QSARINS v 2.2.4 Manual

Index

HOW TO USE QSARINS (version 2.2.4, 2019)

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 and in the new version **"**QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS", *Journal of Computational Chemistry*, *Software News and Updates*, Gramatica et al., 2014, *35*, 1036–1044) can be used after obtaining a license from Prof. Paola Gramatica in the registration step in www.qsar.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:

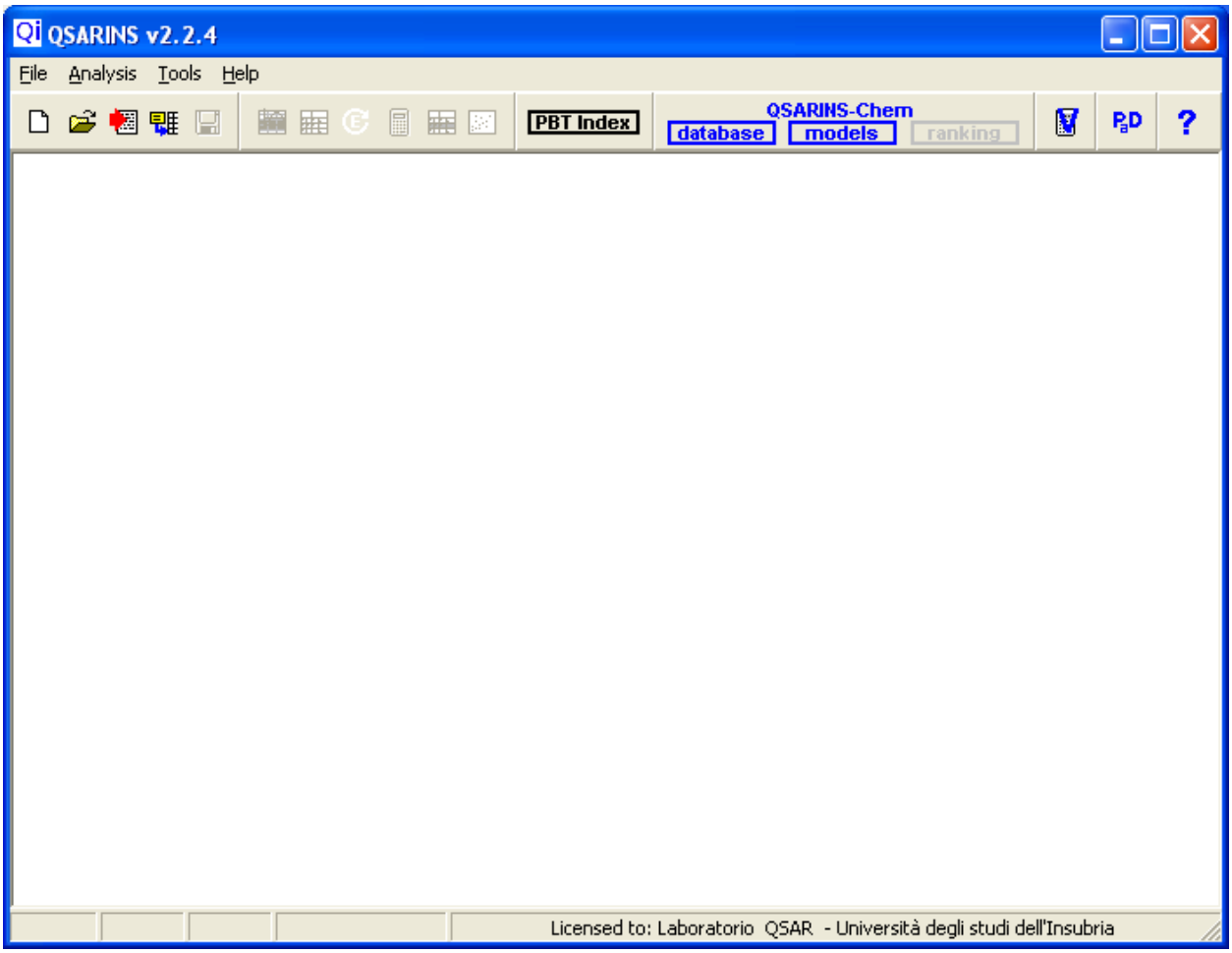


Figure 1*. Main window of QSARINS*

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->Calculate descriptors and import): calculate descriptors using PaDEL-Descriptor software and directly import the data in QSARINS

 (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 Tools->Erase calculated models): erase previously calculated models allowing new variable selection and models calculation (see below)

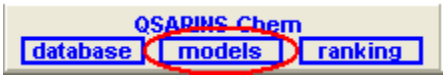
 (or Analysis->Variable selection and models calculation): set the preferred technique and parameters for model calculation (this option is disabled until the data setup has been performed)

 (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 Analysis->View statistics of models from randomized datasets): shows graph and statistics of models generated by randomized datasets.

(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 Tools > Validate experimental vs. predicted data): calculates various validation criteria from columns of experimental vs. predicted data and displays relative graphs.

 (or Tools > Run PaDEL-Descriptor): run PaDEL-Descriptor software.

 (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.

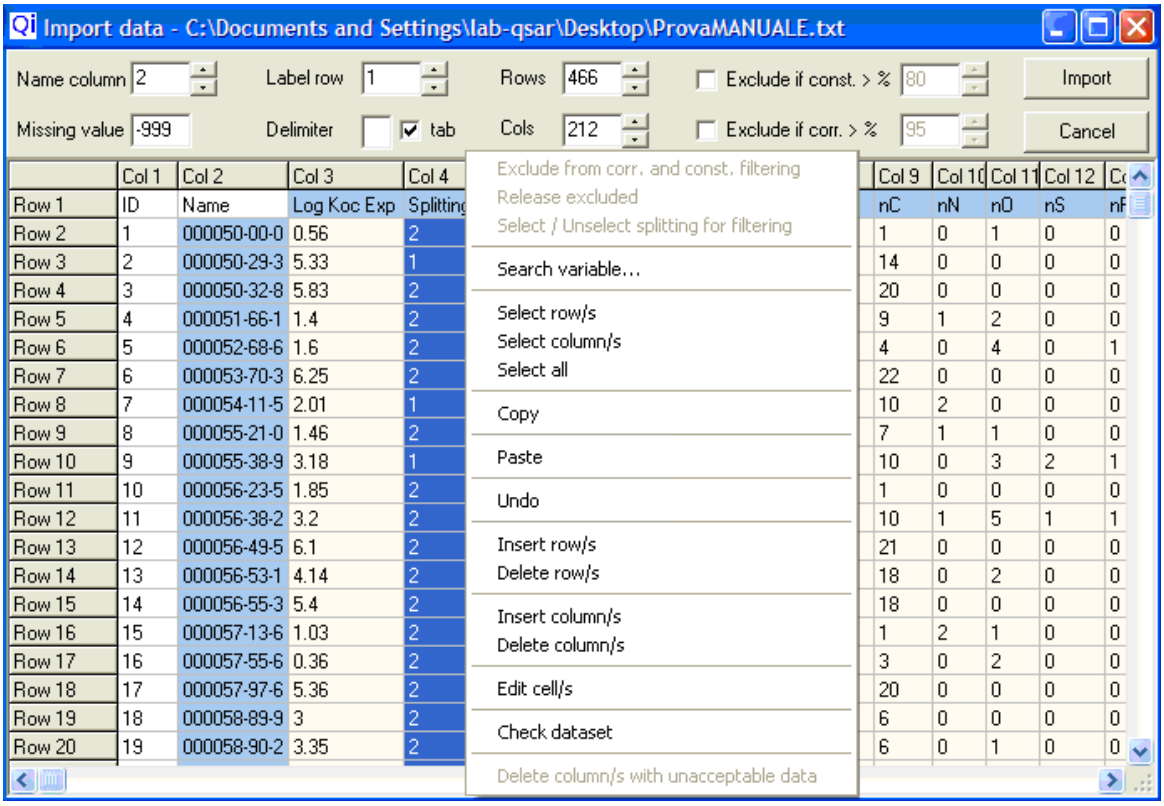
(Note: PaDEL-Descriptor (Yap, 2011), DRAGON software (DRAGON, 2007) and QSPR-THESAURUS online platform (QSPR-THESAURUS, 2011), three packages used by the Insubria group for the calculation of the descriptors, generate output files that are compatible with the aforementioned dataset format). The data 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}x2^{31} elements)

File > Import, or click the icon 

The following dialog will appear:



	Col 1	Col 2	Col 3	Col 4	Col 9	Col 10	Col 11	Col 12	Co
Row 1	ID	Name	Log Koc Exp	Splitting	nC	nN	nO	nS	nF
Row 2	1	000050-00-0	0.56	2	1	0	1	0	0
Row 3	2	000050-29-3	5.33	1	14	0	0	0	0
Row 4	3	000050-32-8	5.83	2	20	0	0	0	0
Row 5	4	000051-66-1	1.4	2	9	1	2	0	0
Row 6	5	000052-68-6	1.6	2	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	10	1	5	1	1
Row 13	12	000056-49-5	6.1	2	21	0	0	0	0
Row 14	13	000056-53-1	4.14	2	18	0	2	0	0
Row 15	14	000056-55-3	5.4	2	18	0	0	0	0
Row 16	15	000057-13-6	1.03	2	1	2	1	0	0
Row 17	16	000057-55-6	0.36	2	3	0	2	0	0
Row 18	17	000057-97-6	5.36	2	20	0	0	0	0
Row 19	18	000058-89-9	3	2	6	0	0	0	0
Row 20	19	000058-90-2	3.35	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). When data are imported the two boxes “Rows” and “Cols” (columns) are automatically set to the pertinent value. The user can change these values at will, for example to add new chemicals (rows) and descriptors or splitting data (columns).

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%). Important note: these two options disable all editing functions (see below) of the cells, i.e. data cannot be edited/pasted, rows and columns number cannot be added, and so on. For this reason, please, be sure having done with dataset editing, otherwise it will be necessary to import it from scratch for having editing options enabled again.

Important note: 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.

After the filtering options in the popup menu (Figure 2) there are some additional tools in the following order:

“Search variable”: search any variable among the columns of data.

“Select row/s” and “Select col/s”: once selected the desired cell(s), these options extend the selection to the whole rows/columns depending of the chosen option.

“Select all”: select the whole dataset.

“Copy”: copy the selected data into the clipboard.

“Paste”: paste data from the clipboard. Data will be pasted starting from the top-left selected cell in the data grid.

“Undo”: undo last operation.

“Insert row/s”: after having selected the desired number or cells, this option adds the same number of rows starting from the bottom of the selection.

“Delete row/s”: delete the selected rows.

“Insert column/s”: after having selected the desired number or cells, this option adds the same number of columns starting from the right of the selection.

“Delete columns”: delete the selected columns.

“Edit cell/s”: fill the selected cells with the value typed in the input dialog box (“Insert new data”).

“Check dataset”: check the consistency of the dataset. All columns containing invalid data (blank, for example) will be highlighted in red. The user can correct the invalid values and check again.

“Delete column/s with unacceptable data”: this option is enabled when unacceptable data are found using the previous option, highlighting in red the corresponding columns. This option automatically deletes the columns containing unacceptable data.

2. Calculate descriptors and import

For descriptors calculation, QSARINS uses the latest available version, at the time of this release, of PaDEL-Descriptor software (Yap, 2011).

Before continuing in this description, it is here recalled that in order to run PaDEL-Descriptor the Java runtime environment (TM Sun Microsystems, Inc., see http://www.java.com) must be installed on your system.

To calculate the descriptors and automatically put the resulting data in the Import dataset session, the user must select the following option from the main window:

File > Calculate descriptors and import, or click the icon 

The user will then asked to select the folder containing the structural files, which must be in one of the format allowed by PaDEL-Descriptor (suggested formats for optimal calculations are MDL MOL and SMILES-.smi). After selecting the folder, a dialog box appears to remember the user loading the “qsarins.cfg” file when the main screen of PaDEL-Descriptor appears. This file is essential to automatically link the output or PaDEL-Descriptor with QSARINS (See Figure 3 to see how to load the configuration file).

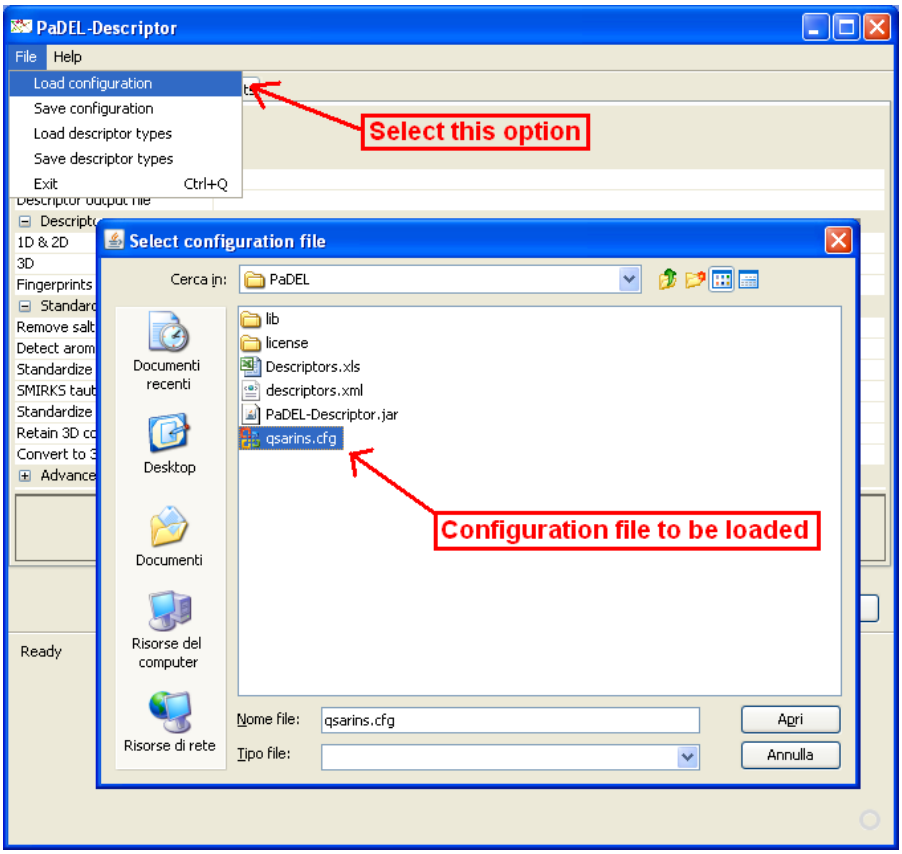


***Figure 3**. Loading of the QSARINS configuration file for PaDEL-Descriptor.*

At this point the user can change the PaDEL-Descriptor options at wish, provided that the Fields “Molecules directory/file” and “Descriptor output file” are not modified, otherwise QSARINS cannot get automatically the calculated descriptors. The selected options can be stored, for future sessions, overwriting the file “qsarins.cfg” (using the “Save configuration” option from the “File” menu of PaDEL-Descriptor). Once the configuration file is loaded, the user can proceed with the calculations of PaDEL-Descriptor; once finished, the descriptors will be automatically loaded in the Import data dialog.

2.1 Updating PaDEL-Descriptor software

After this release of QSARINS it is likely that, after some time, new versions of PaDEL-Descriptor software will be available from the hosting website (http://padel.nus.edu.sg). Provided that the future versions will not change too much from the current software structure (usually authors tend not changing it for compatibility reasons with the previous versions), the user can update (for the QSARINS environment) the version of PaDEL-Descriptor in the following way:

1) Locate the PaDEL folder within QSARINS folder (e.g. Desktop/QSARINS_V_22).

2) Make a copy of the PaDEL folder in another place so, in case of mistakes or incompatibilities with the new version of PaDEL-Descriptor, it can be easily restored.

3) Delete the content of the PaDEL folder.

