs. WCM. predicted by LOO




Insubria graph of average hat diagonal elements vs. ACM predicted by model equation




Insubria graph of average hat diagonal elements vs. WCM predicted by model equation

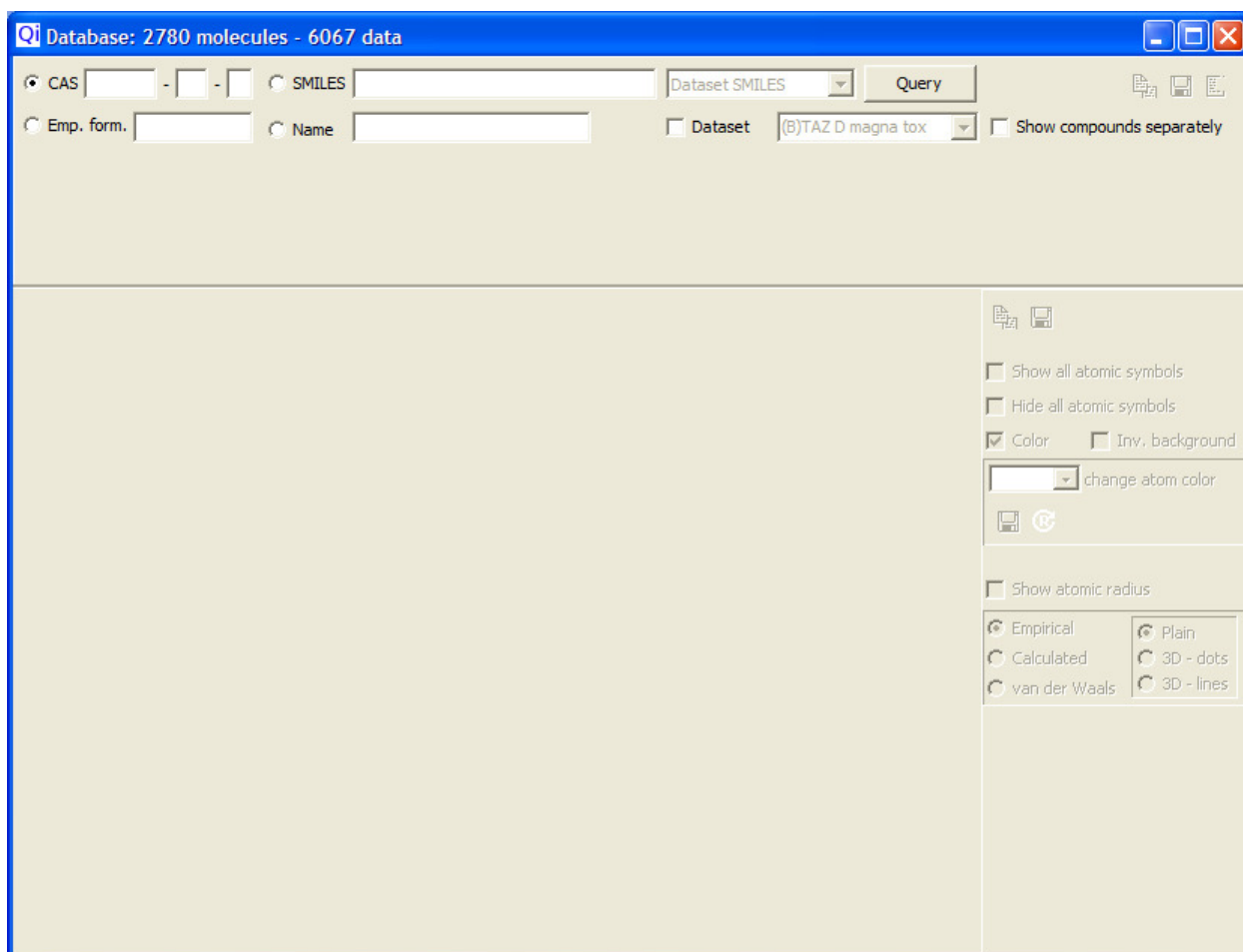
## 9. QSARINS-Chem module

A section of the QSARINS main interface is devoted to a module called QSARINS-Chem

() that allows the management of a database of chemicals, the application of already developed and stored models and the calculation of ranking for the data of interest in the imported dataset. The database management is already provided with several datasets of chemical structures (Hyperchem, .hin, and MDL MOL format files) with some end-points of environmental interest (collected and modeled in the Insubria group in 18 years). The models section includes more than 20 new QSAR/QSPR models, based on descriptors calculated with the on-line open-source PaDEL descriptor version 2.18, ready to be applied to new chemicals. These models are accompanied by their QMRF (QSAR Model Reporting Format) and .sdf (with modeling descriptors). User defined datasets and models can be also easily added.

### 9.1 Database

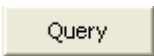
Pressing the  icon (or, alternatively, selecting Tools->Query database menu item) in the main screen, the following dialog will appear, where it is possible to query the compounds included in the database:

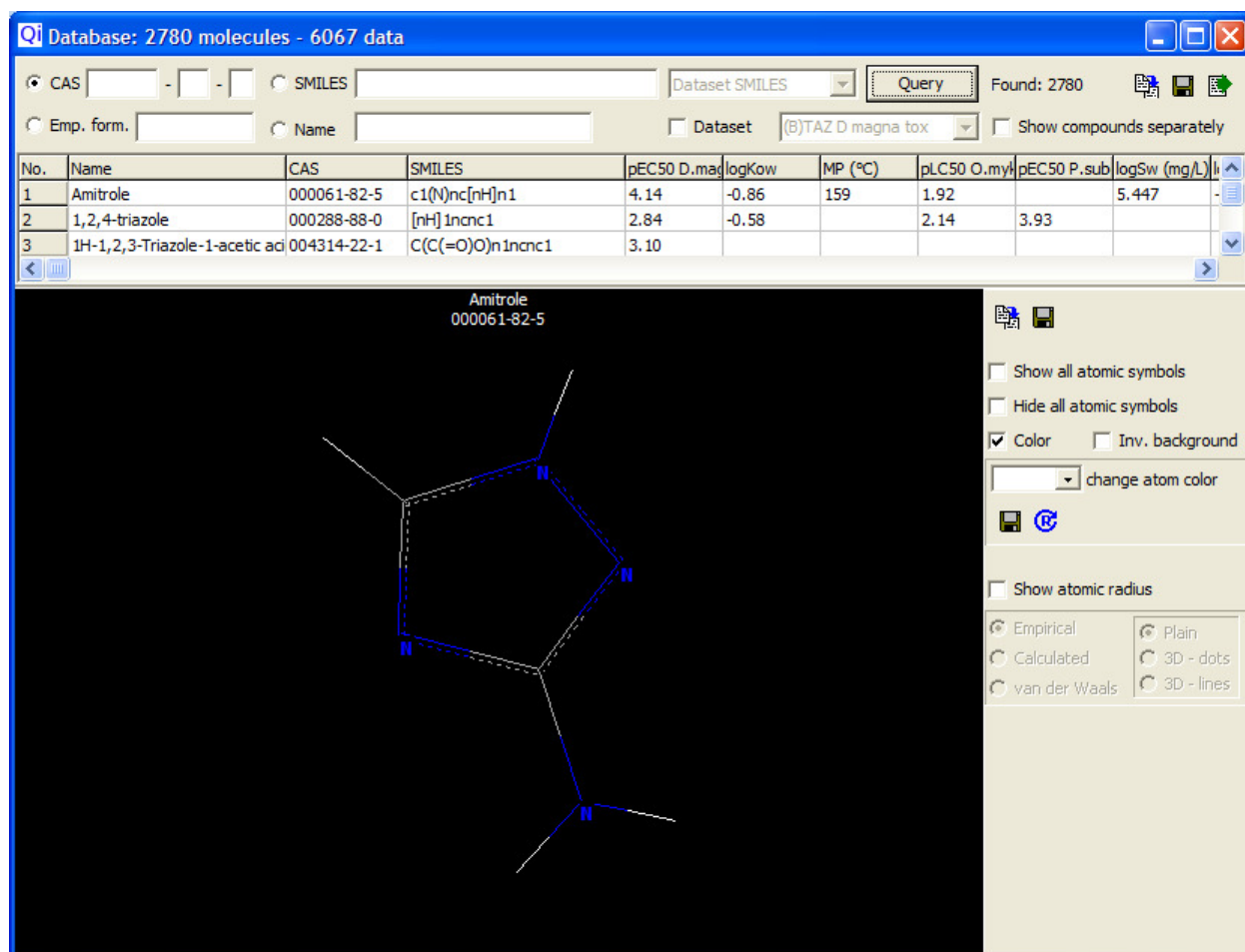


**Figure 34.** Database dialog at its first appearance

The window title reports how many compounds are available in the database, as well as the number of experimental data here collected (note: if one compound is present in different datasets, each dataset possesses its own version of the same compound).

#### 9.1.1 Querying the database

Pressing the  button leaving all the other fields blank, as shown in Figure 34, QSARINS will display all the available molecules (the total number of them is reported at the right of the Query button), enabling also all pertinent options, as shown in the following figure:



**Figure 35.** Database dialog when chemicals are queried

Every row in the grid is a chemical extracted from the database. The first column (No.) is an arbitrarily assigned increasing number, the second (Name) is the name of the compound, the third is the CAS (when available), the fourth is the SMILES (generated in batch using Open Babel 2.3.2, starting from the AM1 minimized HyperChem structure), the fifth and the following are the experimental responses of the corresponding dataset, when available.

As it can be noted in the first row in Figure 35, for each molecule there may be more than one experimental response. In this case, before visualization, the molecule is automatically detected and extracted from the database using its CAS number, and the name and SMILES shown are the ones of the first dataset where the compound is found (in this case pEC50 *D.magna*). This behavior allows to condensate the output to the smallest dimension; in case the user likes to see all the compounds in separate lines with their own “single” experimental responses, which could be useful, for example, to

verify a new user-defined dataset (see section 9.1.3 “Writing a user-defined dataset”), the option ☒ Show compounds separately must be checked.

The whole database can be explored in different ways: by CAS , SMILES, empirical formula or by its name.

To query the compounds by CAS, the corresponding radio button must be pressed ( ☒ CAS  -  -  ). The CAS number must be entered in the corresponding edit boxes. If the

CAS is precisely entered in its form 6+2+1, the corresponding compound will be extracted from the database. If the CAS is partially entered, all the compounds that share the entered part will be extracted:



for example ☒ CAS  -  -  will extract all the molecules that share 000050-00 whatever the


last box value is, ☒ CAS  -  -  will extract all molecules that share 000050 with the last number being 0, whatever the middle value is. If a number is partially entered, e.g. 50 in the first edit box, all molecules having the sequence 50 (five-zero) in the first part of the CAS will be extracted (e.g. 000050-29-3, 025057-89-0 and so on). It is important to note that some compounds included in the database do not have any CAS number; to overcome this potential problem, for these chemicals we set an arbitrary and fictitious CAS number, composed of the correct form 6+2+1. The first six numbers are written using the notation “XXXUNK”; the next two numbers are related to an arbitrary ID that identify any single dataset included in the database (for example, the dataset for BCF-Lu is identified with ID 01, dataset for Pimephales toxicity with ID 02, dataset for Koc with ID 03 and so on until the last dataset, with ID 48, in no particular order). The last number denotes the progressive number of chemicals without CAS in that particular dataset. Summing up, for example, if the Koc dataset (ID 03) has three different chemicals without CAS number, their assigned fictive CAS in QSARINS-Chem will be: XXXUNK-03-1, XXXUNK-03-2 and XXXUNK-03-3.

To query the compounds by SMILES, the ☒ SMILES  radio button must be selected:

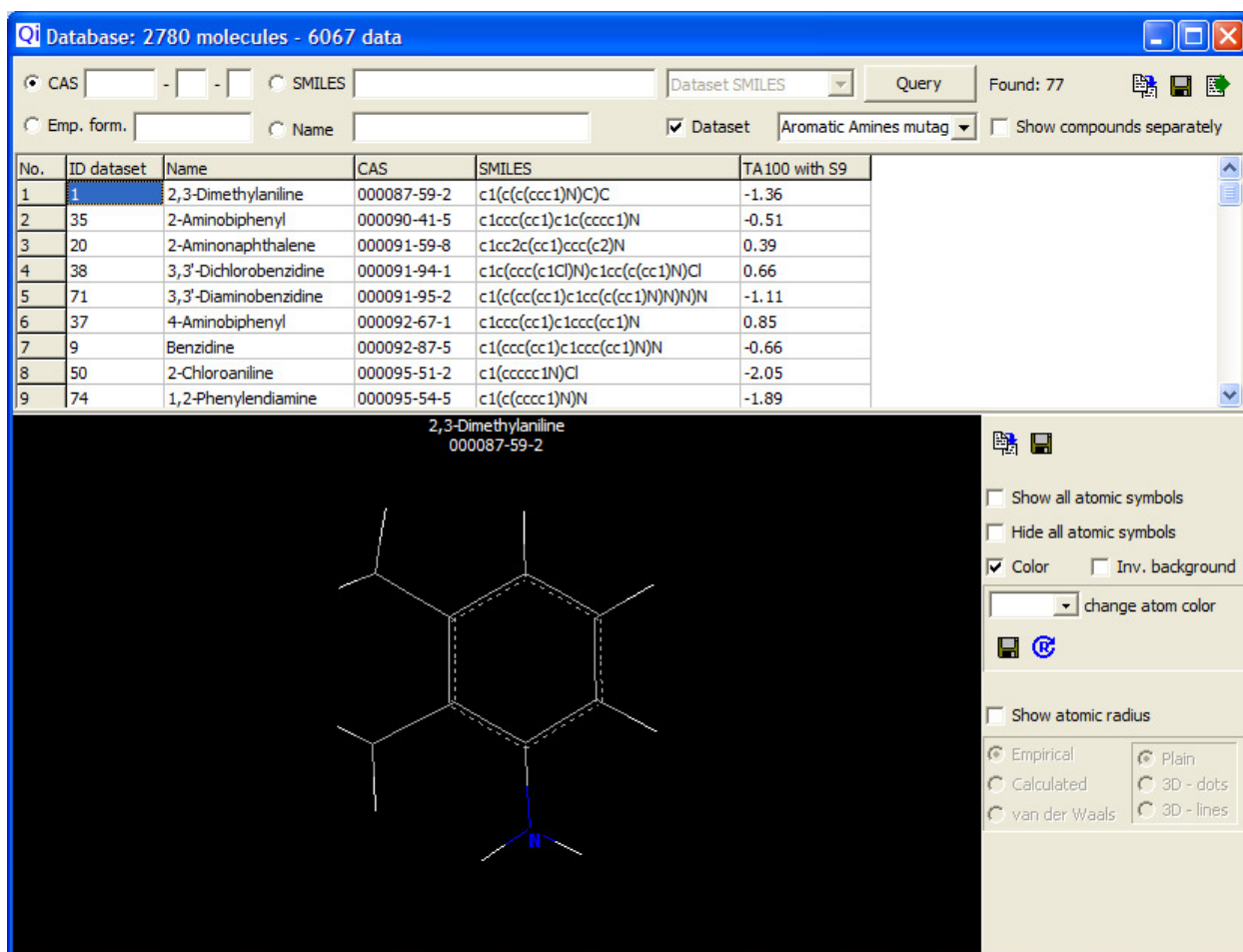
as a consequence the following  list box, which allows the user to select different input SMILES types (Pubchem – Epi suite or Dataset SMILES, namely those calculated with Open Babel 2.3.2) will be activated. If the “Dataset SMILES” option is selected, all the SMILES in the dataset that matches exactly the one entered will be displayed (this option is the fastest because it does not need to

parse the SMILES and the molecular structure files). The second option “Pubchem - EPI suite” parses the SMILES generated by the corresponding web sites and tools. In this case QSARINS examines the files and, following the canonical SMILES rules, determines the topology of the molecule: all compounds sharing the same topology are extracted from the database and visualized.




The remaining ways to query the database are by the empirical formula (  Emp. form.  ) and by name (  Name  ). Concerning the latest option, it suffices to enter even a partial name: all compounds sharing the same part of the name will be extracted.




Instead of using and querying all available datasets, it is possible to query a single dataset, using the  Dataset option. Once selected, the corresponding list box at its right will be activated, allowing the user to select one of the stored datasets. Once done, the output will be similar to the one in Figure 35, except that a new column containing the compound ID given in the original dataset (ID dataset column) will be shown, as in Figure 36.

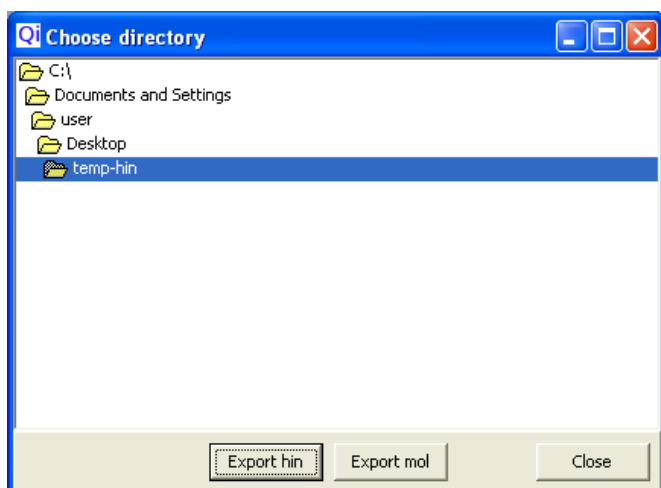




**Figure 36.** Database dialog using the Dataset option

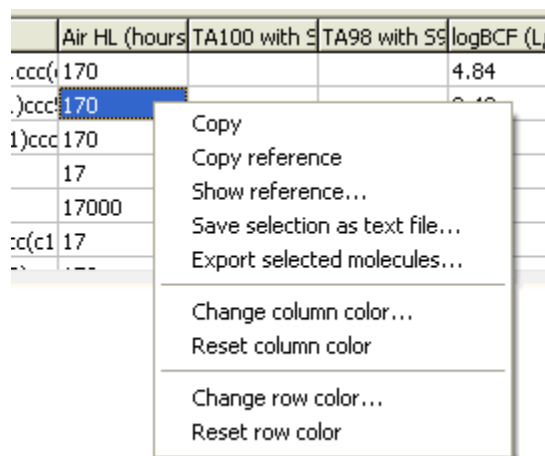
After querying the dataset, if compounds are found, some supporting tools will be activated (  ).