4) Download and extract the new PaDEL-Descriptor zipped file in the PaDEL folder. It should result in something like below:

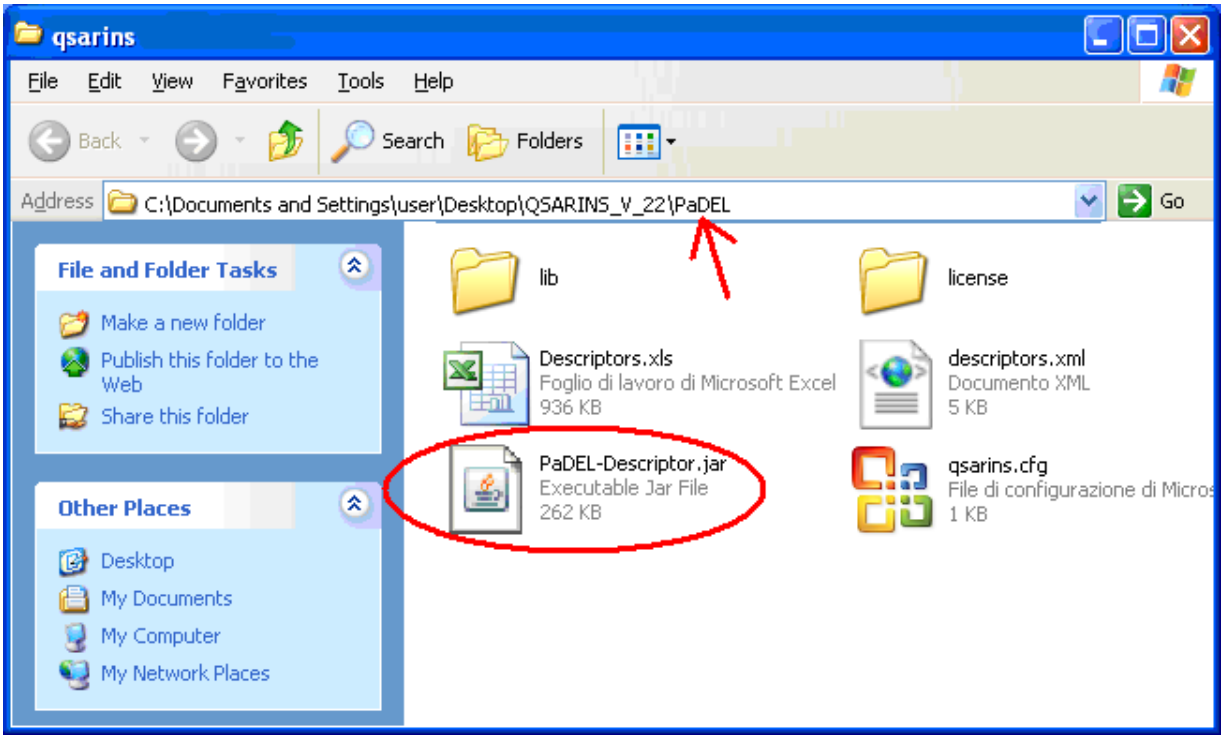


***Figure 4**. Update of PaDEL-Descriptor for QSARINS environment.*

The name “PaDEL-Descriptor.jar” (i.e. the “double clickable” icon that run PaDEL-Descriptor) is essential to be so, otherwise QSARINS will not recognize it (in case in the new version the name has been changed to something else, renaming it “PaDEL-Descriptor.jar” will be likely not to be a concern). It is also essential that “PaDEL-Descriptor.jar” is located in the “PaDEL” folder, as exemplified in Figure 4, otherwise QSARINS will not find it. Other files and subfolders of PaDEL-Descriptor software may change, but it is likely that they will not impact on our purpose.

5) Run PaDEL-Descriptor by double clicking on “PaDEL-Descriptor.jar”

6) Locate and select “Save configuration” option (currently under File menu). Save the current configuration as “qsarins.cfg” in the PaDEL folder.

3. 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:

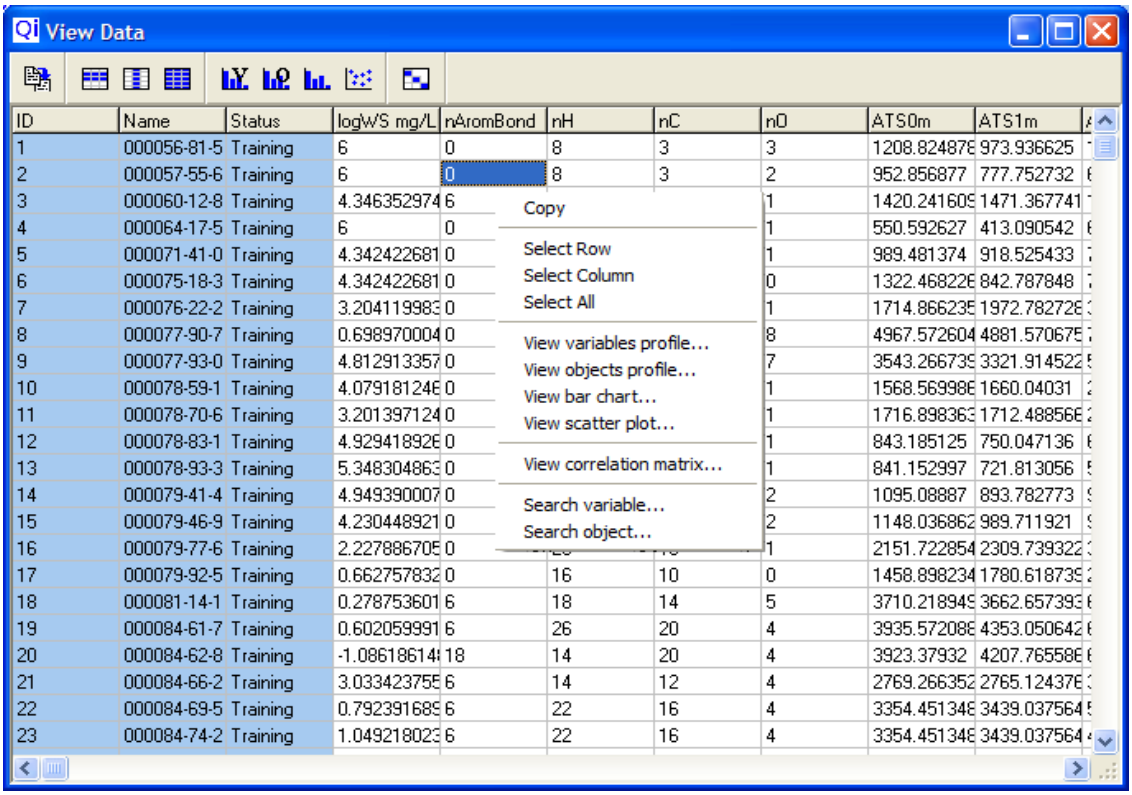


***Figure 5**. View data window*

The main window (Figure 5) 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 5, 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 5 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 6):

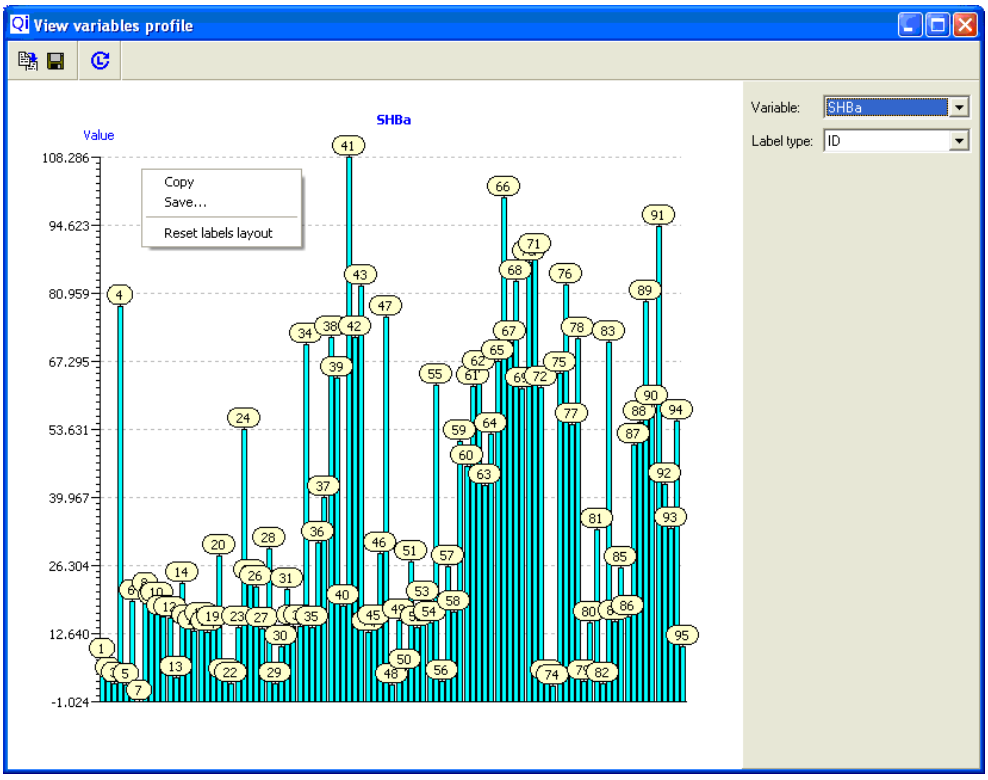


***Figure 6**. Graph of variable profile*

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 5 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 7).

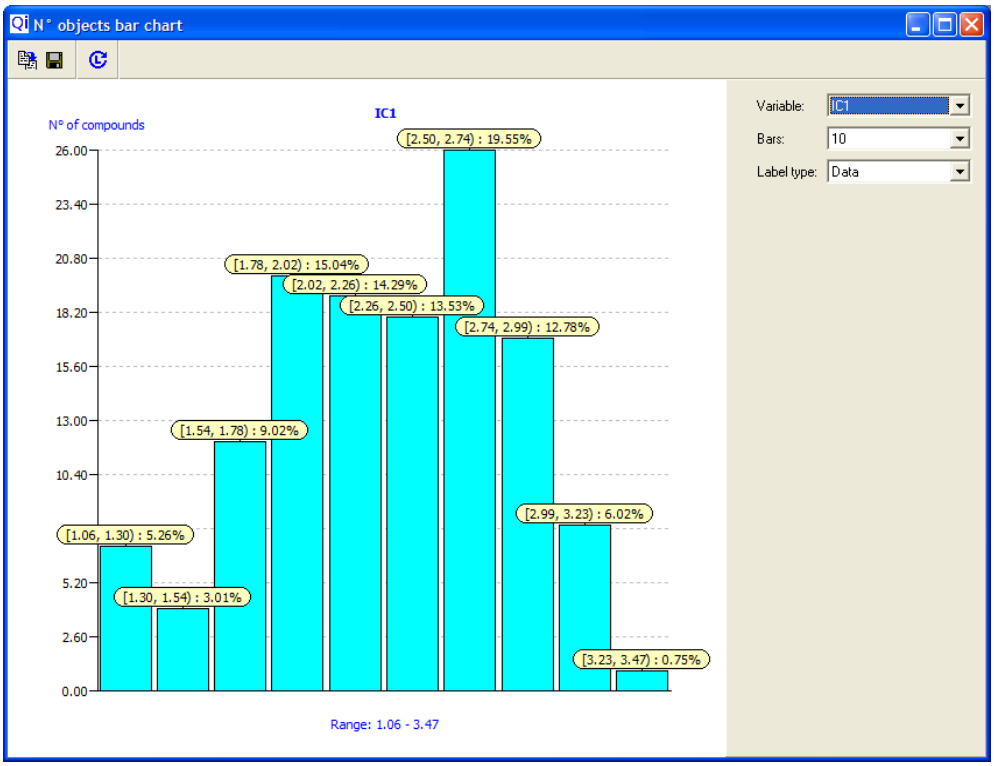


***Figure 7**. Histogram bar chart*

The fourth icon (, or “View scatter plot” in the popup menu) in Figure 5 displays a scatter plot of a couple of data columns chosen by the user, as in the following example (Figure 8):

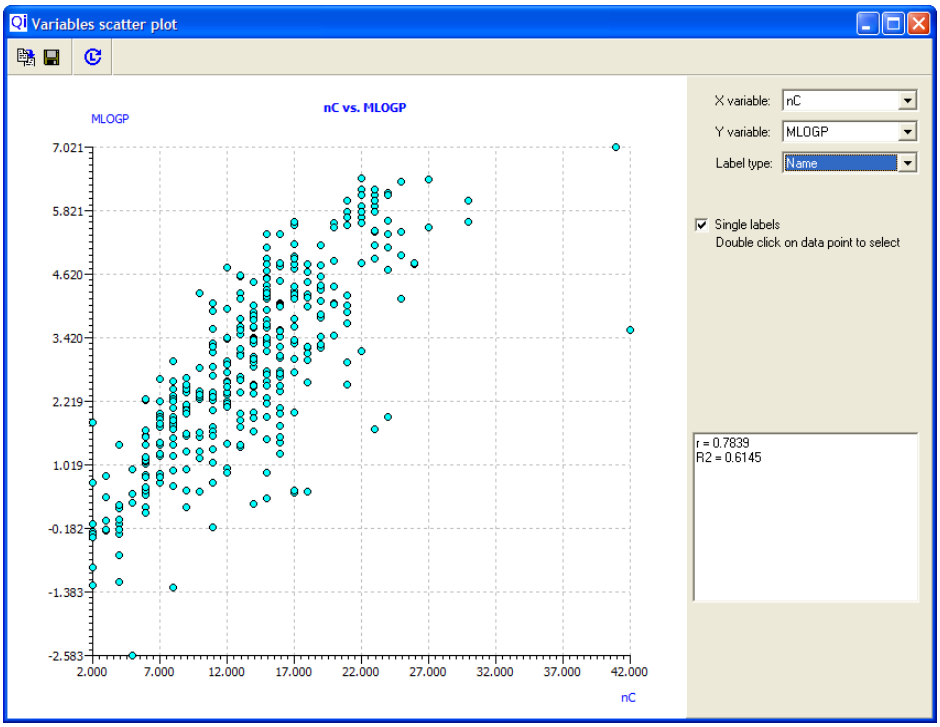


***Figure 8**. Variables scatter plot*

The 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. Additionally, the correlations (r and R^2) between the plotted variables are displayed in the white box.

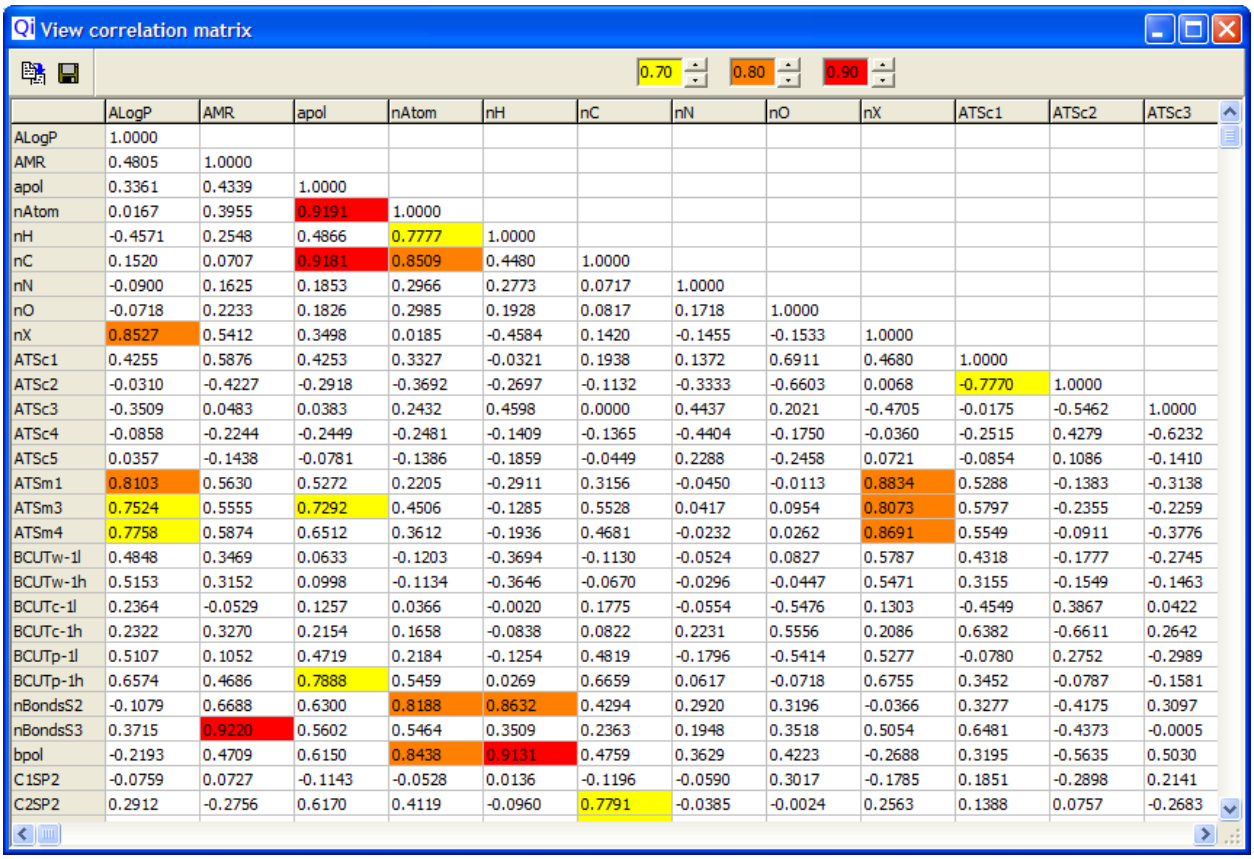
The last icon ( , or “View correlation matrix” in the popup menu) in Figure 5 shows a correlation matrix of all the imported data (Figure 9). The high correlations values are here highlighted with different colors:

- 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 9), allowing to highlight any degree of correlation among the data.