The option  copies into the clipboard all the compounds information as showed in the data grid (additionally, for every response is reported the bibliographical reference) while  does the same but saving everything into a text file. The  option exports the selected compounds files structures (.hin or MDL MOL) in a user selected folder, as showed here:



If more than one response is found for a compound, the structure file corresponding to the first dataset is saved (for example, for the first compound found as showed in Figure 35, only the structure corresponding to pEC50 *D.magna* is saved). A note of warning: in QSARINS-Chem all the original structures were drawn, minimized and computed as Hyperchem .hin files, and all the MDL MOL structures have been recalculated automatically by a conversion software (Open Babel 2.3.2). In case of doubts is better to refer first to the Hyperchem files.

So far we have dealt with general tools supporting the queried molecules, but it is also possible to get more focused data and information using the popup menu:

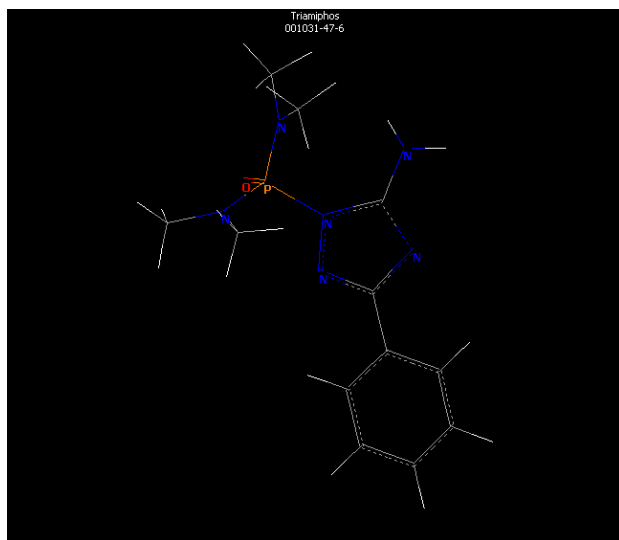


“Copy” will copy only the selected data (one or more grid cells) into the clipboard, adding the headers when needed (e.g. the response name with the measure unit). If a cell containing a response is selected, “Copy reference” is activated and this option will copy only the bibliographical reference into the

clipboard (if more than one cell/data are selected, only the first on the left will be considered). The same can be done with “Show reference...”, except that in this case the reference will be shown in a dialog box. “Save selection as text file...” acts as the “Copy” option except that the output will be directed to a text file. The last option, “Export selected molecules” will save the structure files (.hin or MDL MOL) as already shown. The only difference consists in having the possibility to save exactly the wanted compound(s) (not just the first one found, as shown so far) in a row containing more than one response. To do that, select the cell(s) corresponding to the response(s) of interest and then “Export selected molecules” option. “Change column color...” and “Change row color...” allow the user to choose a personalized color for the wanted column/row in order to emphasize the data of interest. “Reset column color” and “Reset row color” reset the selected column/row to the default grid color.

### 9.1.2 Visualize chemical structures

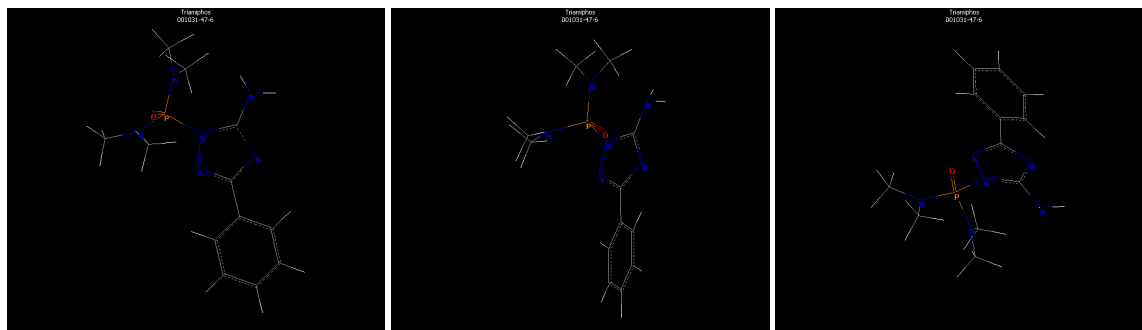
In QSARINS it is also possible to visualize the 3D structure of the chemicals as in the following example:



**Figure 37.** Example of a 3D view of a chemical

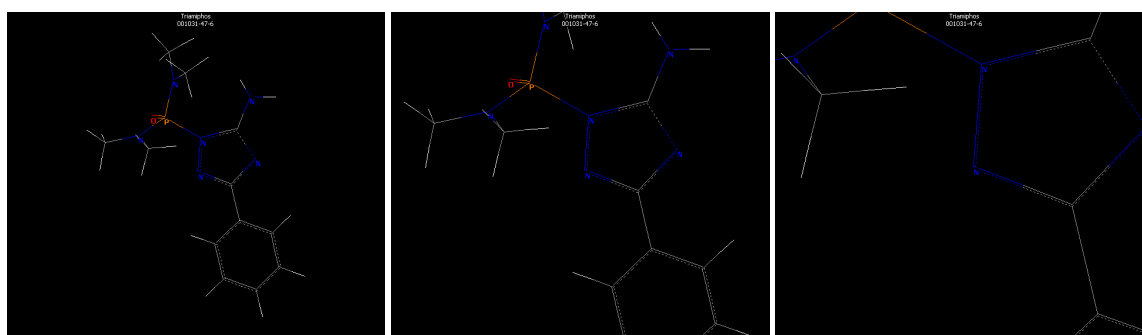
At the top of the image is reported the chemical name and below the CAS registry number.

Pressing the left key and moving the mouse, the 3D structure can be rotated:



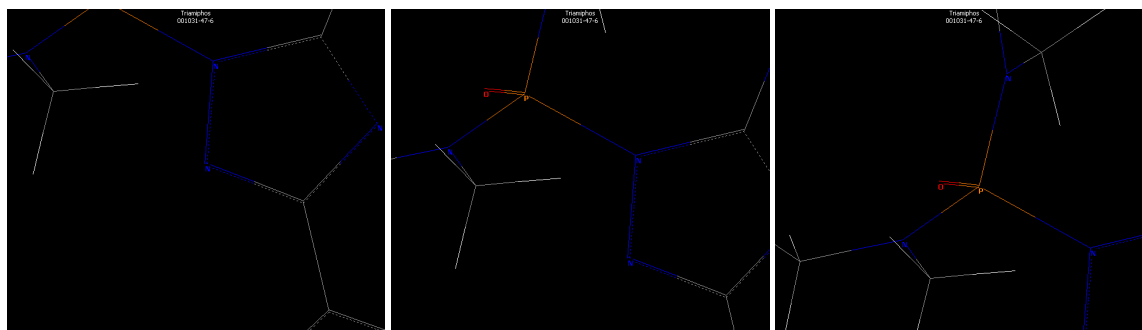
**Figure 38.** Rotating the 3D view of a chemical

Pressing the left mouse key holding the Ctrl key down on the keyboard it can be scaled:





**Figure 39.** Zooming the 3D view of a chemical

Pressing the left mouse key while keeping the shift key down the structure can be translated:



**Figure 40.** Translating the 3D view of a chemical



The compound image can be copied in the clipboard pressing the  icon near the upper right corner or pressing the right mouse button and selecting “Copy” from the popup menu, while the image can be saved (as .jpg or .bmp) pressing the  icon or selecting the “Save image...” from the popup menu.

The compound image can be tuned in different ways: “Show all atomic symbols” displays all the atom symbols (including C and H, which display is disabled by default), while “Hide all atomic symbols” does the opposite. The “Color” option is selected by default, displaying the molecules in a black background (as shown in Figures 37-40). When “Color” is unselected, all the atoms and bonds are displayed in black while the background becomes white (this option should help for black and white documents). The “Inv. background” button swaps the background and structure colors.

The element colors follow the CPK rule, but with slight differences in color saturations (e.g. Cl and F) that are here introduced for the user convenience. The colors can also be changed (except for C and H) at will using the “change atom color” list box. If this option is selected the system color dialog chooser will appear, allowing the change in the easiest way. If needed the new color setup can be saved pressing the

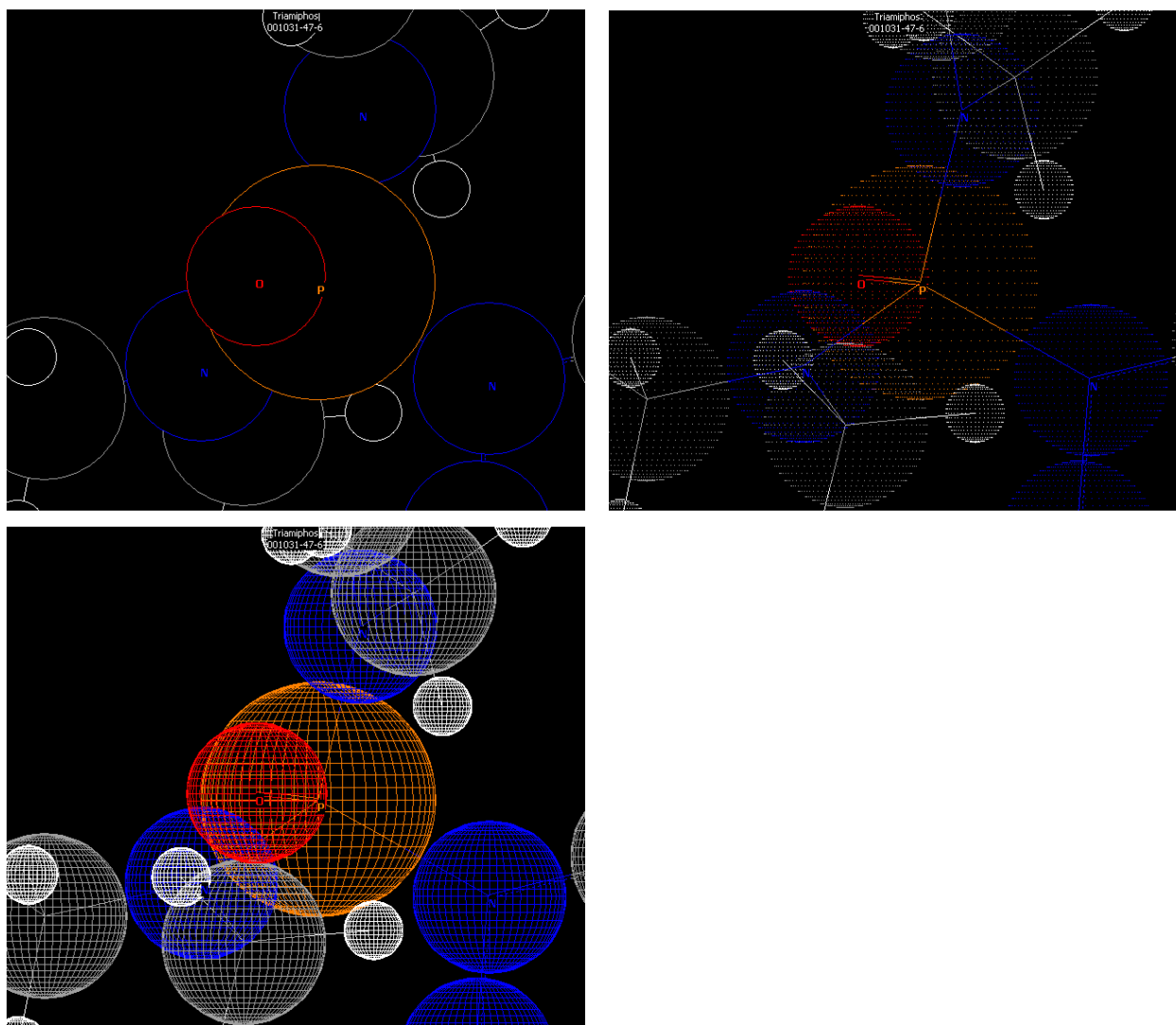


icon below the list box. At the following run of QSARINS this will become the used color setup.

Pressing the  icon, the atom color map will be reset as default: after that, pressing the  icon, the default map will be saved.

The “Show atomic radius” option allows displaying the radius of the atoms as: Empirical, Calculated and van der Waals.

The atom radius can be displayed in three ways: plain, 3D – dots and 3D – lines, as in the following figure:



**Figure 41.** Display of the atomic radius in the 3D view of a chemical. Upper left: plain. Upper right: 3D-Dots. Lower: 3D-Lines.

The plain option is used when a quick display is needed, while the others are more demanding in terms of calculations.

### 9.1.3 Prepare a user-defined dataset

It is possible to add a user defined dataset to QSARINS, following few simple rules. To make easier the process learning, it is here explained how the “BFR Henry constant” database has been built (this dataset has been chosen because of its small size).

The first step is to create a spread-sheet (using Open office, for example), as done for the “BFR Henry constant” dataset in the following example:

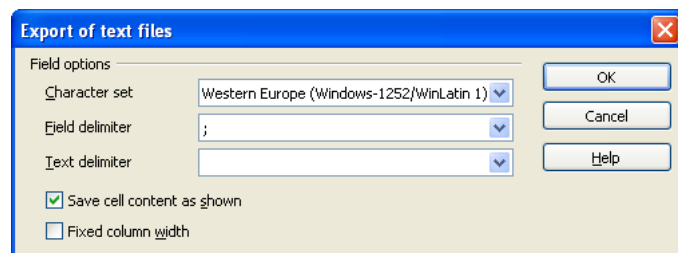
	A	B	C	D	E	F
1	ID	Name	CAS	Smiles	logH (Pa m3/mol, 25 °C)	QSAR & Comb.Sci. 28, 2009, pp 790-796
2	28	2,4,4'-triBDE	041318-75-6	<chem>c1(cc(ccc1Occc(cc1)Br)Br)Br</chem>	0.684	
3	47	2,2',4,4'-tetraBDE	005436-43-1	<chem>c1(cc(ccc1Occc(cc1)Br)Br)Br</chem>	-0.071	
4	99	2,2',4,4',5-pentaBDE	060348-60-9	<chem>c1(cc(c(cc1Occc(cc1)Br)Br)Br)Br</chem>	-0.222	
5	100	2,2',4,4',6-pentaBDE	189084-64-8	<chem>c1(cc(cc(c1Occc(cc1)Br)Br)Br)Br</chem>	-0.62	
6	153	2,2',4,4',5,5'-hexaBDE	068631-49-2	<chem>c1(cc(c(cc1Occc(cc1)Br)Br)Br)Br</chem>	-0.585	
7	154	2,2',4,4',5,6'-hexaBDE	207122-15-4	<chem>c1(cc(c(cc1Occc(cc1)Br)Br)Br)Br</chem>	-1.097	
8	209	2,2',3,3',4,4',5,5',6,6'-decaBDE	001163-19-5	<chem>c1(c(c(c(c(c1Occc(cc1)Br)Br)Br)Br)Br)Br</chem>	-1.398	

**Figure 42.** Example of a spreadsheet for writing a user-defined dataset

The first column (A) contains the arbitrary ID of the compound used in the original modeling process of the dataset. The second column (B) contains the name of the compound, the third (C) the CAS number (if not available, the user could add an arbitrary CAS following the 6+2+1 rule; see above in the paragraph 9.1.1), the fourth (D) the SMILES. The fifth column (E) reports the experimental value of the response: it is also important to write carefully in the first row the name of the response (in this case also with the measure unit, “logH (Pa m3/mol, 25 °C)”), because it will appear in QSARINS-Chem in the main view of the database. The first row of the last column (F) must contain the reference of bibliography (if there is none, just write “No reference” or let blank).

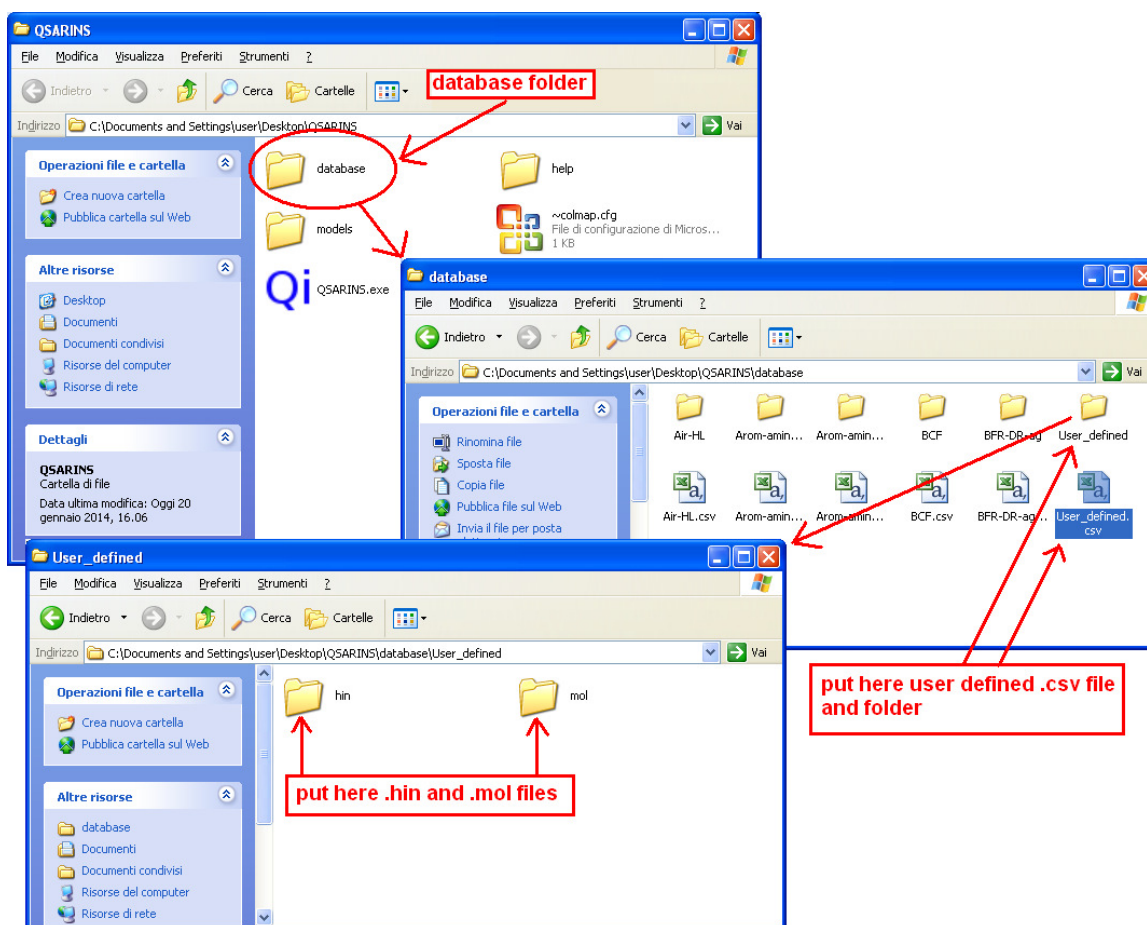
Once set up the first row of the spread-sheet, the subsequent rows must contain the corresponding data and information. Since only the first row can contain bibliographical data, all subsequent row cells corresponding to Column F must be blank (if filled with data they will be anyway ignored by QSARINS).