	ALogP	AMR	apol	nAtom	nH	nC	nN	nO	nX	ATSc1	ATSc2	ATSc3
ALogP	1.0000											
AMR	0.4805	1.0000										
apol	0.3361	0.4339	1.0000									
nAtom	0.0167	0.3955	0.9191	1.0000								
nH	-0.4571	0.2548	0.4866	0.7777	1.0000							
nC	0.1520	0.0707	0.9181	0.8509	0.4480	1.0000						
nN	-0.0900	0.1625	0.1853	0.2966	0.2773	0.0717	1.0000					
nO	-0.0718	0.2233	0.1826	0.2985	0.1928	0.0817	0.1718	1.0000				
nX	0.8527	0.5412	0.3498	0.0185	-0.4584	0.1420	-0.1455	-0.1533	1.0000			
ATSc1	0.4255	0.5876	0.4253	0.3327	-0.0321	0.1938	0.1372	0.6911	0.4680	1.0000		
ATSc2	-0.0310	-0.4227	-0.2918	-0.3692	-0.2697	-0.1132	-0.3333	-0.6603	0.0068	-0.7770	1.0000	
ATSc3	-0.3509	0.0483	0.0383	0.2432	0.4598	0.0000	0.4437	0.2021	-0.4705	-0.0175	-0.5462	1.0000
ATSc4	-0.0858	-0.2244	-0.2449	-0.2481	-0.1409	-0.1365	-0.4404	-0.1750	-0.0360	-0.2515	0.4279	-0.6232
ATSc5	0.0357	-0.1438	-0.0781	-0.1386	-0.1859	-0.0449	0.2288	-0.2458	0.0721	-0.0854	0.1086	-0.1410
ATSm1	0.8103	0.5630	0.5272	0.2205	-0.2911	0.3156	-0.0450	-0.0113	0.8834	0.5288	-0.1383	-0.3138
ATSm3	0.7524	0.5555	0.7292	0.4506	-0.1285	0.5528	0.0417	0.0954	0.8073	0.5797	-0.2355	-0.2259
ATSm4	0.7758	0.5874	0.6512	0.3612	-0.1936	0.4681	-0.0232	0.0262	0.8691	0.5549	-0.0911	-0.3776
BCUTw-1l	0.4848	0.3469	0.0633	-0.1203	-0.3694	-0.1130	-0.0524	0.0827	0.5787	0.4318	-0.1777	-0.2745
BCUTw-1h	0.5153	0.3152	0.0998	-0.1134	-0.3646	-0.0670	-0.0296	-0.0447	0.5471	0.3155	-0.1549	-0.1463
BCUTc-1l	0.2364	-0.0529	0.1257	0.0366	-0.0020	0.1775	-0.0554	-0.5476	0.1303	-0.4549	0.3867	0.0422
BCUTc-1h	0.2322	0.3270	0.2154	0.1658	-0.0838	0.0822	0.2231	0.5556	0.2086	0.6382	-0.6611	0.2642
BCUTp-1l	0.5107	0.1052	0.4719	0.2184	-0.1254	0.4819	-0.1796	-0.5414	0.5277	-0.0780	0.2752	-0.2989
BCUTp-1h	0.6574	0.4686	0.7888	0.5459	0.0269	0.6659	0.0617	-0.0718	0.6755	0.3452	-0.0787	-0.1581
nBondsS2	-0.1079	0.6688	0.6300	0.8188	0.8632	0.4294	0.2920	0.3196	-0.0366	0.3277	-0.4175	0.3097
nBondsS3	0.3715	0.9220	0.5602	0.5464	0.3509	0.2363	0.1948	0.3518	0.5054	0.6481	-0.4373	-0.0005
bpol	-0.2193	0.4709	0.6150	0.8438	0.9131	0.4759	0.3629	0.4223	-0.2688	0.3195	-0.5635	0.5030
C1SP2	-0.0759	0.0727	-0.1143	-0.0528	0.0136	-0.1196	-0.0590	0.3017	-0.1785	0.1851	-0.2898	0.2141
C2SP2	0.2912	-0.2756	0.6170	0.4119	-0.0960	0.7791	-0.0385	-0.0024	0.2563	0.1388	0.0757	-0.2683

***Figure 9**. 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.

4. 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 5.1, “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 10) will appear:

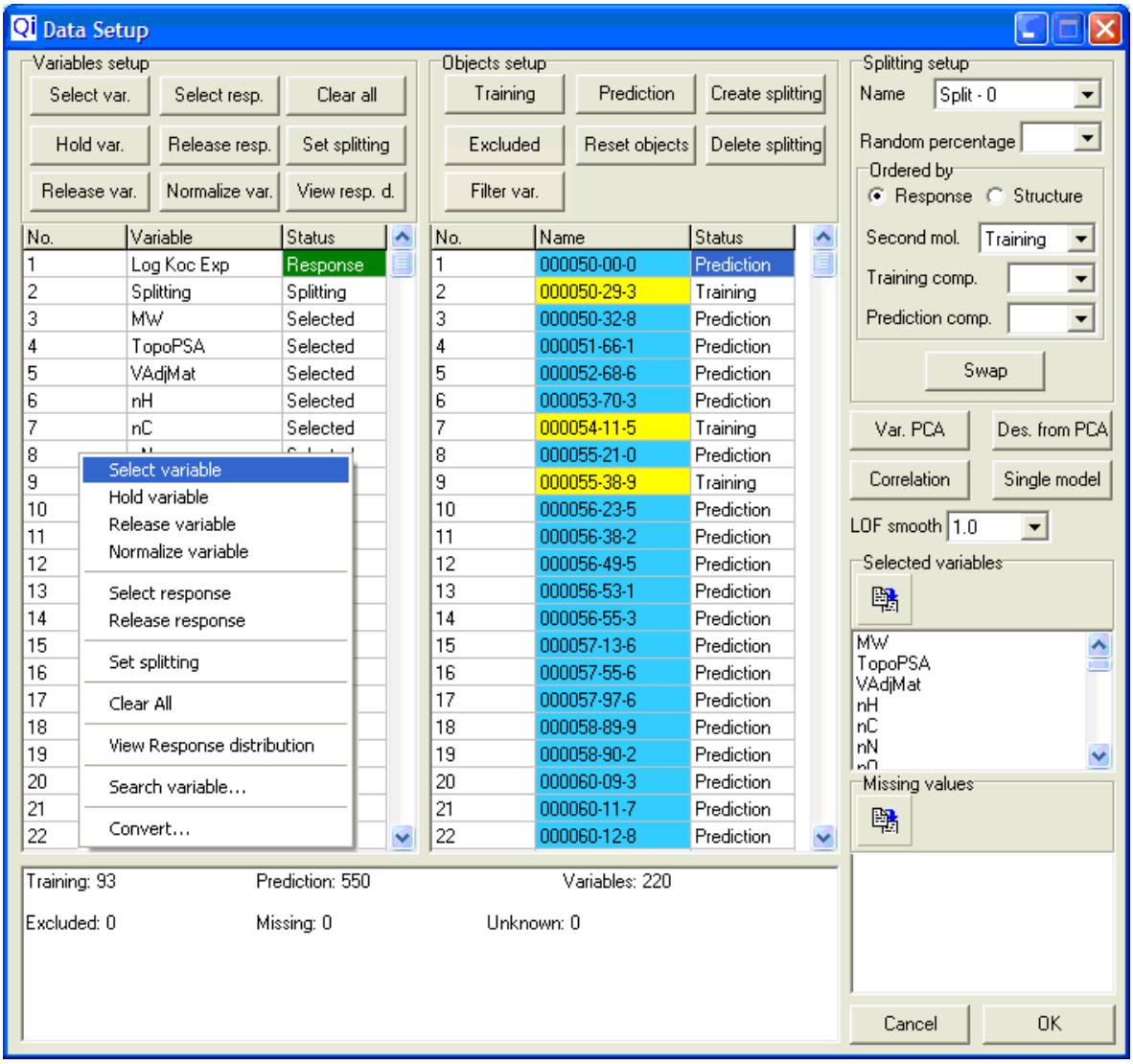
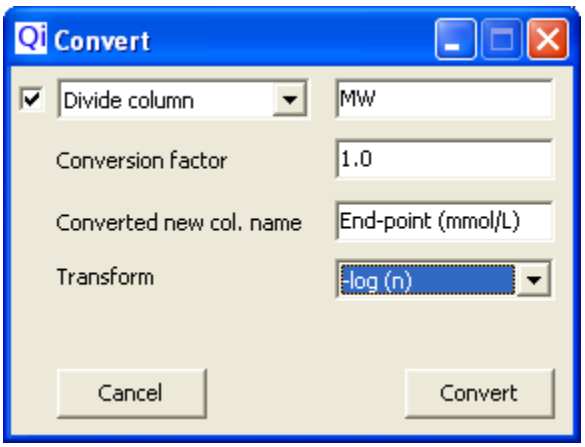


***Figure 10**. Data setup window*

The box on the upper left of Figure 10 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/excluded) by means of the specific buttons

("Select/Hold/Release Var", "Select/Release Resp" and "Set splitting") or using the pop-up menu (mouse right click). By double-clicking on a single cell of the data grid the options Selected/ Hold/Release(empty cell) are cyclically proposed. 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 5.2: "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 "Search variable" option in the pop-up menu allow searching the corresponding variable.

The latest option of the pop-up menu is for unit conversion ("Convert…"). After selecting the variable to be converted (e.g. a response) the following dialog box (with example setup) appears:

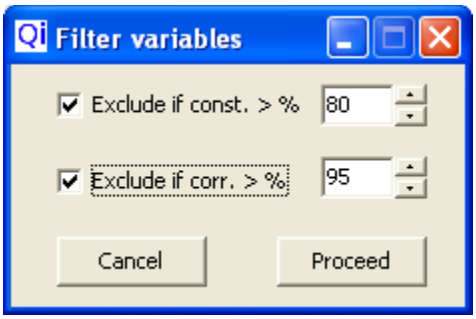


The first list box determines whether the selected column has to divided or multiplied by the corresponding descriptor column (in the example is molecular weight, called "MW" in the dataset). The "Conversion factor" determines by which constant the previous result must be multiplied (in the example the conversion is from g/l to mmol/l). The "Converted column name" is the name of the new converted column that will be generated. The last option is used, if needed, to transform the result as +/- ln or log, or by 1/n (in the example is –log (n)). If a transformation is requested, an additional untransformed column will be automatically added to the dataset.

The subsequent data grid on the right of Figure 10 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 (note: by double-clicking on the cells, training/prediction/excluded are cyclically proposed), 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 last option above the grid is “Filter var.” (Filter variables), that allows filtering the descriptors according to the current splitting. In fact, even though descriptors can be filtered in the import data section (see 1.2 Importing the dataset), afterwards different object splittings can be applied. As a consequence, descriptors that were below the correlation and/or constancy thresholds after dataset import, could fall above the thresholds. This could cause problems in PCA display (data constancy) and/or models with highly correlated descriptors.

Pressing to the corresponding button or menu item, the following dialog will appear.

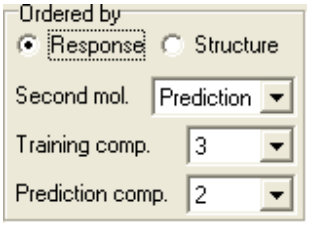


After selecting and adjusting the options to the wanted levels, pressing “Proceed” will automatically unselect the descriptors if they are above the threshold/s.

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:



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.

Concerning all splitting options, the molecules with the lowest and the highest experimental response value are always set automatically as training, to keep the whole response domain while developing the models. Should the user decide otherwise, the status of these molecules must be changed manually.

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 11). 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:

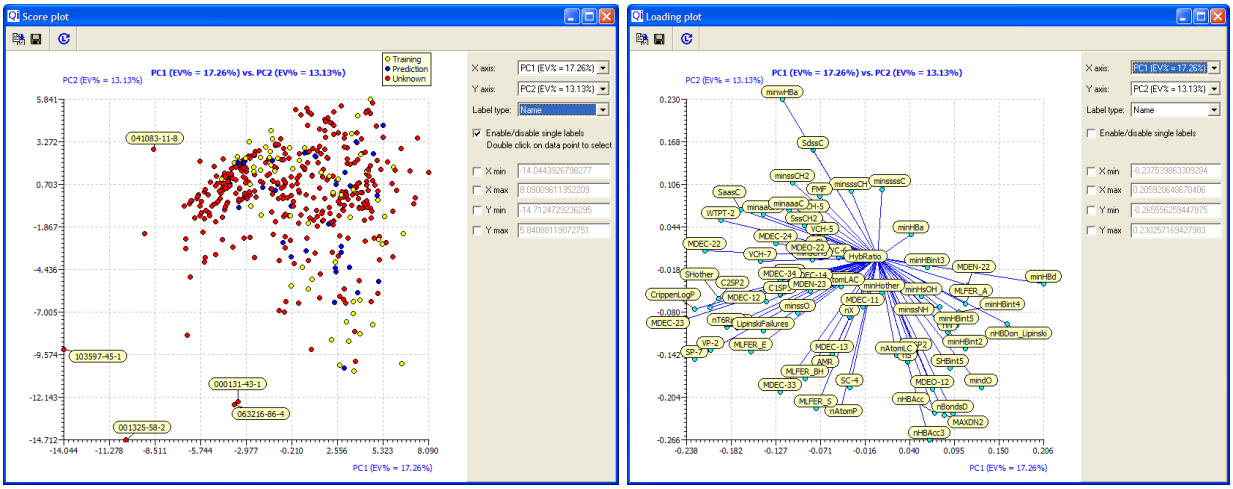


***Figure 11**. PCA (Score plot and Loading plot) of the selected descriptors*

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, 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 (PCR). 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 (Section 3 “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 (10, “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 10 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).

5. Variable selection and Model calculation

5.1. 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 12) will appear:

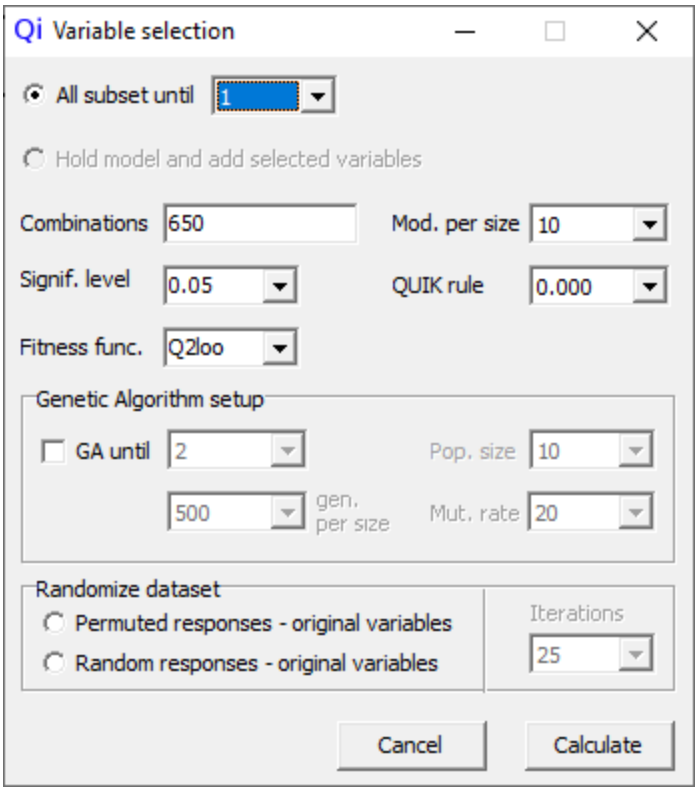


***Figure 12**. Variable selection dialog*

The user can apply the following approaches for variable selection (Figure 12, note: "Randomize dataset" is explained in section 8.)