Once finished filling, the spreadsheet must be saved in the “database” folder of QSARINS (Figure 43) as a .csv text file, with a semicolon (;) as delimiter and no text delimiter, as in the following example (in this case LibreOffice Calc, other software may have a different dialog layout):



The file name of the prepared file should be representative of the user dataset, since it will appear as an option in the dataset interface of QSARINS-Chem (as for example Figure 36, “Arom Amines mutag”): for example “BFR Henry constant.csv” (is better to avoid the use of dots in dataset names).

Now the user must insert the representative structure files of its own dataset and the corresponding folder must be created within the “database” folder, as in the figure :



**Figure 43.** Dataset files and folders location

The name of the new folder **must** match the name of the prepared .csv file without the extension, i.e. “BFR Henry constant”. Within this folder, two new subfolders named “mol” and “hin” must be created. The corresponding structure files (MDL MOL and HyperChem files) must be put in these folders.

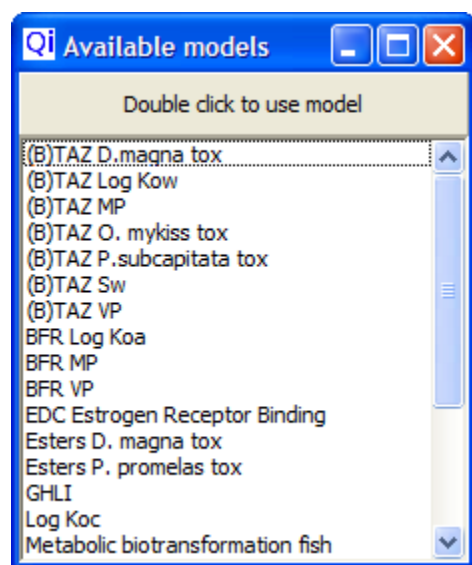


Concerning the structure file, the ones used by QSARINS for querying and displaying, are the Hyperchem (.hin) while the MDL MOL (.mol) are optionally used only for exporting purpose.

The names of the .hin and .mol data are fundamental and **must** match the CAS reported in the supporting files because this is the way QSARINS can locate them. For example the CAS 041318-75-6 in the example (Figure 42) must have a file named 041318-75-6.hin in the hin folder, and a file named 041318-75-6.mol in the mol folder.

### 9.2 Apply developed model

All the stored models (see Section 7, “Analysis of single models”, for further information) can be used for the prediction of new chemicals, pressing the **models** icon of the main screen, or selecting “Tools->Apply developed models” menu item (Figure 1). The following dialog, where all the available and stored models are listed, will appear (Figure 44):



**Figure 44.** List of available models


Double-clicking the preferred model in the list, another dialog appears, which allows setting the input of new chemicals for their predictions (Figure 45). A file containing the list of chemicals with relative values of descriptors and the response must be prepared beforehand, then it must be copied and pasted in the


screen sheet (if experimental data are not available, insert -999). This screen sheet could be also manually filled. Concerning the data format, while editing, it is obligatory to use the dot (.) as the decimal separator.


Firstly, the user must set the number of “Rows” in order to match the number of chemicals to be tested (if the number of chemicals to be predicted is 100, the user must set the number of Rows to 100).


**Figure 45.** Dialog for adding external chemicals to existing model

The first group of icons in Figure 45 (     ) are the ordinary operations of:


 : copy selected data to the clipboard

 : paste selected data from the clipboard

: cut selected data (and paste to the clipboard)

: delete selected data



: undo last operation

by pressing the  icon, the previous functions will alternatively enabled or disabled. When disabled, the user is allowed to manually edit the grid fields, instead of just paste them from the clipboard.


On the user preference a popup menu with all the listed options can be used.

The “Rows” options must match the number of chemicals to be tested, as mentioned before.

Pressing the “Accept” button, the predictions will be explored using the same dialog as explained in full details in the “Analysis of single models” (section 7)

In QSARINS there is the possibility to redevelop, check, validate and apply to any chemicals single MLR models, previously developed also by other tools. In QSARINS-Chem more than 20 QSAR/QSPR models, developed by the Insubria group with the freely available PaDEL-Descriptor (open source software for the calculation of molecular descriptors and fingerprints) are now implemented and available, with their QMRF and .sdf (also with modeling descriptors) files, exportable by pressing  and  icons that will be appear in the Single Model dialog box (Figure 22).

#### 9.2.1 Apply developed model: the PBT Index example

The PBT Index, included in the stored models available in QSARINS (Figure 44), is also achievable from the main software window (Figure 1) , where the  button (or Tools->Apply PBT index model) applies the model, developed and proposed in international journal by the Insubria research group (Ester Papa and Paola Gramatica, QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure, *Green Chemistry*, 2010, 12, 836-843; DOI: 10.1039/b923843c, Hot Article).

The original and published model for the prediction of the cumulative Persistence-Bioaccumulation-Toxic (PBT) behaviour of chemicals, based on DRAGON descriptors (5.5 version, 2007), is the following:

$$\text{PBT Index} = -1.44 (\pm 0.10) + 0.65 (\pm 0.03) \text{ nX} + 0.22 (\pm 0.01) \text{ nBM} - 0.39 (\pm 0.06) \text{ nHDon} - 0.07 (\pm 0.03) \text{ MAXDP}$$

$$n = 180, R^2 = 88.40\%; Q^2_{\text{LOO}} = 87.72\%; Q^2_{\text{LMO}} = 87.73\%; R^2_{\text{YS}} = 0.02; \text{RMSE} = 0.52$$

The molecular descriptors were selected by Genetic Algorithm and their ability in predicting chemicals not participating to model development (training: 54 chemicals) was preliminarily verified on external chemicals (126 chemicals): ( $Q^2_{\text{ext (F1)}} = 80.72\%$ ;  $R^2_{\text{ext}} = 89.27\%$ ;  $\text{RMSE}_p = 0.72$ )

The model implemented in QSARINS software is the following, based on a freely available software for descriptors calculation PaDEL-Descriptor (Yap, 2011):

$$\text{PBT Index} = -1.46 (\pm 0.19) + 0.64 (\pm 0.05) \text{ nX} + 0.22 (\pm 0.02) \text{ nBondsM} - 0.39 (\pm 0.13) \text{ nHBDon\_Lipinski} - 0.06 (\pm 0.06) \text{ MAXDP2}$$

$$n = 180, R^2 = 88.89\%; Q^2_{\text{LOO}} = 88.25\%; Q^2_{\text{LMO}} = 88.28\%; R^2_{\text{YS}} = 0.02; \text{RMSE}_{\text{TR}} = 0.51; Q^2_{\text{ext (F1)}}^* = 89\%; \text{RMSE}_p^* = 0.49.$$

\*: split model, with 92 compounds in the Training set and 88 compounds in the Prediction set.

Below, we report an example of how to apply the PBT Index model to a set of external compounds with or without experimental data.

1) Calculate the modeling descriptors of the PBT Index equation (nX, nBondsM, nHBDon\_Lipinski, MAXDP2) using the open source **PaDEL-Descriptor software v.2.16** or later versions.

(<http://padel.nus.edu.sg/software/padeldescriptor/index.html>);

2) Choose and enter in PBT Index model of QSARINS and set the number of rows according to the number of your set of chemicals (see Figure 45);

3) Paste the names (or CAS numbers), calculated descriptors and responses (if available) (see Figure 45);

4) Apply the PBT Index model simply pushing the button “Accept” (see Figure 45);

5) Equation and statistical parameters of the model (including all the external validation criteria), predicted data and hat values of the Insubria dataset (180 compounds in this case) and your dataset will be shown (ID > 180, in this case);

6) The user can now visualize the following graphs: Scatter plot, Williams plot and Residual plot for the compounds with experimental data; Insubria graph for compounds with and without experimental data (“unknown”). For the explanation of the graphs, see the Section 7” Analysis of single models”;

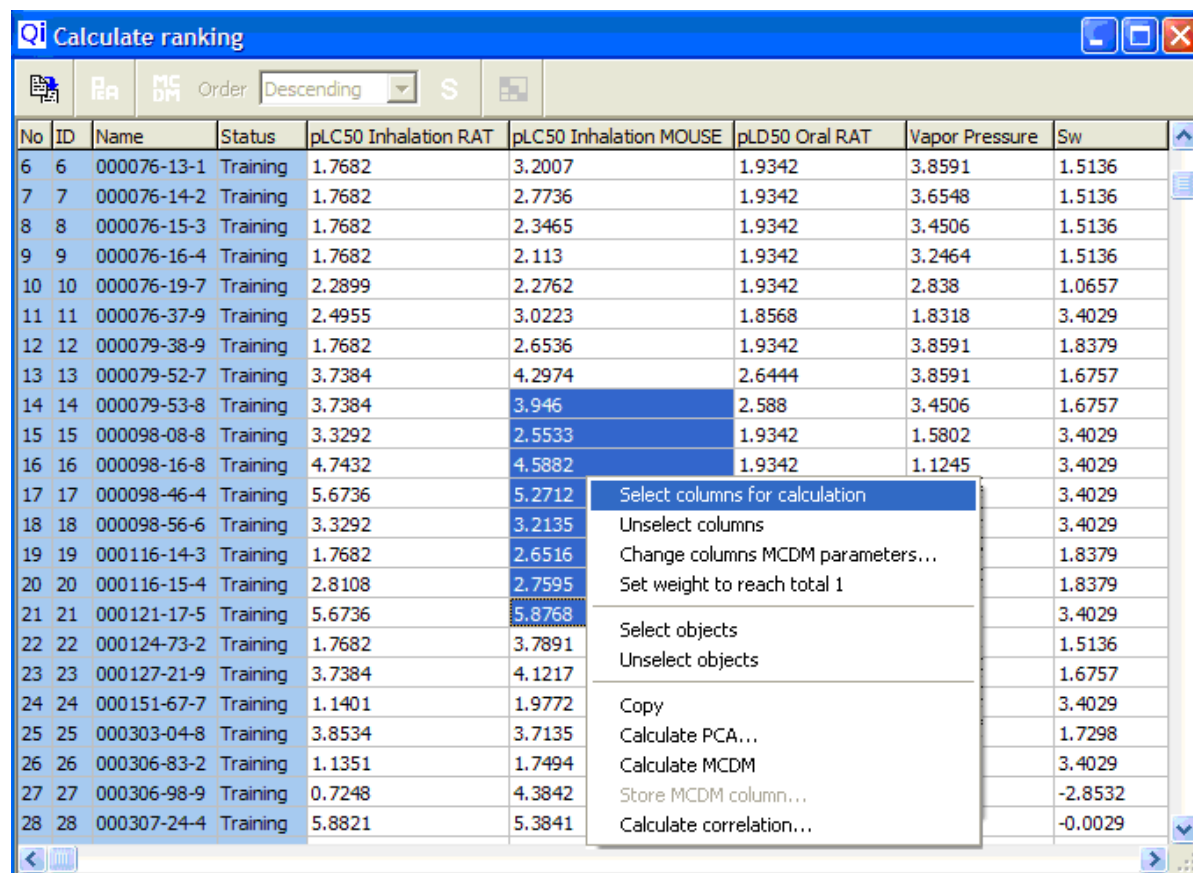
7) The user can save and/or copy the results.

Any other proprietary single model, can be used by QSARINS if it is saved in the appropriate models folder (see section 7, “Analysis of single models” to see how to generate single models and save them in the models folder) making it available in the box called by the **models**, or by the “Tools->Apply developed models.”, from the QSARINS main window, for further application.

In this version of QSARINS more than 20 QSAR/QSPR models, developed by the Insubria group with the freely available PaDEL-Descriptor software for the calculation of molecular descriptors and fingerprints, are now implemented.




### 9.3. Ranking

Once imported a dataset, pressing the **ranking** icon (or, alternatively, selecting Tools->Calculate ranking menu item) in the main screen, the following dialog will appear, where it is possible to calculate the ranking of the user-selected columns of the input dataset by Principal Component Analysis (PCA) and/or Multi-Criteria Decision Making (MCDM):



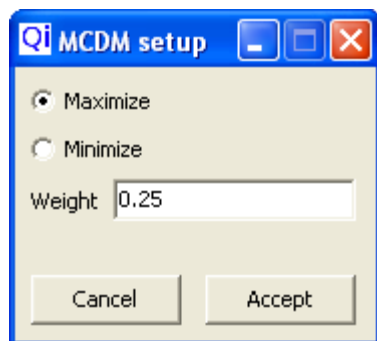
**Figure 46.** Dialog for calculating the ranking of the molecules

In order to calculate the MCDM and the PCA on the imported data, it is necessary to select the column of interest by the popup menu option "Select column/s for calculation" ("Unselect columns does the opposite). All selected columns will be coloured in yellow.

Once this is done, the following icons will be activated for the selected data/columns:  for the Principal Component Analysis,  for the Multi-Criteria Decision Making (MCDM) and  for the correlation matrix. The corresponding popup menu items are also activated (Calculate PCA, Calculate MCDM and Calculate correlation). There is also the possibility to exclude unwanted rows from computations (that will be marked in gray), because as default they are all selected. To select or unselect rows for calculations the options “Select Objects” and “Unselect Objects” of the popup menu must be used (Figure 46).




Using the selected columns and rows of data, the Principal Component Analysis calculates the loadings and the scores, as already explained in the “Data setup” section. The correlation matrix is calculated over the selected columns and rows and visualized as explained in the “View Data” section.

The purpose of the MCDM is to sort and rank the studied objects (in our case, chemicals) according to a score that is calculated, by means of a method called function of dominance, using the weight and the optimality of the user-selected data column (in our case descriptors or responses). The meaning of the column weight can be roughly summarized as “how much it influences the calculation of the score”. The sum of the weights of the selected columns must be 1 and, in fact, by default the value assigned to every column is  $1 / n$ . of columns. Anyway, the values of the weights can be changed by the user selecting the option “Change columns MCDM parameters...” of the popup menu and using the dialog that follows, as reported here:



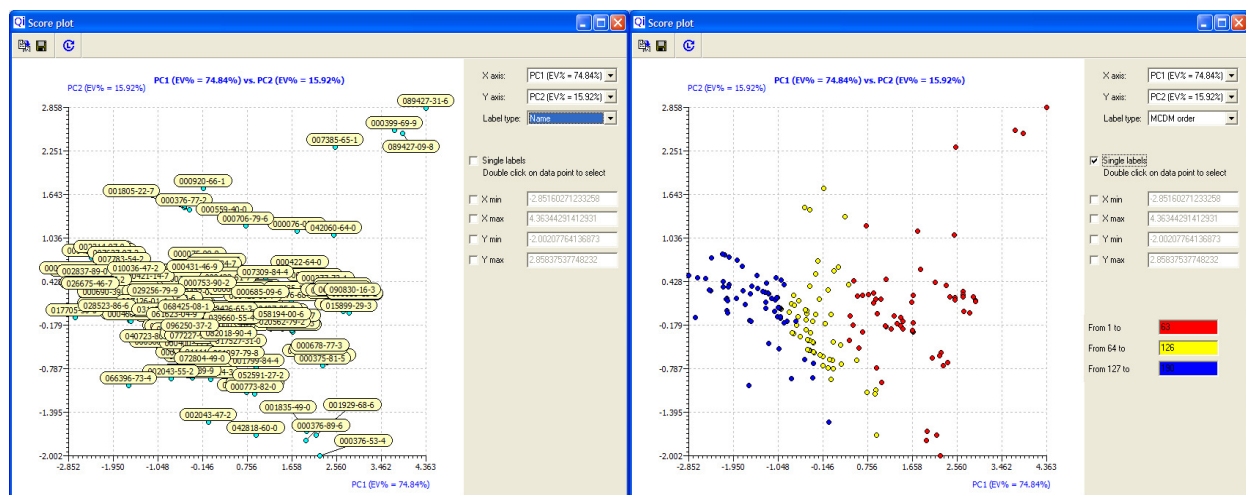
The “Maximize” and “Minimize” options are related to the MCDM optimality. This options serves “to tell” to the algorithm, in calculating the scores, whether the values (selected for the ranking) of the molecules of a column are those near the minimum or the maximum value. For example, if the selected column is toxicity, the user could like that the best values are those near the minimum.

Concerning changes in the user-defined weights, QSARINS controls their coherence asking new values until the total it is not negative and smaller than zero. Since at the end it could be difficult to reach a total of 1 (usually because of decimal numbers approximations) the option “Set weight to reach total 1” has been introduced. To clarify its use here it follows a simplified example: consider having three columns weighted 0.33. The total weight would be 0.99, but the user needs to reach exactly 1. Instead of manually weighting one of the column 0.34 (this is a trivial example, but in practice it can become tedious when there are lots of columns to be controlled and summed up) the user can just click the aforementioned option on a column of choice: its value will then automatically be trimmed to reach a total of 1.

Once the desired columns are selected and the weights/optimalities are assigned according to the user needs, pressing the  button (or selecting the corresponding “Calculate MCDM” option from the popup menu) the ranking scores will be calculated. After that, the rightmost column of the data grid called “MCDM score” will be filled with the scores. The rows excluded from the MCDM calculations obtain a fixed value of 0. Selecting the “Ascending” or “Descending” options from the “Order” drop list at the right of the  button the objects scores will be sorted accordingly. Once the scores are calculated, the  icon (and the corresponding popup menu “Store MCDM column”) is also activated and its purpose is to store permanently the MCDM score column in the dataset, for future use. If this step is fulfilled, the column will be stored and visualized in the data grid and a new empty MCDM score column will be also generated at the rightmost side.

If the PCA Score plot, calculated here for ranking purposes, is visualized before the calculation of MCDM, every object (chemical) can be marked with its ID or name (name in Figure 47, left part).