- 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 4, "Data setup") 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 5.2, "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.

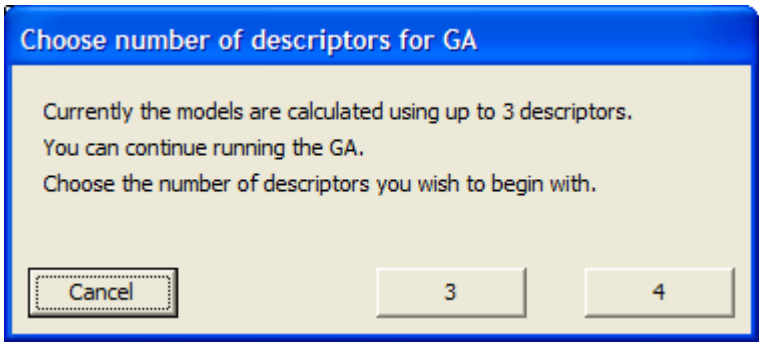
In addition, for the whole variable selection procedure, the user can set:

- the fitness function (Q^2_{LOO}or R^2_{adj} to be maximized, LOF or $\mathrm{RMSE_{CV}}$ 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 (Kxy) is lower than or too similar to the inter-correlation among the descriptors (Kxx). This difference should be the highest as possible. (it is suggested to set a value ≥ 0.05, i.e. Kxy-Kxx ≥ 0.05)
- The minimum acceptable significance level for all equation coefficients and intercept.
- 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 5.2, “Model calculation”) 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 12) 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 12).

5.2. Model calculation

Once the variable selection setup has been completed and the model calculation begins, the following dialog box (Figure 13) appears:

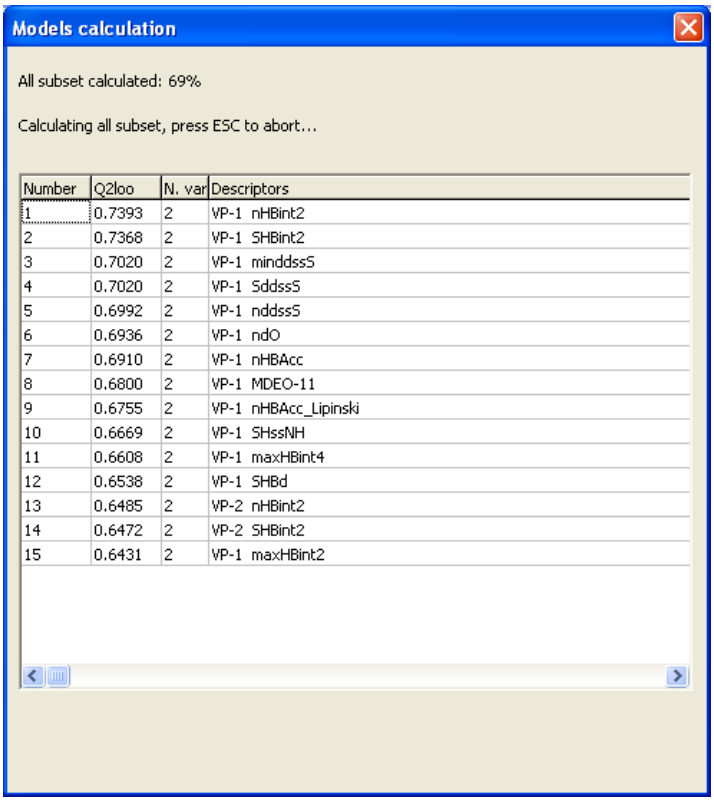


***Figure 13**. All Subset-Models calculation dialog*

The best 15 models, according to the chosen fitness function, are displayed (Figure 13) 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 14):

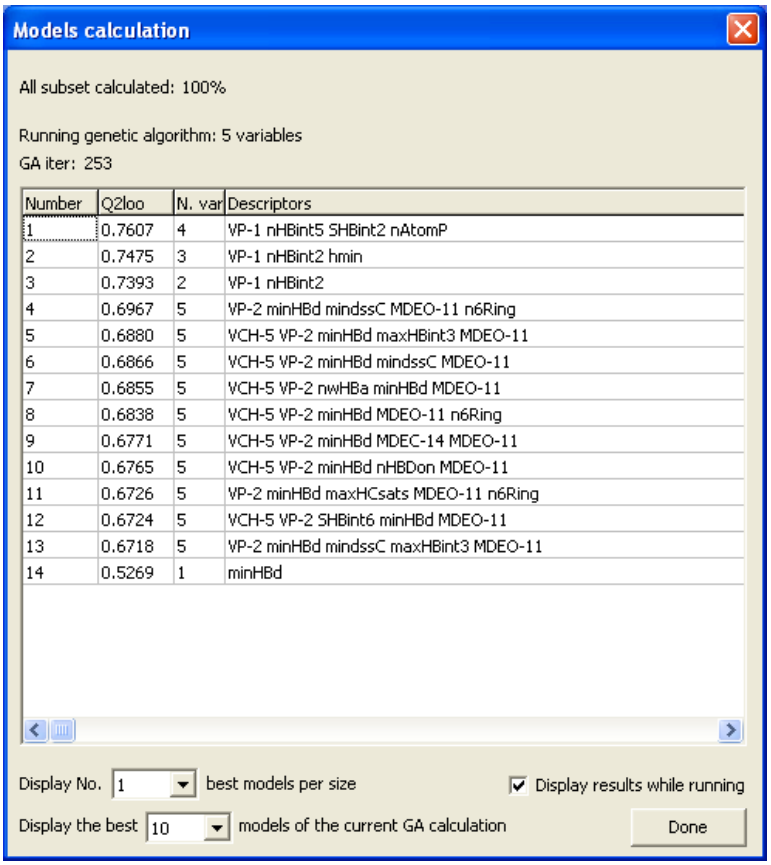


***Figure 14**. Genetic Algorithm-Models calculation dialog*

In this screen, the user can visualize the results of GA while it is running, selecting the “Display results while running” checkbox (Figure 14). 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 14 there is one model for every size till 5 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.

6. 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 (Figures 15a and 15b):

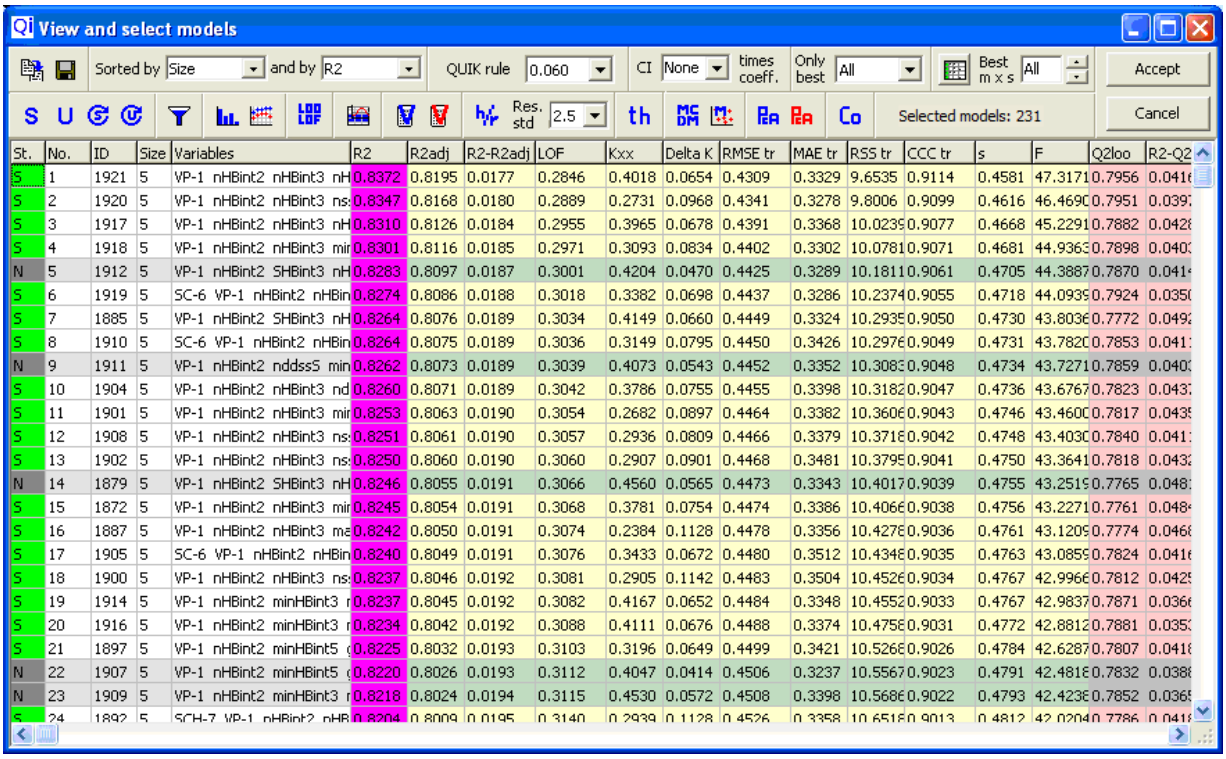
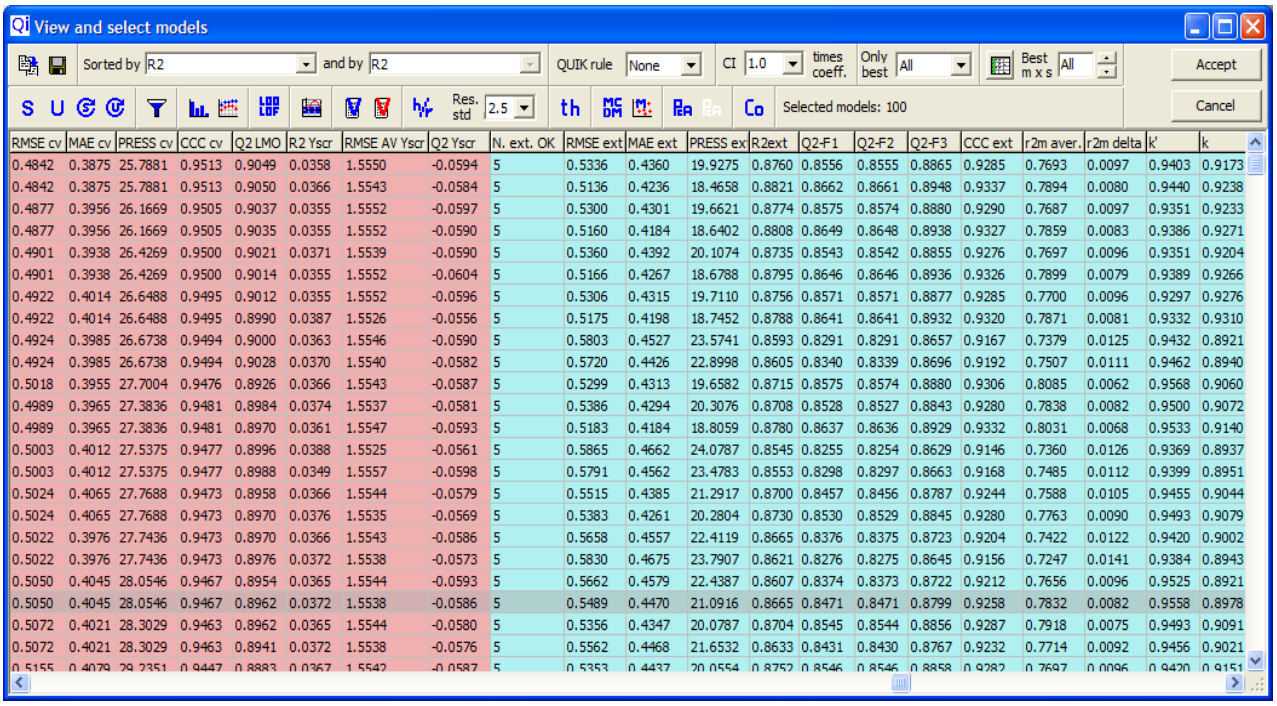


***Figure 15a**. Population of calculated models (with Fitting and some CV criteria)*



RMSE cv	MAE cv	PRESS cv	CCC cv	Q2 LMO	R2 Yscr	RMSE AV Yscr	Q2 Yscr	N. ext. OK	RMSE ext	MAE ext	PRESS ex	R2ext	Q2-F1	Q2-F2	Q2-F3	CCC ext	r2m aver.	r2m delta	k'	k
0.4842	0.3875	25.7881	0.9513	0.9049	0.0358	1.5550	-0.0594	5	0.5336	0.4360	19.9275	0.8760	0.8556	0.8555	0.8865	0.9285	0.7693	0.0097	0.9403	0.9173
0.4842	0.3875	25.7881	0.9513	0.9050	0.0366	1.5543	-0.0584	5	0.5136	0.4236	18.4658	0.8821	0.8662	0.8661	0.8948	0.9337	0.7894	0.0080	0.9440	0.9238
0.4877	0.3956	26.1669	0.9505	0.9037	0.0355	1.5552	-0.0597	5	0.5300	0.4301	19.6621	0.8774	0.8575	0.8574	0.8880	0.9290	0.7687	0.0097	0.9351	0.9233
0.4877	0.3956	26.1669	0.9505	0.9035	0.0355	1.5552	-0.0590	5	0.5160	0.4184	18.6402	0.8808	0.8649	0.8648	0.8938	0.9327	0.7859	0.0083	0.9386	0.9271
0.4901	0.3938	26.4269	0.9500	0.9021	0.0371	1.5539	-0.0590	5	0.5360	0.4392	20.1074	0.8735	0.8543	0.8542	0.8855	0.9276	0.7697	0.0096	0.9351	0.9204
0.4901	0.3938	26.4269	0.9500	0.9014	0.0355	1.5552	-0.0604	5	0.5166	0.4267	18.6788	0.8795	0.8646	0.8646	0.8936	0.9326	0.7899	0.0079	0.9389	0.9266
0.4922	0.4014	26.6488	0.9495	0.9012	0.0355	1.5552	-0.0596	5	0.5306	0.4315	19.7110	0.8756	0.8571	0.8571	0.8877	0.9285	0.7700	0.0096	0.9297	0.9276
0.4922	0.4014	26.6488	0.9495	0.8990	0.0387	1.5526	-0.0556	5	0.5175	0.4198	18.7452	0.8788	0.8641	0.8641	0.8932	0.9320	0.7871	0.0081	0.9332	0.9310
0.4924	0.3985	26.6738	0.9494	0.9000	0.0363	1.5546	-0.0590	5	0.5803	0.4527	23.5741	0.8593	0.8291	0.8291	0.8657	0.9167	0.7379	0.0125	0.9432	0.8921
0.4924	0.3985	26.6738	0.9494	0.9028	0.0370	1.5540	-0.0582	5	0.5720	0.4426	22.8998	0.8605	0.8340	0.8339	0.8696	0.9192	0.7507	0.0111	0.9462	0.8940
0.5018	0.3955	27.7004	0.9476	0.8926	0.0366	1.5543	-0.0587	5	0.5299	0.4313	19.6582	0.8715	0.8575	0.8574	0.8880	0.9306	0.8085	0.0062	0.9568	0.9060
0.4989	0.3965	27.3836	0.9481	0.8984	0.0374	1.5537	-0.0581	5	0.5386	0.4294	20.3076	0.8708	0.8528	0.8527	0.8843	0.9280	0.7838	0.0082	0.9500	0.9072
0.4989	0.3965	27.3836	0.9481	0.8970	0.0361	1.5547	-0.0593	5	0.5183	0.4184	18.8059	0.8780	0.8637	0.8636	0.8929	0.9332	0.8031	0.0068	0.9533	0.9140
0.5003	0.4012	27.5375	0.9477	0.8996	0.0388	1.5525	-0.0561	5	0.5865	0.4662	24.0787	0.8545	0.8255	0.8254	0.8629	0.9146	0.7360	0.0126	0.9369	0.8937
0.5003	0.4012	27.5375	0.9477	0.8988	0.0349	1.5557	-0.0598	5	0.5791	0.4562	23.4783	0.8553	0.8298	0.8297	0.8663	0.9168	0.7485	0.0112	0.9399	0.8951
0.5024	0.4065	27.7688	0.9473	0.8958	0.0366	1.5544	-0.0579	5	0.5515	0.4385	21.2917	0.8700	0.8457	0.8456	0.8787	0.9244	0.7588	0.0105	0.9455	0.9044
0.5024	0.4065	27.7688	0.9473	0.8970	0.0376	1.5535	-0.0569	5	0.5383	0.4261	20.2804	0.8730	0.8530	0.8529	0.8845	0.9280	0.7763	0.0090	0.9493	0.9079
0.5022	0.3976	27.7436	0.9473	0.8970	0.0366	1.5543	-0.0586	5	0.5658	0.4557	22.4119	0.8665	0.8376	0.8375	0.8723	0.9204	0.7422	0.0122	0.9420	0.9002
0.5022	0.3976	27.7436	0.9473	0.8976	0.0372	1.5538	-0.0573	5	0.5830	0.4675	23.7907	0.8621	0.8276	0.8275	0.8645	0.9156	0.7247	0.0141	0.9384	0.8943
0.5050	0.4045	28.0546	0.9467	0.8954	0.0365	1.5544	-0.0593	5	0.5662	0.4579	22.4387	0.8607	0.8374	0.8373	0.8722	0.9212	0.7656	0.0096	0.9525	0.8921
0.5050	0.4045	28.0546	0.9467	0.8962	0.0372	1.5538	-0.0586	5	0.5489	0.4470	21.0916	0.8665	0.8471	0.8471	0.8799	0.9258	0.7832	0.0082	0.9558	0.8978
0.5072	0.4021	28.3029	0.9463	0.8962	0.0365	1.5544	-0.0580	5	0.5356	0.4347	20.0787	0.8704	0.8545	0.8544	0.8856	0.9287	0.7918	0.0075	0.9493	0.9091
0.5072	0.4021	28.3029	0.9463	0.8941	0.0372	1.5538	-0.0576	5	0.5562	0.4468	21.6532	0.8633	0.8431	0.8430	0.8767	0.9232	0.7714	0.0092	0.9456	0.9021
0.5155	0.4079	29.2351	0.9447	0.8883	0.0367	1.5542	-0.0587	5	0.5353	0.4437	20.0554	0.8752	0.8546	0.8546	0.8858	0.9282	0.7697	0.0096	0.9420	0.9151

***Figure 15b**. 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^2_{LOO} and so on, R^2 in Figure 15a), 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 15a and 15b). During the all subset and GA calculation, the models are labelled with a unique ID (see ID column in Figure 15a) 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 7, "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^2_{LOO} 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 5.1, “Variable selection”). 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 10, “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 filters), 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 number of selected models is available under the “Best m x s” box.

The subsequent row of icons in Figures 15 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, combined modeling, 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 16), where the bars are the relative frequency of the descriptors used by all the models having a size of 3 variables:

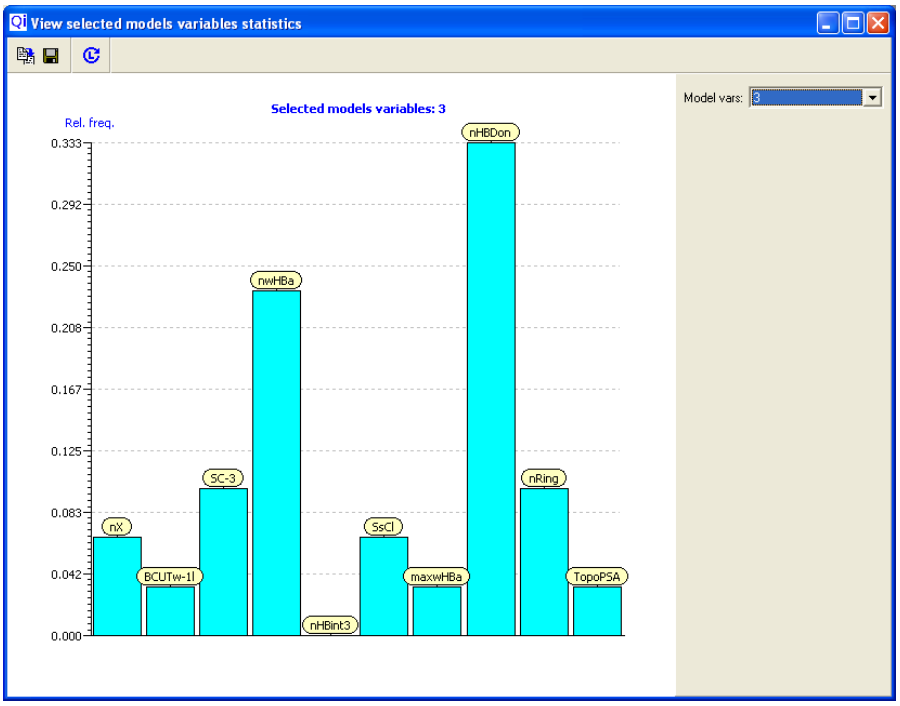


***Figure 16**. Relative frequency of descriptors in the population of calculated models*

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 any overfitting in the resulting populations of models exists. 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 17), Q^2_{LOO}will decrease, showing that the predictive power of the model could be deceptive.

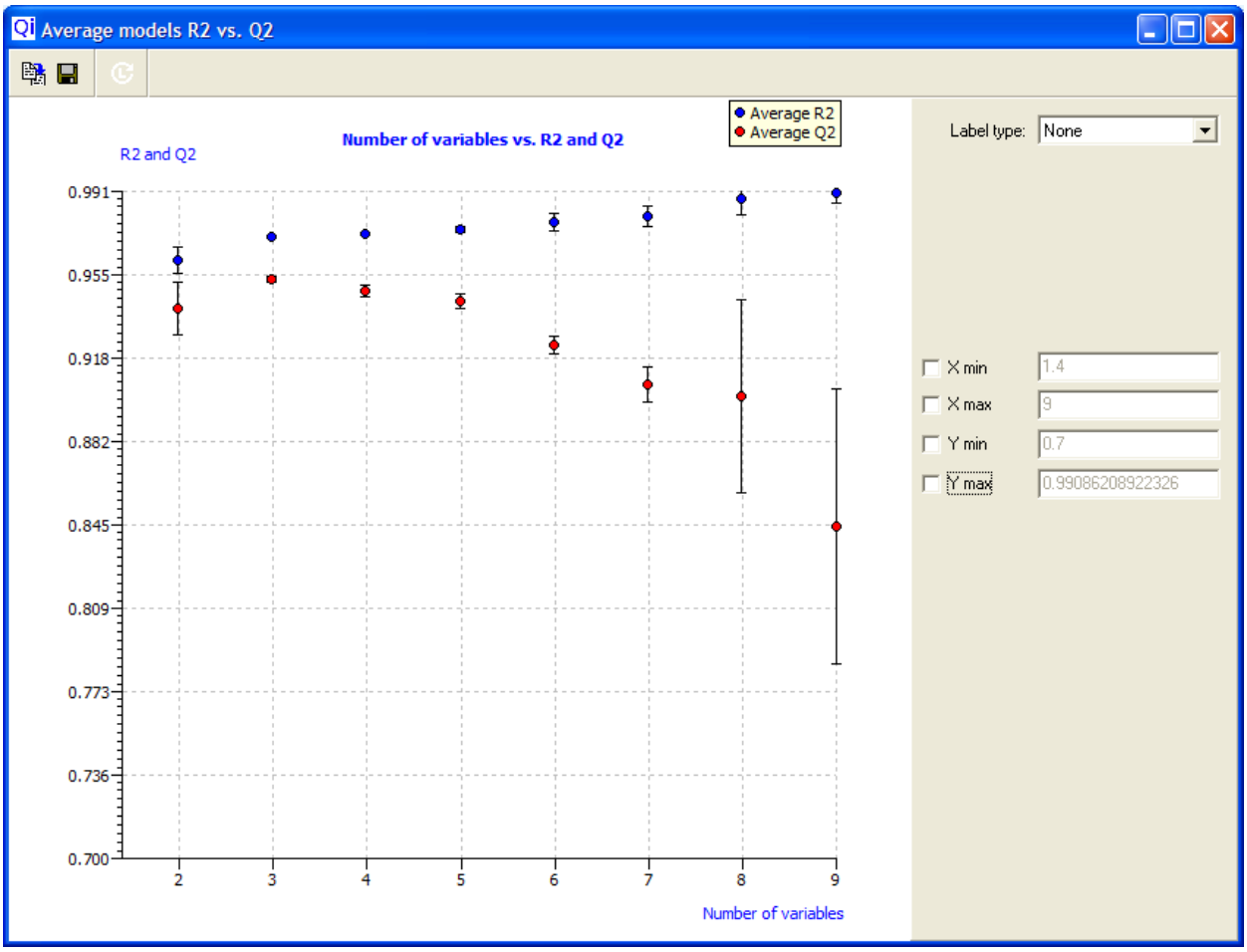


***Figure 17**. Plot of R^2 and Q^2_{LOO} average values vs. number of modeling variables*

: 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).

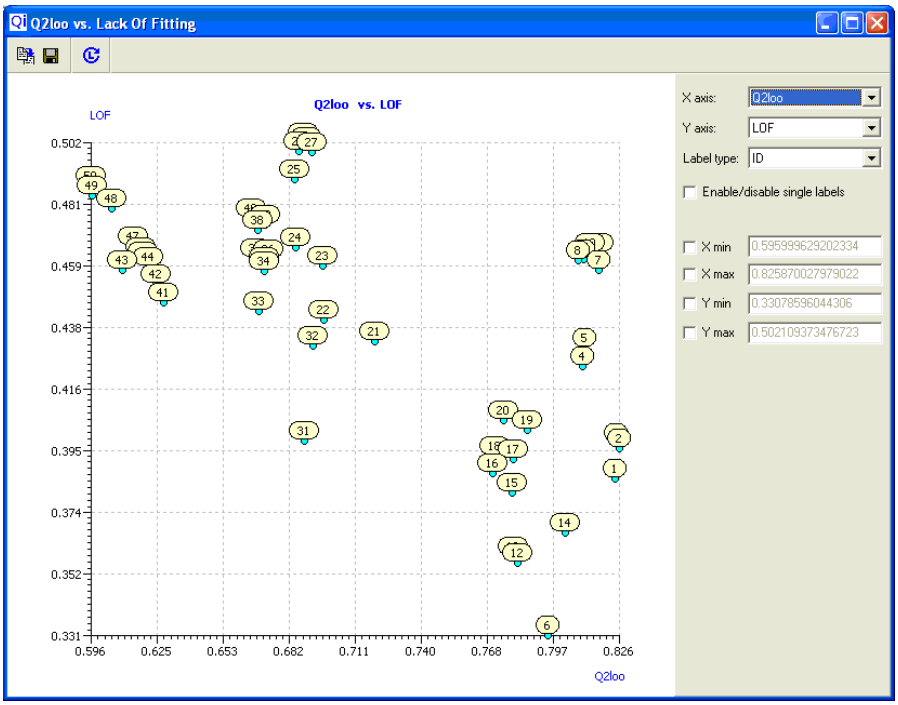


***Figure 18**. Plot of Q^2_{LOO} vs LOF*

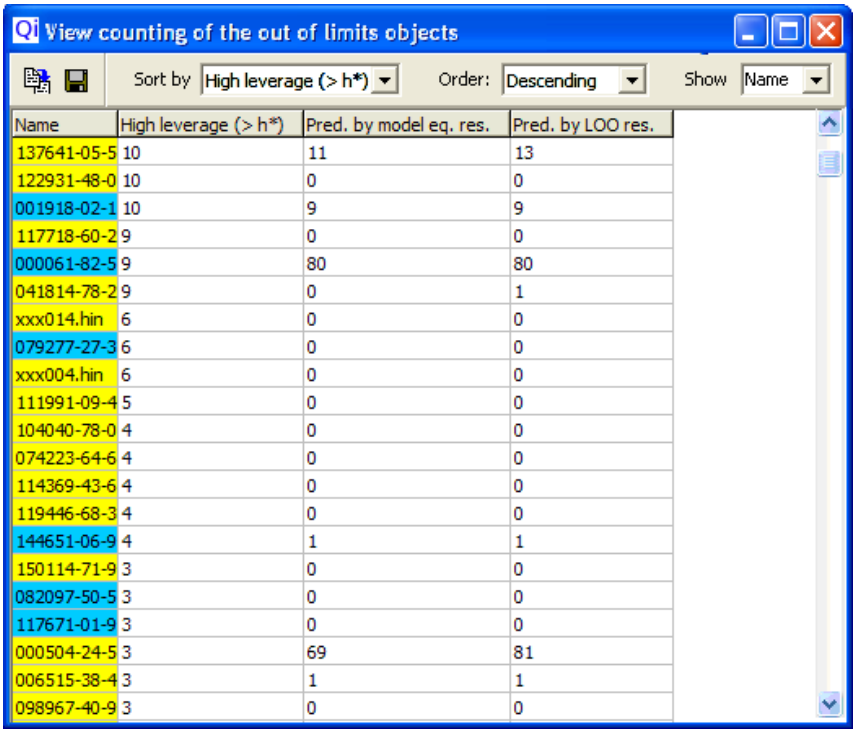
This graph (Figure 18) can help in the selection of the models with the best compromise between high predictivity (high Q^2_{LOO}) and small dimension (low LOF) (as, for example, the model 6 in the figure above).

: visualizes the performances of the selected model, both in tabulated and in graphical form, as explained in Section 10 (“Analysis of Single Models”).

: applies the internal validation tools (LMO, Y-scrambling and random responses/descriptors techniques) on the selected models: see section 7 (“Model validation”) for further details.

: applies the external validation tools (R^2_{EXT}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, *CCC*, $\overline{r^2_m}$ and Δr^2_m) on the selected models: see section 7 ("Model validation") for further details.

: calculates and count the structural outliers (X-outliers, i.e. those with leverage values above the h* value (see Section 10, "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 19):



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 19**. View counting of the out of limits objects.*

The first column in Figure 19 reports the name of the chemicals (CAS number in this example; the user can also decide to visualize the molecule ID, selecting it on “Show” dialog box, in the upper right part of Figure 19). 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 20):

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 20**. 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.

Note that if if the same compound is outlier in all or in most models there is a reasonable doubt that the experimental data is wrong and it should be removed .

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 7 (“Model validation”).

The icons and are used for the selection of the most diverse models to be averaged in the Combined modeling , see section 11 (“Combined modeling”) for further details.

7. Model validation

Since the number of calculations required by the model validation is very large and involve random data, at this point it is suggested to save the project (as .qsi), in order to prevent possible crashes and data loss.

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 21) for internal validation can be visualized (automatically) by the single model option from the “Data setup” (see section 4) or (manually, pressing the icon, or clicking the corresponding popup menu item) from the “View and select models” dialog (see section 6). The first box contains four techniques for internal validations that require a high number of iterations (this is why they are optional).

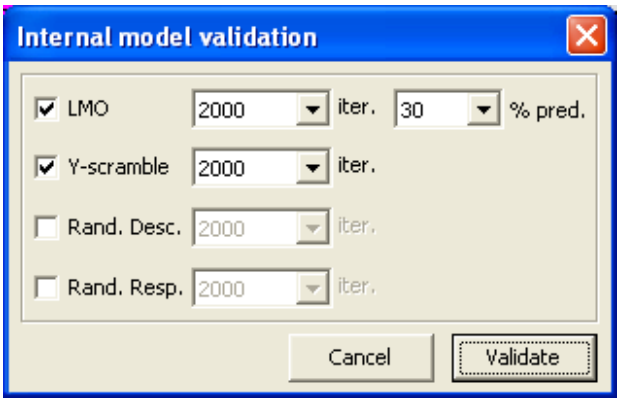


***Figure 21**. Dialog for internal validation*

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^2_{LMO} [Q2LMO] (squared brackets are here used for the criteria names used in QSARINS). Those values must be significantly near to Q^2_{LOO} of the model, and the randomized values ofQ^2_{LMO}, of all the iterations, must be not too dispersed.