**Figure 47.** PCA Score plots calculated for ranking purposes

When the MCDM score has been calculated, on the score plot the three colors corresponding to three different ranking ranges will be activated (Figure 47, right part). By default the ranges equally divides the whole range of the MCDM scores order. For example, in the right Figure 47, 190 molecules are used for ranking: the red color MCDM ordered scores range from 1 to 63, the yellow ones range from 64 to 126 and finally the blue ones from 127 to 190. By clicking within the colored edit boxes, the user can change these ranges. Colored data points, divided by range, allows the user to spot patterns of the objects within the space of the principal component axis. Note that in the PCA score plot there is also the possibility to visualize within the graph the obtained MCDM order value for every object (chemicals, in our case), selecting the “MCDM order” option in the “Label type” drop box (Figure 47) . This means that if an object has obtained the highest MCDM score, it must be the first in the resulting MCDM order, and thus identified with the number 1 (also available in the Ranking dialog, Figure 46, identified as “No”).

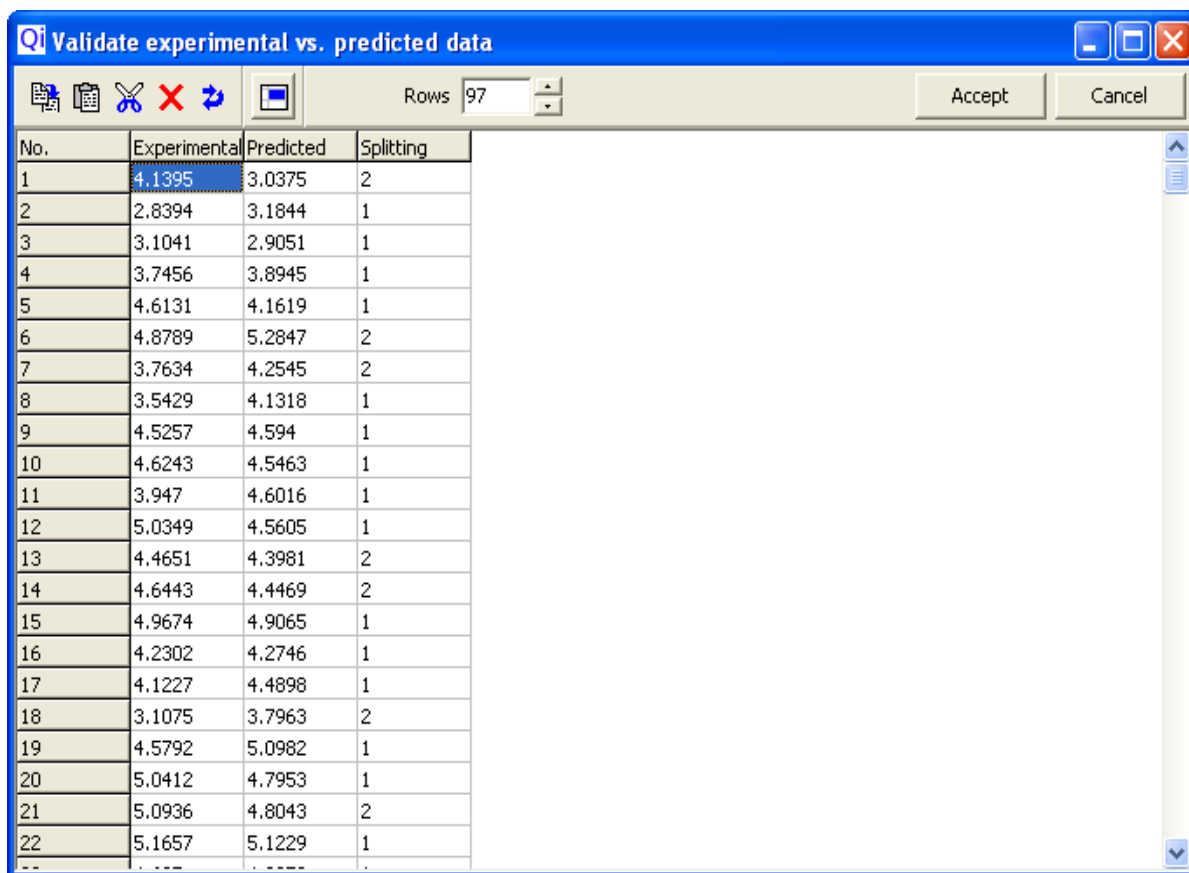
Pressing the  button, all the selected data will be copied in the clipboard.

## 9.4 Validate experimental vs. predicted data

Using the following option it is possible to calculate the validation criteria of a simple double column of data containing experimental and predicted data, which can be the result of whatever user model.

Pressing the  (or selecting Validate experimental vs. predicted data from the main menu) the

following dialog box appears, allowing the user entering experimental and predicted data and, optionally, a splitting column (if not used it must be left blank).



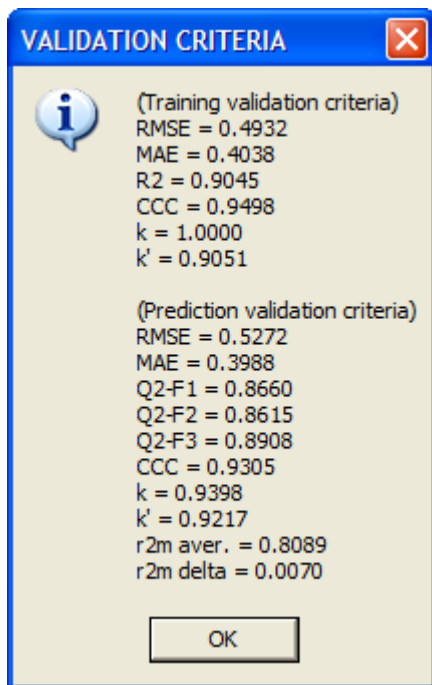
No.	Experimental	Predicted	Splitting
1	4.1395	3.0375	2
2	2.8394	3.1844	1
3	3.1041	2.9051	1
4	3.7456	3.8945	1
5	4.6131	4.1619	1
6	4.8789	5.2847	2
7	3.7634	4.2545	2
8	3.5429	4.1318	1
9	4.5257	4.594	1
10	4.6243	4.5463	1
11	3.947	4.6016	1
12	5.0349	4.5605	1
13	4.4651	4.3981	2
14	4.6443	4.4469	2
15	4.9674	4.9065	1
16	4.2302	4.2746	1
17	4.1227	4.4898	1
18	3.1075	3.7963	2
19	4.5792	5.0982	1
20	5.0412	4.7953	1
21	5.0936	4.8043	2
22	5.1657	5.1229	1

**Figure 48.** Dialog for adding experimental and predicted data for validation.

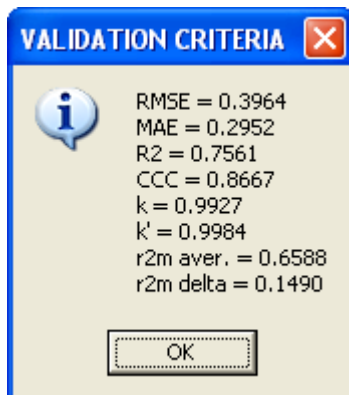
For entering the data, the dialog works in the same manner as in section 9.2 “Apply developed models”, so it is suggested to read it as a reference.

Pressing the button “Accept” the validation criteria are calculated and a dialog box with the results appears.

If a splitting column is provided, the dialog box shows the validation criteria separating those of the training set from those of the prediction set, as in the following example.



If no splitting column is provided, no prediction set validation criteria can be calculated (e.g.  $Q^2_{Fn}$ ), so the list is shorter, as in the following example.



## **Additional information**

QSARINS can be used for every modeling work involving Multiple Linear Regression (MLR) calculations, based on Genetic Algorithm for variable selection and Ordinary Least Squares (OLS) as modelling method. Other chemometric tools (Principal Component Analysis (PCA), Multicriteria Decision Macking (MCDM)) for explorative analysis and ranking are also implemented, thus it is not limited to Quantitative Structure-Activity Relationships (QSAR) studies. The objects studied in QSAR modeling are chemicals, but they could be any kind of objects in other modeling studies.

QSARINS-Chem is the module where 2780 chemicals, studied by the Insubria group, are available with their 3D structure and experimental responses. In addition, 23 QSAR models of environmental end-points, based on a free software for molecular descriptors (PaDEL-Descriptor 2.18) are available. These models, supported by their QMRF, can be applied for any new chemical.

It is important to note that any user can upload personal data sets and models and use QSARINS to manage them for storing, visualization, modeling, ranking etc.