The "Y-scramble" list box calculates iteratively a certain number of models (see corresponding "iter." list box) shuffling at random the experimental responses (Eriksson, et al., 2003). The results ($R^2_{Y-\mathrm{SCRAMBLE}}$, RMSE Average *Y-SCRAMBLE* and $Q^2_{Y-\mathrm{SCRAMBLE}}$) will be stored as [R2Yscr] [RMSE AV Yscr] and [Q2Yscr]. R^2 and Q^2_{LOO} of the model under test must be reasonably higher than the scrambled ones, as well as the RMSE of model under test must be reasonably smaller 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 As before, in both cases R^2 and Q^2_{LOO} of the model under test must be much higher than the ones of the random models ($R^2_{\mathrm{RND-RESP}}$ [R2Yrnd], $Q^2_{\mathrm{RND-RESP}}$ [Q2Yrnd] and $R^2_{\mathrm{RND-DESCR}}$ [R2Xrnd], $Q^2_{\mathrm{RND-DESCR}}$ [Q2Xrnd]). The second case will exclude that the original modeling descriptors are just numbers selected by chance, thus without any structural meaning.

External validation

Internal validation is necessary, but not sufficient (Golbraikh, A., Tropsha, 2002, Tropsha et al. 2003, Gramatica 2007, 2009, 2012-2014), thus in QSARINS rigorous check of the external validation can be performed.

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 4, "Data setup") of the following validation criteria:R^2_{EXT} [R2ext](external determination coefficient), Q^2_{F1} (Shi et al., 2001)[Q2-F1], Q^2_{F2} (Schüürmann et. al, 2008)[Q2-F2], Q^2_{F3} (Consonni et al. 2009, 2010) [Q2-F3], *CCC* (Lin, 1989)(Chirico and Gramatica, 2011, 2012) [CCC ext], $\overline{r^2_m}$ (Ojha et al., 2011) [r2m aver.] and Δr^2_m (Ojha et al., 2011)[r2m delta].

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

The icon th visualizes the following dialog (Figure 22) concerning the external validation criteria thresholds.

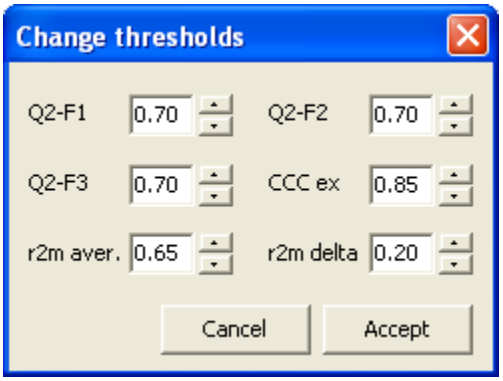


***Figure 22**. Dialog for setting the external validation thresholds*

The user can then set the thresholds of the external validation criteria (Q^2_{F1} [Q2-F1], Q^2_{F2} [Q2-F2], Q^2_{F3} [Q2-F3], *CCC* [CCC ext], $\overline{r^2_m}$ [r2m aver.] and Δr^2_m [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

As suggested in the paper “A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology” (Gramatica and Sangion, 2016), the best models that should be selected are those with the highest values for all the calculated statistical parameters, in agreement.

MC DM: pressing this icon the Multi-Criteria Decision Making score will be calculated (See Section 9)

8. Check of probability of chance correlation in models using variable selection from large pools of descriptors

While performing the variable selection procedure, a well-known issue concerns the descriptor pool size since the larger the number of descriptors, the greater could be the possibility to find a correlation by chance with the response. As a general rule, the smaller the number of molecules used to develop the model, and the larger the size of the descriptor pool, the higher becomes the probability of chance correlation.

It is thus pivotal to calculate the probability of such an event, repeating the descriptors selection procedure many times (i.e. in multiple parallel populations of models) using randomized responses. The distribution of the performances of the best randomized models in each population is then used to calculate the probability of chance correlation for each model (this step is automatically performed by QSARINS).

In more detail, as mentioned above, this procedure generates, for a number of multiple populations of models defined by the user (i.e. iterations), the same number of models as in a regular model development session but using randomized responses. The highest $R^2$and Q^2_{LOO} (the latter is proposed here for coherence with the statistics used in QSARINS) calculated for each iteration are stored and compared with $R^2$and Q^2_{LOO} values of the not randomized model. The difference between the $R^2$and Q^2_{LOO} values for randomized and not randomized models is quantitatively related to the probability of chance correlation.

We want to emphasize that the aforementioned procedure should not be confused with the randomization procedure (Y-scramble) explained in Sections 7 (“Model validation”) and 10 (“Analysis of single models”), where the randomization of the responses is applied only to **a single model, *after* the variable selection procedure**. In this chapter, instead, the focus is on the comparison between the performances of the best models **in n populations generated using the same variable selection procedure, where n-1 populations use randomized responses**. When a statistically relevant number of $R^2$and Q^2_{LOO} values, calculated for the best randomized and not randomized models, are similar, the quality of the not randomized model is dubious.

Here it follows an explanatory example, as step by step guide, where an apparently good model is tested for the possibility of chance correlation. The model used here as an example is

calculated using a small number of molecules (20) and variables selected from a pool of 650 descriptors. Performances of the model are very good since Q^2_{LOO} = 0.91, and R^2=0.94 as well as the coefficients and the intercept are significant (p values <= 0.0002). Furthermore, the classical Y-scrambling procedure, as explained in Section 7 and 10, results in scrambled R^2 = 0.21 and Q^2_{LOO} = -0.45, which indicate the absence of chance correlation.

Let us now apply the additional procedure of randomization described in this chapter.

- The first step is to set up *the same* variable selection procedure which leaded to the not randomized model, as shown in the Figure 23.
- The second step consists of choosing the randomization procedure, in this example "Permuted responses – original variables". Note: randomization techniques are many (Rücker et al., 2007) but, to avoid confusion, in QSARINS only the ones reasonable for QSAR datasets typology and modelling (Katritzky et al., 2008) are available.
- The third step consists of choosing the number of iterations (25 in this example). Note: the whole randomization procedure is *very* time consuming, therefore in this example the number of iterations is relatively low, but anyway reasonable as reported in Rücker et al., 2007.

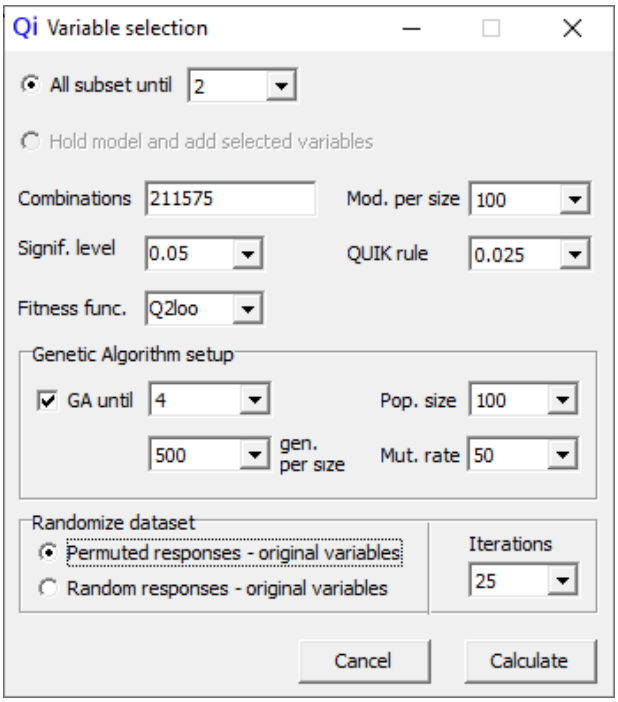
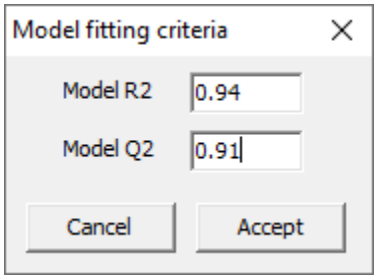


***Figure 23**. Variable selection for the randomization of the dataset*

The fourth step consists of launching the randomization procedure which should generate randomized models with a comparable quality respect to the not randomized model. Therefore, the significance level of the model's coefficients and intercepts chosen ("Signif. level" in the variable selection dialog) must be consistent.

- The fifth and last step compares the statistics among the best models in multiple populations and the not randomized model. Once calculations have been completed, the icon  is activated. Pressing on this icon, or alternatively from the menu Analysis->View statistics of models from randomized datasets, the user will be asked to enter the R^2 and Q^2_{LOO} values of the model under scrutiny, as shown below.



After entering the values, a graph showing the statistics of the models generated by randomized datasets is shown as in Figure 24

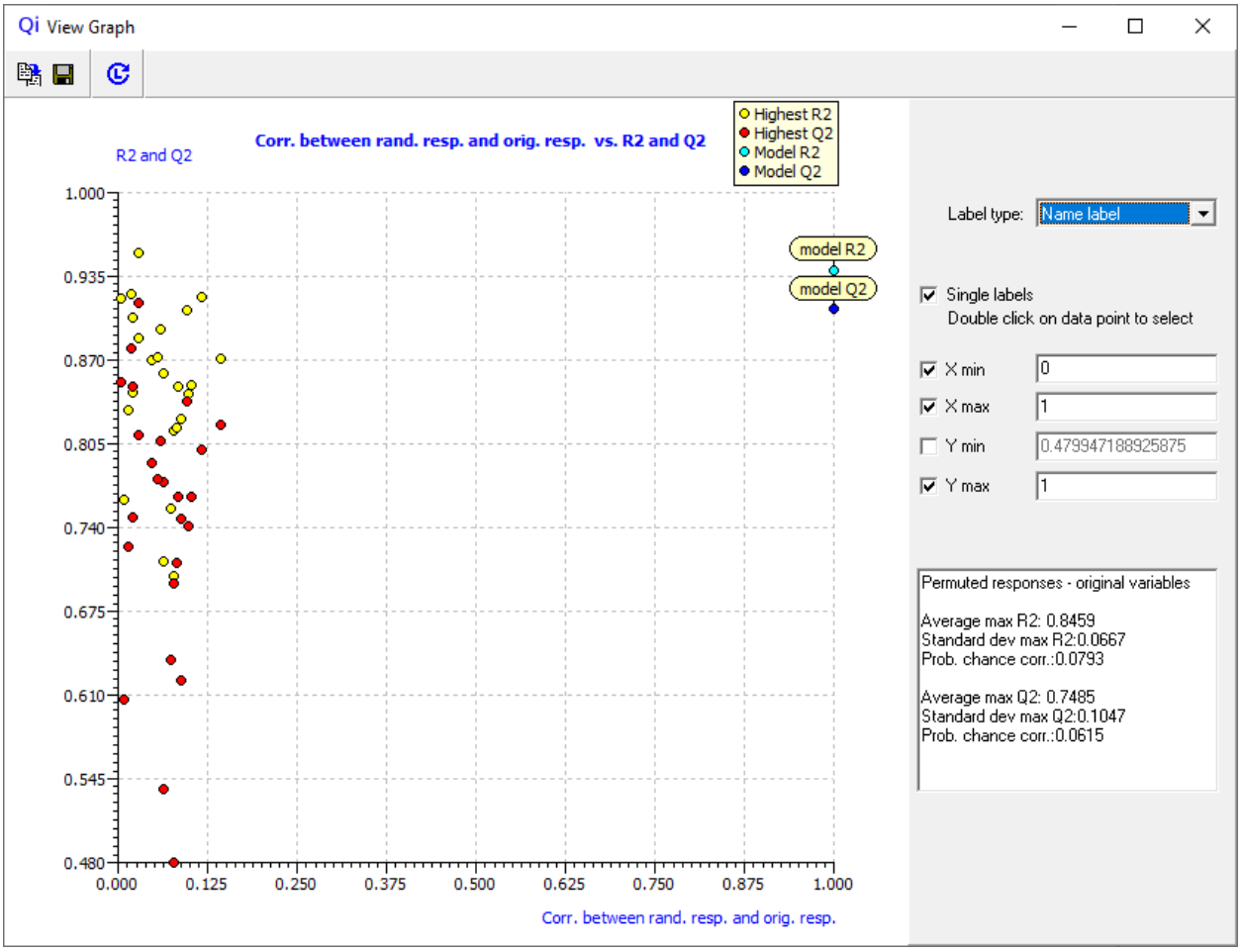


***Figure 24**. Example of graphs and statistics of maximum Q^2_{LOO} and R^2 of the models calculated from randomized dataset.*

On the ordinate axis the highest Q^2_{LOO} and R^2 values of the models generated by randomization, red and blue dots, are plotted while on the abscissa axis the corresponding correlation between the experimental responses and the randomized ones is reported. The latter is an important information, since it is essential that the randomized responses are highly dissimilar from the original (i.e. the correlation between randomized and not randomized responses must be low) to guarantee that the randomization process is effective and does not closely reproduce the sequence of the original responses.

Finally, the probability of chance correlation is reported as “Prob. chance corr.” in the text box in the lower right of the graph. In this example the p-values are: average maximum Q^2_{LOO} = 0.06 and average maximum R^2 = 0.08. If the acceptable level of probability of chance

correlation is less or equal to 0.05 (a commonly used threshold value), as in this example, it is not possible to exclude chance correlation at the selected p-value.

Therefore in this example the quality of the not-randomized model should be considered as dubious, accordingly to the additional randomization procedure.

In this case, if the model was developed using correctly filtered descriptors (avoiding collinearity) and deeply validated, a careful analysis and possibly interpretation of the selected descriptors is necessary to guarantee the reliability of the model, even if not substantially different from a possible chance correlation.

9. Model selection by MCDM

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^2_{adj} and CCC_{TR}, while minimizing R^2-R^2_{adj}), cross validation (maximizing Q^2_{LOO},Q^2_{LMO} and CCC_{cv}, while minimizing $R^2_{Y-\text{SCRAMBLE}}$) and external validation (maximizing Q^2_{F1}, Q^2_{F2}, Q^2_{F3} 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 25):

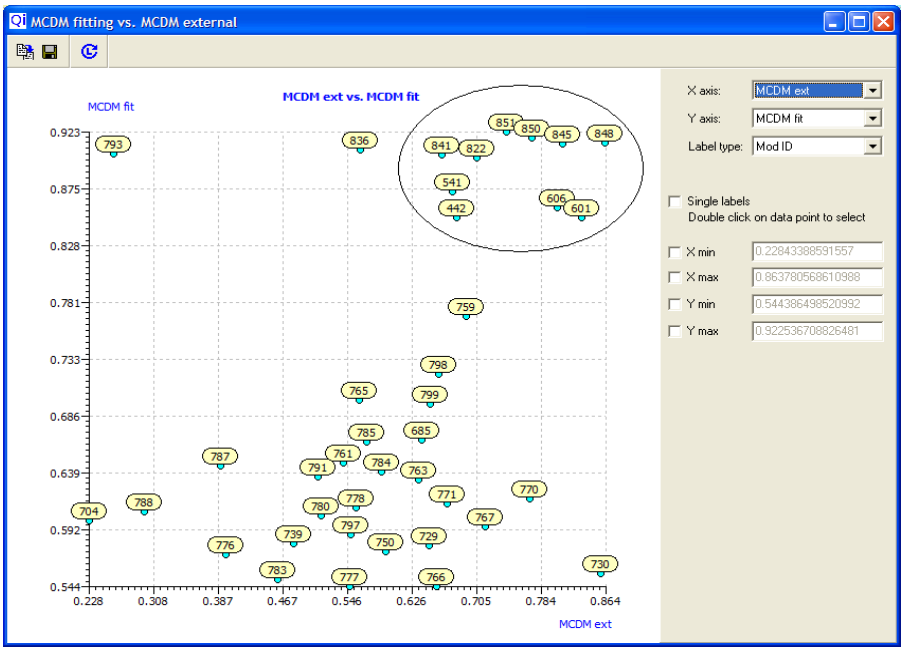


***Figure 25**. Plot of MCDM ext vs. MCDM fit*

In Figure 25, the model 793 has very good fitting performances, but it is not externally predictive, while for instance model 730 has very good performances in external predictivity while it is not satisfactory for fitting. Models in the upper-right zone of the Figure 25 (circled) show the better compromise between fitting and predictivity, and should be selected as the best models.