## **Contacts**

Contacts details for additional information:

- Paola Gramatica (content and general philosophy of the QSAR approach in QSARINS)

e-mail: [paola.gramatica@uninsubria](mailto:paola.gramatica@uninsubria), [gramaticapaola@gmail.com](mailto:gramaticapaola@gmail.com)

- Nicola Chirico (software development and use)

e-mail: [nicola.chirico@uninsubria.it](mailto:nicola.chirico@uninsubria.it)

- Stefano Cassani (practical application of QSARINS for modeling)

e-mail: [stefano.cassani@uninsubria.it](mailto:stefano.cassani@uninsubria.it)

## References

- Atkinson, A.C., 1985, Plots, transformations, and regression : an introduction to graphical methods of diagnostic regression analysis in Plots, Clarendon Press, New York : Oxford University Press.
- Chirico, N., Gramatica, P. 2011. Real External Predictivity of QSAR models: How to evaluate it? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *J. Chem. Inf. Model.* 51, 2320-2335.
- Chirico, N., Gramatica, P. 2012. Real External Predictivity of QSAR Models. Part 2. New inter-comparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* 52, 2044-2058.
- Consonni, V. Ballabio, D., Todeschini R. 2009. Comments on definition of Q2 parameter for QSAR validation. *J. Chem. Inf. Model.* 49, 1669-1678.
- Consonni, V., Ballabio, D. Todeschini, R. 2010. Evaluation of model predictive ability by external validation techniques. *J. Chemom.*, 24, 194–201.
- Friedman, J. H. , 1991. Multivariate adaptive regression splines. *Ann. Stat.*, 19, 1-141.
- DRAGON software v.5.5, Todeschini, R., Consonni, V., Mauri, A., Pavan, M. 2007, Talete srl, Milan, Italy, [www.talete.mi.it](http://www.talete.mi.it)
- Golbraikh, A., Tropsha, A. 2002. Beware of q2. *J. Mol. Graphics Model.*, 20, 269-276.
- Gramatica, P., Pilutti, P., Papa, E., 2004. Validated QSAR Prediction of OH Tropospheric degradability: splitting into training-test set and consensus modeling. *J. Chem Inf. Comput. Sci.* 44, 1794-1802.
- Gramatica, P. 2007. Principles of QSAR models validation: internal and external. *QSAR & Comb.Sci.* 26(5), 694-701
- Gramatica P. 2009. Chemometric Methods and Theoretical Molecular Descriptors in Predictive QSAR Modeling of the Environmental Behaviour of Organic Pollutants, Chapter 12 in *Recent Advances in QSAR Studies* , Tomasz Puzyn - Jerzy Leszczynski - Mark T.D. Cronin Eds., (Challenges and Advances in Computational Chemistry and Physics), Springer-Verlag New York Inc, Nov. pp. 327-366.
- Gramatica P. 2012. Modeling Chemicals in the Environment. Chap. 17 in *Drug Design Strategies-Quantitative Approaches*, D.J.Livingstone and A.M.Davies Eds., RSC Pub., pp. 458-478.

Gramatica, P., 2013. On the Development and Validation of QSAR Models. Chap. 21 in *Computational Toxicology: Volume II, Methods in Molecular Biology*, vol. 930, Brad Reisfeld and Arthur N. Mayeno (eds.), DOI 10.1007/978-1-62703-059-5\_21, Springer Science+Business Media, LLC, N.Y. (USA).

Gramatica P., Cassani S., Roy P.P., Kovarich S., Yap C. W., Papa E. 2012. QSAR Modeling is not « Push a button and find a correlation » : a case study of toxicity of (benzo-)triazoles on algae. *Mol. Inf.*, 31, 817 – 835, DOI: 10.1002/minf.201200075

Gramatica, P., Chirico, N., Papa, E., Cassani, S. Kovarich, S., 2013. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, *J. Comput. Chem. (Software News and Updates)*, 2013, 34, 2121-2132 DOI:10.1002/jcc.23361

Gramatica, P., Cassani, S. Chirico, N., 2014. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS *J. Comput. Chem. (Software News and Updates)*, Early View, DOI: 10.1002/jcc.23576

Haupt, R.L., Haupt, S.E., 2004. Practical Genetic Algorithms. 2<sup>nd</sup> Ed. Wiley-Interscience, New Jersey, United States.

Jackson, J.E., A User's Guide to Principal Component. 1991, Wiley, New York, United States.

Keller, H.R., Massart, D.L., Brans, J.P.. 1991. Multicriteria decision making: a case study. *Chemom. Int. Lab. Syst.* 11, 175-189.

Li, J., Lei, B., Liu, H., Li, S., Yao, X., Liu, M., Gramatica, P., 2008. QSAR Study of Malonyl-CoA Decarboxylase Inhibitors Using GA-MLR and a New Strategy of Consensus Modeling. *J. Comput. Chem.* 29, 2636-2647.

Lin, L.I. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45, 255-268.

Ojha, P.K., Mitra, I., Das, R.N., Roy, K. 2011. Further exploring r<sup>2</sup><sub>m</sub> metrics for validation of QSPR models. *Chemom. Intell. Lab. Syst.*, 107, 194-205.

Papa, E., Gramatica, P. 2010. QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure. *Green Chemistry*, 12, 836-843, DOI: 10.1039/b923843c

QSPR-THESAURUS Online Platform, 2011. Available at: <http://qspr-thesaurus.eu/login/show.do?render-mode=full>

Rücker, C., Rücker, G., Meringer, M. 2007.  $\gamma$ -Randomization and Its Variants in QSPR/QSAR, *J. Chem. Inf. Model.* 47, 2345-2357.

Schüürmann, G., Ebert, R., Chen, J., Wang, B., Kühne R. 2008. External Validation and Prediction Employing the Predictive Squared Correlation Coefficients Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.*, 48, 2140–2145.




Shi, L.M.; Fang, H., Tong, W.; Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.L., Sheehan, D.M. 2001. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci.*, 41, 186-195.

Todeschini, R., Consonni, V., Maiocchi, A. 1999. The K correlation index: theory development and its application in chemometrics. *Chemom. Int. Lab. Syst.* 46, 13-29.




Wehrens, R., Putter, H., Buydens, L.M.C. 2000. The bootstrap: a tutorial. *Chemom. Int. Lab. Syst.* 54, 35-52.

Yap, C.W. 2011. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32, 1466-1474. Available online at <http://padel.nus.edu.sg/software/padeldescriptor/index.html>

## Figure Captions

<b>Figure 1.</b> Main window of QSARINS	3
<b>Figure 2.</b> Import data dialog	7
<b>Figure 3.</b> View data window	9
<b>Figure 4.</b> Graph of variable profile	10
<b>Figure 5.</b> Histogram bar chart	11
<b>Figure 6.</b> View variables scatter plot	12
<b>Figure 7.</b> Correlation matrix of the imported data	13
<b>Figure 8.</b> Data setup window	14
<b>Figure 9.</b> PCA (Score plot and Loading plot) of the chemicals in the selected descriptors space	17
<b>Figure 10.</b> Variable selection dialog	18
<b>Figure 11.</b> All Subset-Models calculation dialog	21
<b>Figure 12.</b> Genetic Algorithm-Models calculation dialog	22
<b>Figure 13a.</b> Population of calculated models (with Fitting and some CV criteria)	23
<b>Figure 13b.</b> Population of calculated models (with some CV and external criteria)	24
<b>Figure 14.</b> Relative frequency of descriptors in the population of calculated models	27
<b>Figure 15.</b> Plot of $R^2$ and $Q^2_{LOO}$ average values vs. number of modeling variables	28
<b>Figure 16.</b> Plot of $Q^2_{LOO}$ vs LOF	29
<b>Figure 17.</b> View counting of the out of limit objects	30
<b>Figure 18.</b> Tabulated values of structural and response outliers for each model	31
<b>Figure 19.</b> Dialog for internal validation	33
<b>Figure 20.</b> Dialog for setting the external validation thresholds	34
<b>Figure 21.</b> Plot of MCDM ext vs. MCDM fit	36
<b>Figure 22.</b> Single model dialog box	37
<b>Figure 23.</b> Single model file (.smd) folder location	38
<b>Figure 24.</b> Single model dialog box, where PC columns became available	41
<b>Figure 25.</b> Scatter plot of experimental vs. predicted data by LOO (press  icon)	42
<b>Figure 26.</b> Scatter plot of the residuals of the predicted data by LOO (press  icon)	43
<b>Figure 27.</b> Williams plot using data predicted by LOO (press  icon)	44



<b>Figure 28.</b> Example of Insubria graph (press  icon)	45
<b>Figure 29.</b> Scatter plot of LMO models compared to the QSAR model (press  icon)	46
<b>Figure 30.</b> Scatter plot of Y-scrambled models compared to the QSAR model (press  icon)	47
<b>Figure 31.</b> PCA based on residuals of the selected models	48
<b>Figure 32.</b> PCA based on the descriptors of the selected models	49
<b>Figure 33.</b> Consensus model dialog box	50
<b>Figure 34.</b> Database dialog at its first appearance	53
<b>Figure 35.</b> Database dialog when chemicals are queried	54
<b>Figure 36.</b> Database dialog using the Dataset option	57
<b>Figure 37.</b> Example of a 3D view of a chemical	59
<b>Figure 38.</b> Rotating the 3D view of a chemical	60
<b>Figure 39.</b> Zooming the 3D view of a chemical	60
<b>Figure 40.</b> Translating the 3D view of a chemical	60
<b>Figure 41.</b> Display of the atomic radius in the 3D view of a chemical. Upper left: plain. Upper right: 3D-Dots. Lower: 3D-Lines.	62
<b>Figure 42.</b> Example of a spreadsheet for writing a user-defined dataset	63
<b>Figure 43.</b> Dataset files and folders location	64
<b>Figure 44.</b> List of available models	65
<b>Figure 45.</b> Dialog for adding external chemicals to existing model	66
<b>Figure 46.</b> Dialog for calculating the ranking of the molecules	70
<b>Figure 47.</b> PCA Score plots calculated for ranking purposes	73
<b>Figure 48.</b> Dialog for adding experimental and predicted data for validation	74