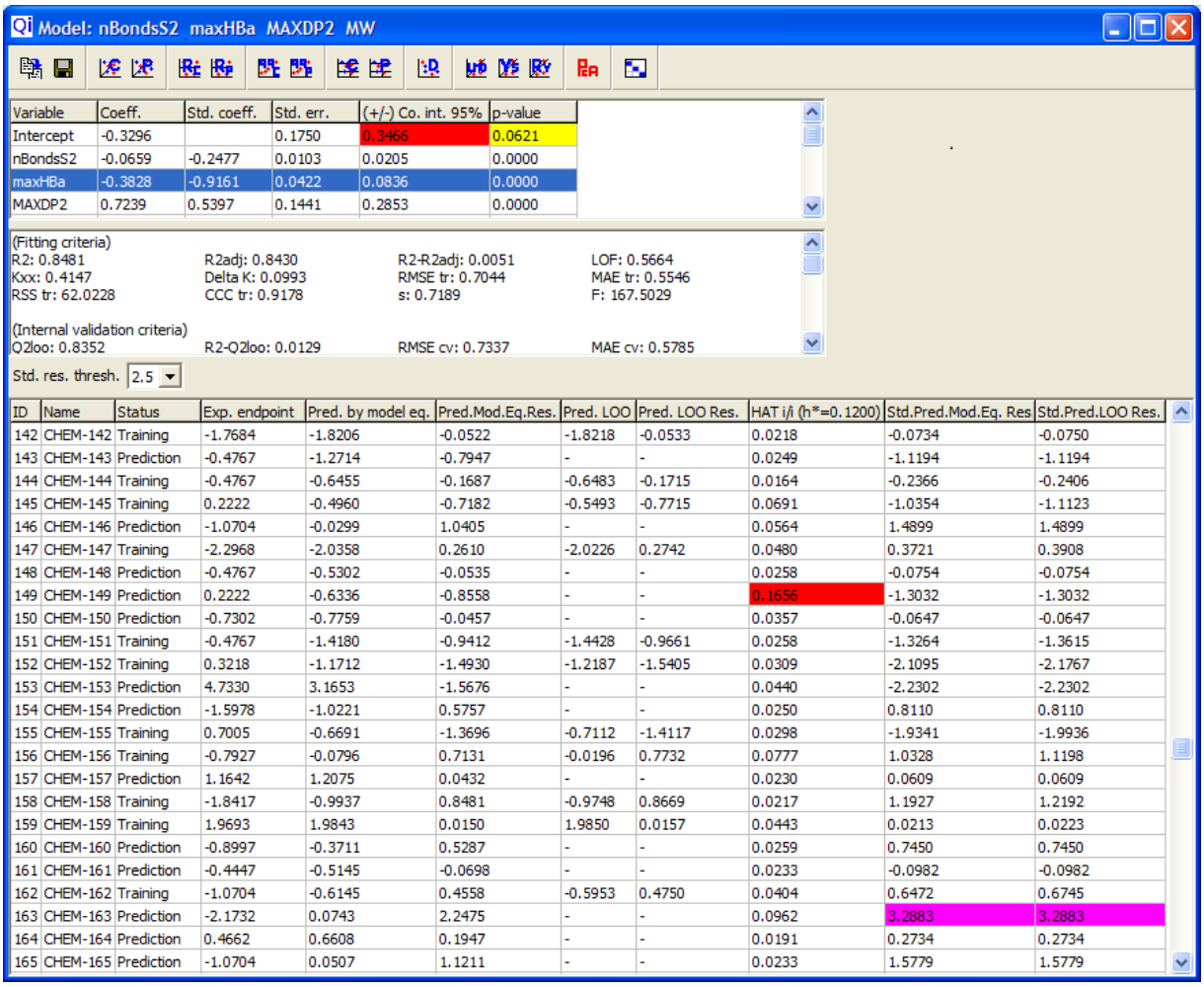
10. Analysis of single models

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 4) 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 6, “View and select models”

Here it follows an example of the single model dialog box (Figure 26):



Variable	Coeff.	Std. coeff.	Std. err.	(+/-) Co. int. 95%	p-value
Intercept	-0.3296		0.1750	0.3466	0.0621
nBondsS2	-0.0659	-0.2477	0.0103	0.0205	0.0000
maxHBa	-0.3828	-0.9161	0.0422	0.0836	0.0000
MAXDP2	0.7239	0.5397	0.1441	0.2853	0.0000

(Fitting criteria)
R2: 0.8481 R2adj: 0.8430 R2-R2adj: 0.0051 LOF: 0.5664
Kxx: 0.4147 Delta K: 0.0993 RMSE tr: 0.7044 MAE tr: 0.5546
RSS tr: 62.0228 CCC tr: 0.9178 s: 0.7189 F: 167.5029

(Internal validation criteria)
Q2loo: 0.8352 R2-Q2loo: 0.0129 RMSE cv: 0.7337 MAE cv: 0.5785

Std. res. thresh. 2.5

ID	Name	Status	Exp. endpoint	Pred. by model eq.	Pred.Mod.Eq.Res.	Pred. LOO	Pred. LOO Res.	HAT i/i (h*=0.1200)	Std.Pred.Mod.Eq. Res	Std.Pred.LOO Res.
142	CHEM-142	Training	-1.7684	-1.8206	-0.0522	-1.8218	-0.0533	0.0218	-0.0734	-0.0750
143	CHEM-143	Prediction	-0.4767	-1.2714	-0.7947	-	-	0.0249	-1.1194	-1.1194
144	CHEM-144	Training	-0.4767	-0.6455	-0.1687	-0.6483	-0.1715	0.0164	-0.2366	-0.2406
145	CHEM-145	Training	0.2222	-0.4960	-0.7182	-0.5493	-0.7715	0.0691	-1.0354	-1.1123
146	CHEM-146	Prediction	-1.0704	-0.0299	1.0405	-	-	0.0564	1.4899	1.4899
147	CHEM-147	Training	-2.2968	-2.0358	0.2610	-2.0226	0.2742	0.0480	0.3721	0.3908
148	CHEM-148	Prediction	-0.4767	-0.5302	-0.0535	-	-	0.0258	-0.0754	-0.0754
149	CHEM-149	Prediction	0.2222	-0.6336	-0.8558	-	-	0.1656	-1.3032	-1.3032
150	CHEM-150	Prediction	-0.7302	-0.7759	-0.0457	-	-	0.0357	-0.0647	-0.0647
151	CHEM-151	Training	-0.4767	-1.4180	-0.9412	-1.4428	-0.9661	0.0258	-1.3264	-1.3615
152	CHEM-152	Training	0.3218	-1.1712	-1.4930	-1.2187	-1.5405	0.0309	-2.1095	-2.1767
153	CHEM-153	Prediction	4.7330	3.1653	-1.5676	-	-	0.0440	-2.2302	-2.2302
154	CHEM-154	Prediction	-1.5978	-1.0221	0.5757	-	-	0.0250	0.8110	0.8110
155	CHEM-155	Training	0.7005	-0.6691	-1.3696	-0.7112	-1.4117	0.0298	-1.9341	-1.9936
156	CHEM-156	Training	-0.7927	-0.0796	0.7131	-0.0196	0.7732	0.0777	1.0328	1.1198
157	CHEM-157	Prediction	1.1642	1.2075	0.0432	-	-	0.0230	0.0609	0.0609
158	CHEM-158	Training	-1.8417	-0.9937	0.8481	-0.9748	0.8669	0.0217	1.1927	1.2192
159	CHEM-159	Training	1.9693	1.9843	0.0150	1.9850	0.0157	0.0443	0.0213	0.0223
160	CHEM-160	Prediction	-0.8997	-0.3711	0.5287	-	-	0.0259	0.7450	0.7450
161	CHEM-161	Prediction	-0.4447	-0.5145	-0.0698	-	-	0.0233	-0.0982	-0.0982
162	CHEM-162	Training	-1.0704	-0.6145	0.4558	-0.5953	0.4750	0.0404	0.6472	0.6745
163	CHEM-163	Prediction	-2.1732	0.0743	2.2475	-	-	0.0962	3.2883	3.2883
164	CHEM-164	Prediction	0.4662	0.6608	0.1947	-	-	0.0191	0.2734	0.2734
165	CHEM-165	Prediction	-1.0704	0.0507	1.1211	-	-	0.0233	1.5779	1.5779

***Figure 26**. Single model dialog box*

The first two icons on the first row in Figure 26 allow to copy ( ) on the clipboard and save models data ( ), the subsequent icons ( ) display the model performances in graphical form, the icon displays the PCA (score plot and loadings plot) of the model descriptors (useful for the check of the structural domain of the selected model), while the last icon ( ) displays the correlation matrix of the modeling descriptors. If additional information (QMRF and/or the related .sdf file with modeling descriptors) about the single model file (.smd, see below) is available, a and icons will appear to the right.

It is here recalled that in addition to the icons the user can use the popup menu to select the tools/options.

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 12.2, “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 27):

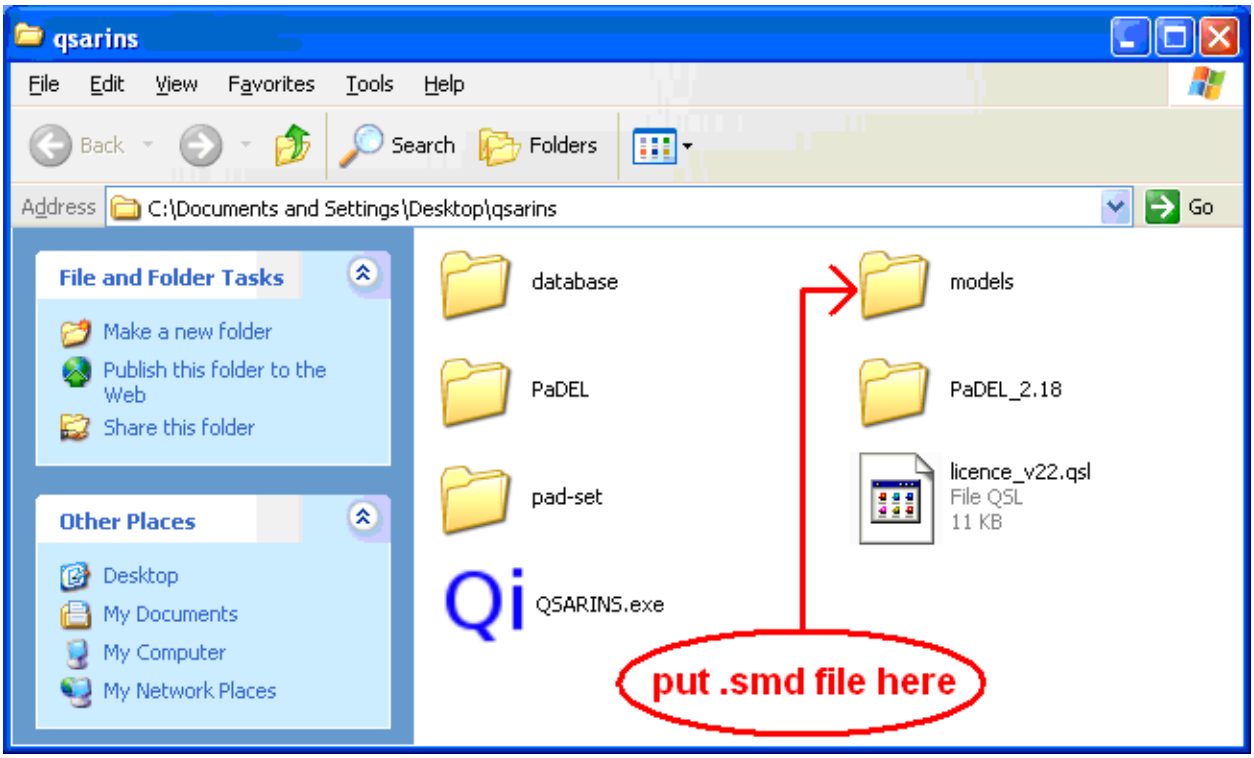


***Figure 27**. Single model file (.smd) folder location*

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 26 displays the model equation coefficients, the standardized coefficients, standard error, the confidence intervals (Co. int. 95%) of the regression coefficients

and p-value. It 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 26). 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 is 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^2_{adj} [R2adj], $R^2 - R^2_{\text{adj}}$ [R2-R2adj], *LOF* (Friedman lack of fit criteria) [LOF], K_x [Kxx] (inter-correlation among descriptors), ΔK (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^2_{LOO} [Q2loo], $R^2 - Q^2_{\text{LOO}}$ [R2-Q2loo], RMSE [RMSE cv], MAE [MAE cv], PRESS [PRESS cv], CCC [CCC cv], Q^2_{LMO} [Q2LMO], $R^2_{Y-\text{SCRAMBLE}}$ [R2Yscr], RMSE Average *Y-SCRAMBLE* [RMSE AV Yscr], $Q^2_{Y-\text{SCRAMBLE}}$ [Q2Yscr], $R^2_{\text{RND}-\text{DESCR}}$ [R2Xrnd], $Q^2_{\text{RND}-\text{DESCR}}$ [Q2Xrnd], $R^2_{\text{RND}-\text{RESP}}$ [R2Yrnd], $Q^2_{\text{RND}-\text{RESP}}$ [Q2Yrnd].

External validation criteria: RMSE [RMSE ext], MAE [MAE ext], PRESS [PRESS ext], R^2_{EXT} [R2ext], Q^2_{F1} [Q2-F1], Q^2_{F2} [Q2-F2], and Q^2_{F3} [Q2-F3], CCC [CCC ext], $\overline{r^2_m}$ [r2m aver.] and Δr^2_m [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^2_m statistics (Ojha,et al.2011), 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

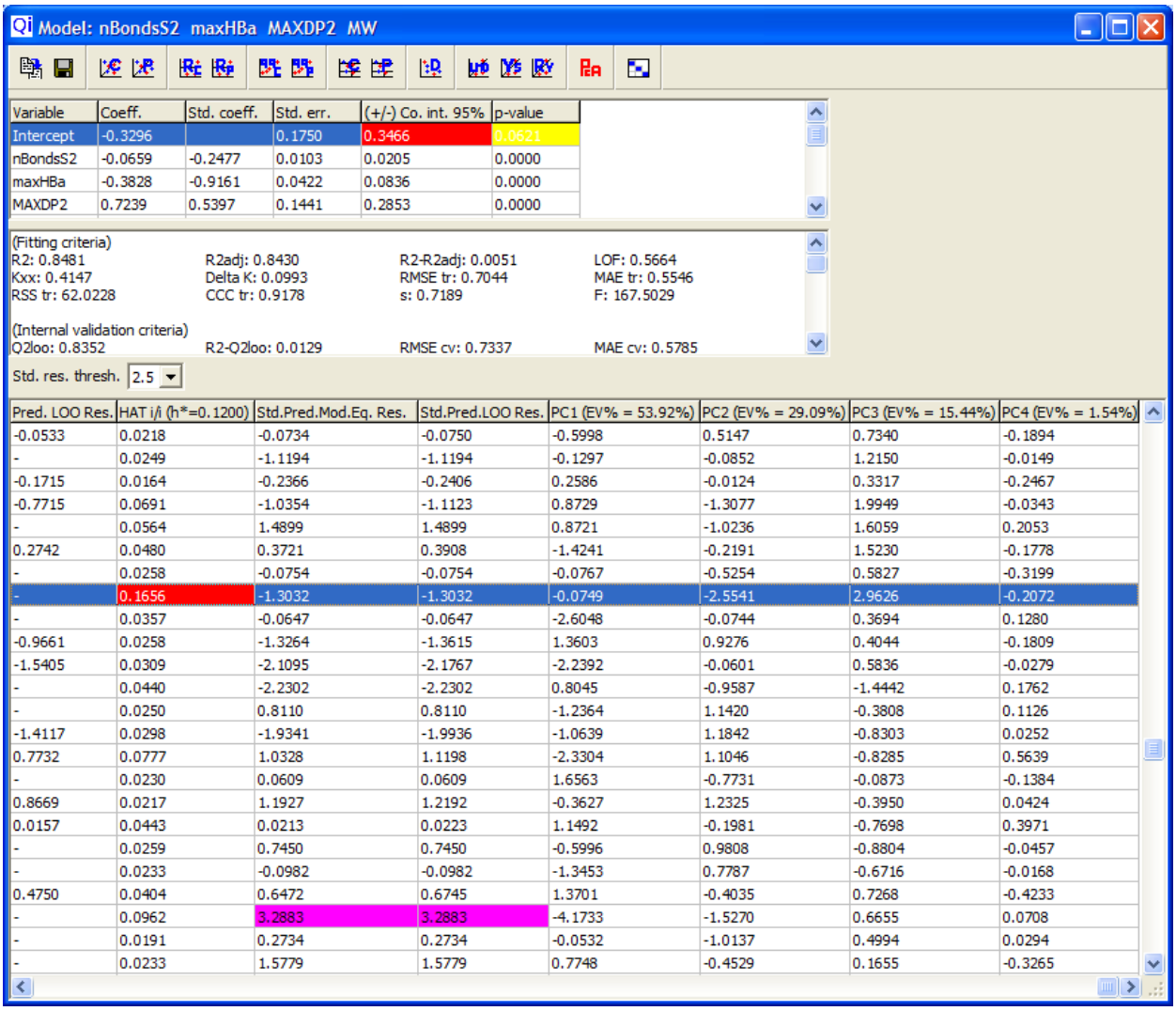
Pred(x) vs. Exp(y): R2: 0.7928 R2o: 0.7912 k: 0.9991 Clos: 0.0020 r2m: 0.7610

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, except 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 26 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 28).



Variable	Coeff.	Std. coeff.	Std. err.	(+/-) Co. int. 95%	p-value
Intercept	-0.3296		0.1750	0.3466	0.0621
nBondsS2	-0.0659	-0.2477	0.0103	0.0205	0.0000
maxHBa	-0.3828	-0.9161	0.0422	0.0836	0.0000
MAXDP2	0.7239	0.5397	0.1441	0.2853	0.0000

(Fitting criteria)
R2: 0.8481 R2adj: 0.8430 R2-R2adj: 0.0051 LOF: 0.5664
Kxx: 0.4147 Delta K: 0.0993 RMSE tr: 0.7044 MAE tr: 0.5546
RSS tr: 62.0228 CCC tr: 0.9178 s: 0.7189 F: 167.5029

(Internal validation criteria)
Q2loo: 0.8352 R2-Q2loo: 0.0129 RMSE cv: 0.7337 MAE cv: 0.5785

Std. res. thresh. 2.5

Pred. LOO Res.	HAT i/i (h*=0.1200)	Std.Pred.Mod.Eq. Res.	Std.Pred.LOO Res.	PC1 (EV% = 53.92%)	PC2 (EV% = 29.09%)	PC3 (EV% = 15.44%)	PC4 (EV% = 1.54%)
-0.0533	0.0218	-0.0734	-0.0750	-0.5998	0.5147	0.7340	-0.1894
-	0.0249	-1.1194	-1.1194	-0.1297	-0.0852	1.2150	-0.0149
-0.1715	0.0164	-0.2366	-0.2406	0.2586	-0.0124	0.3317	-0.2467
-0.7715	0.0691	-1.0354	-1.1123	0.8729	-1.3077	1.9949	-0.0343
-	0.0564	1.4899	1.4899	0.8721	-1.0236	1.6059	0.2053
0.2742	0.0480	0.3721	0.3908	-1.4241	-0.2191	1.5230	-0.1778
-	0.0258	-0.0754	-0.0754	-0.0767	-0.5254	0.5827	-0.3199
-	0.1656	-1.3032	-1.3032	-0.0749	-2.5541	2.9626	-0.2072
-	0.0357	-0.0647	-0.0647	-2.6048	-0.0744	0.3694	0.1280
-0.9661	0.0258	-1.3264	-1.3615	1.3603	0.9276	0.4044	-0.1809
-1.5405	0.0309	-2.1095	-2.1767	-2.2392	-0.0601	0.5836	-0.0279
-	0.0440	-2.2302	-2.2302	0.8045	-0.9587	-1.4442	0.1762
-	0.0250	0.8110	0.8110	-1.2364	1.1420	-0.3808	0.1126
-1.4117	0.0298	-1.9341	-1.9936	-1.0639	1.1842	-0.8303	0.0252
0.7732	0.0777	1.0328	1.1198	-2.3304	1.1046	-0.8285	0.5639
-	0.0230	0.0609	0.0609	1.6563	-0.7731	-0.0873	-0.1384
0.8669	0.0217	1.1927	1.2192	-0.3627	1.2325	-0.3950	0.0424
0.0157	0.0443	0.0213	0.0223	1.1492	-0.1981	-0.7698	0.3971
-	0.0259	0.7450	0.7450	-0.5996	0.9808	-0.8804	-0.0457
-	0.0233	-0.0982	-0.0982	-1.3453	0.7787	-0.6716	-0.0168
0.4750	0.0404	0.6472	0.6745	1.3701	-0.4035	0.7268	-0.4233
-	0.0962	3.2883	3.2883	-4.1733	-1.5270	0.6655	0.0708
-	0.0191	0.2734	0.2734	-0.0532	-1.0137	0.4994	0.0294
-	0.0233	1.5779	1.5779	0.7748	-0.4529	0.1655	-0.3265

Figure 28 *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 29)

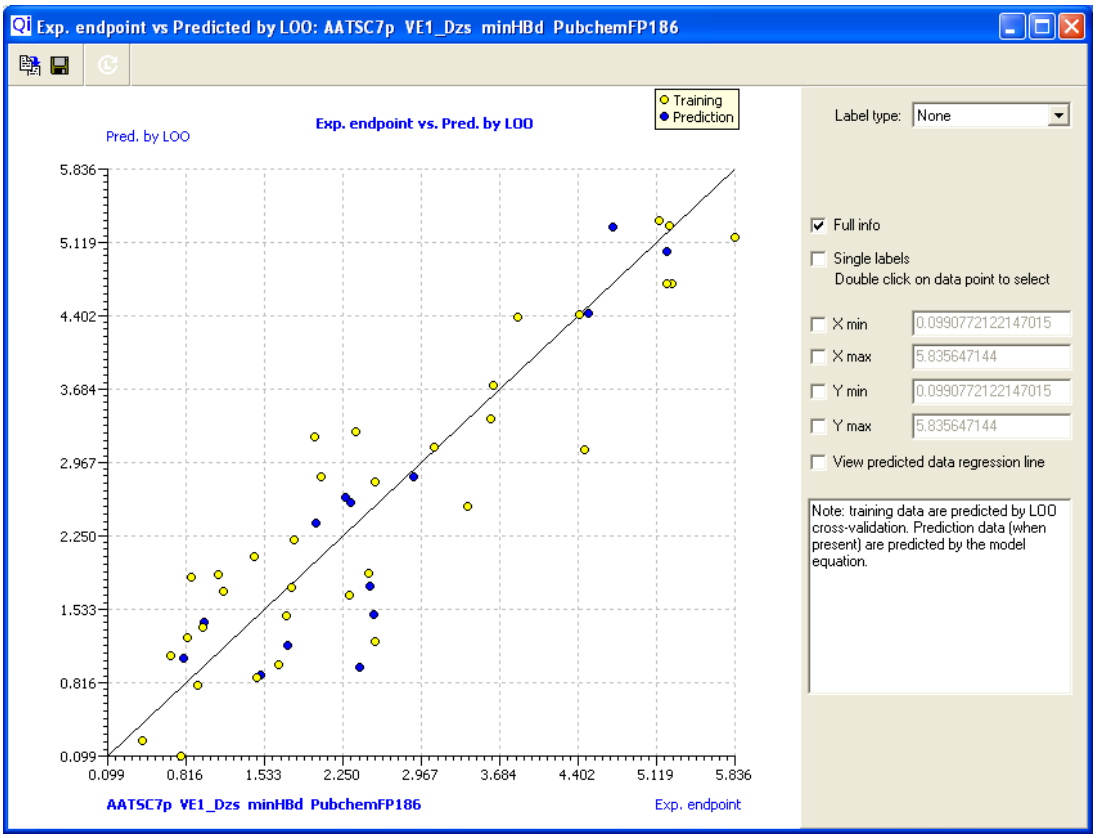


Figure 29 *Scatter plot of experimental vs. predicted data by LOO (press*  *icon)*

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.

	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 30)

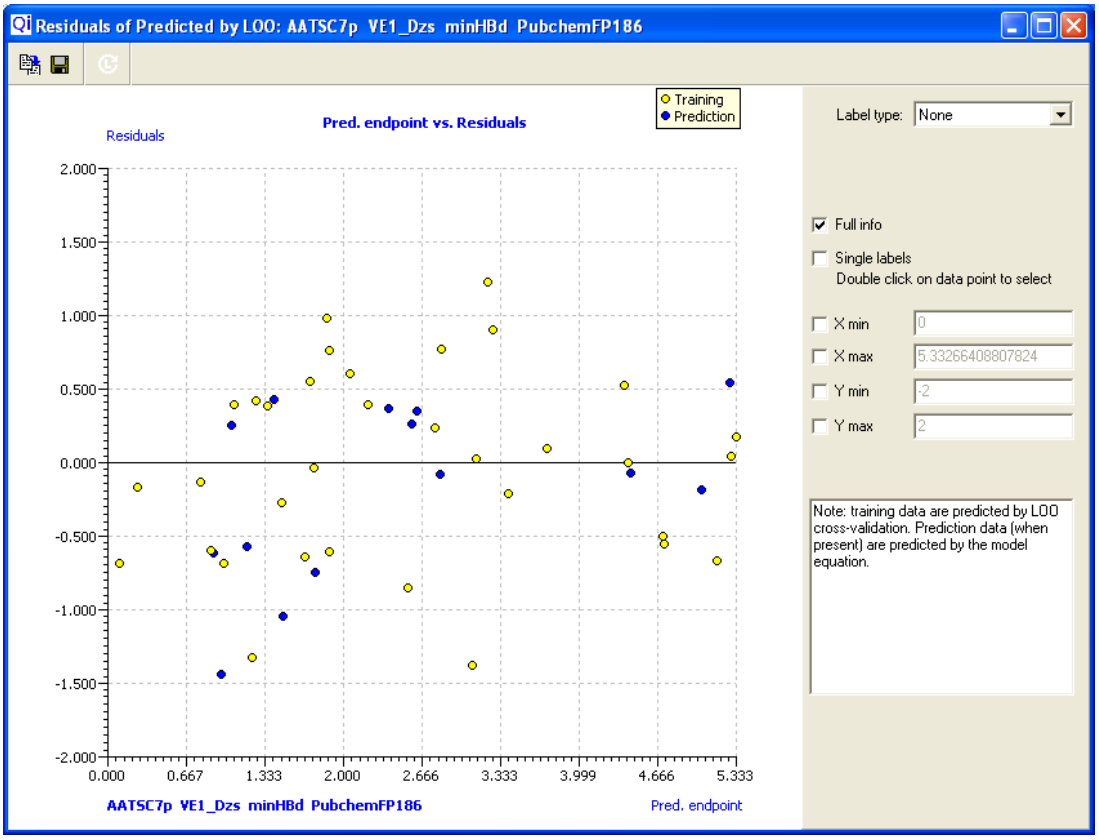


***Figure 30**. Scatter plot of the residuals of the predicted data by LOO (press*  *icon)*

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.

	q-q plot of experimental endpoint vs. residuals from the predictions by model equation
	q-q plot of experimental endpoint vs. residuals from the LOO predictions (Figure 31)

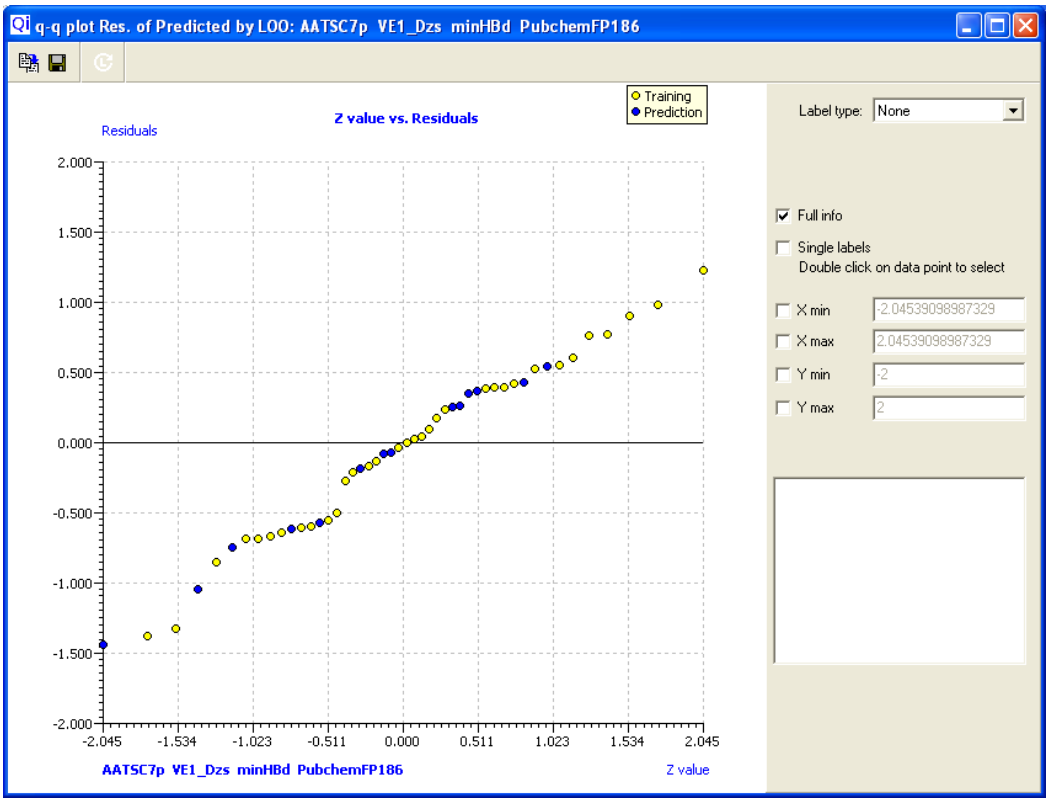


***Figure 31**. q-q plot of the residuals of the predicted data by the model equation (press*  *icon)*

The reported values are the values of the theoretical quantiles (Z values) on the abscissa and the values of the residuals of the predictions on the ordinate. In Figure 31 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 32)

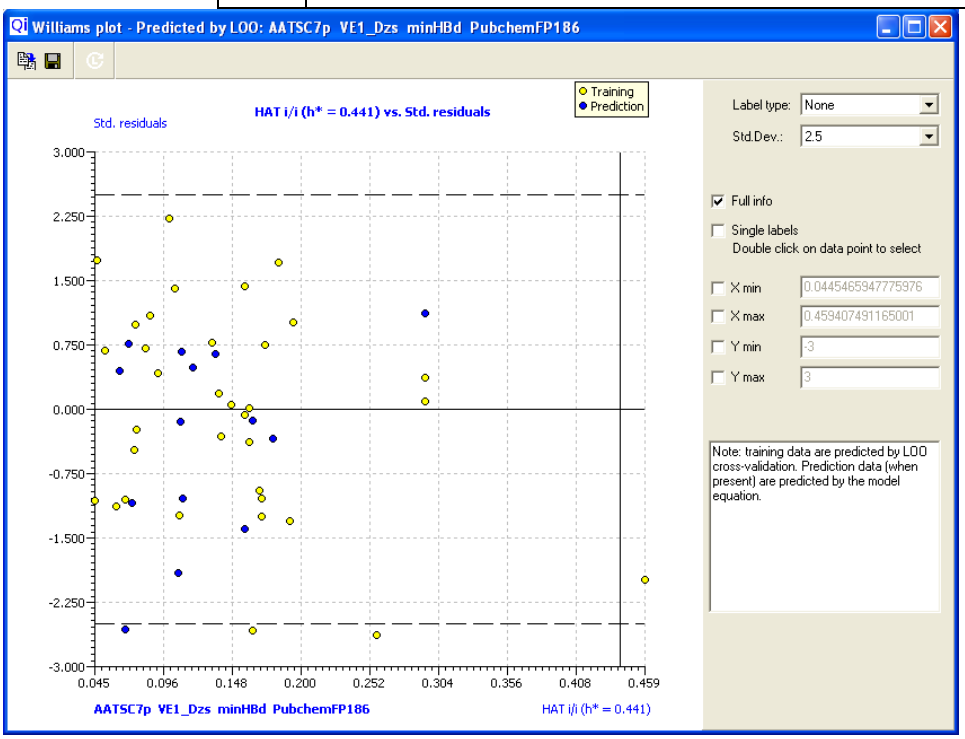


Figure 32 *Williams plot using data predicted by LOO (press*  *icon)*

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.

In the Williams plot, 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 33)

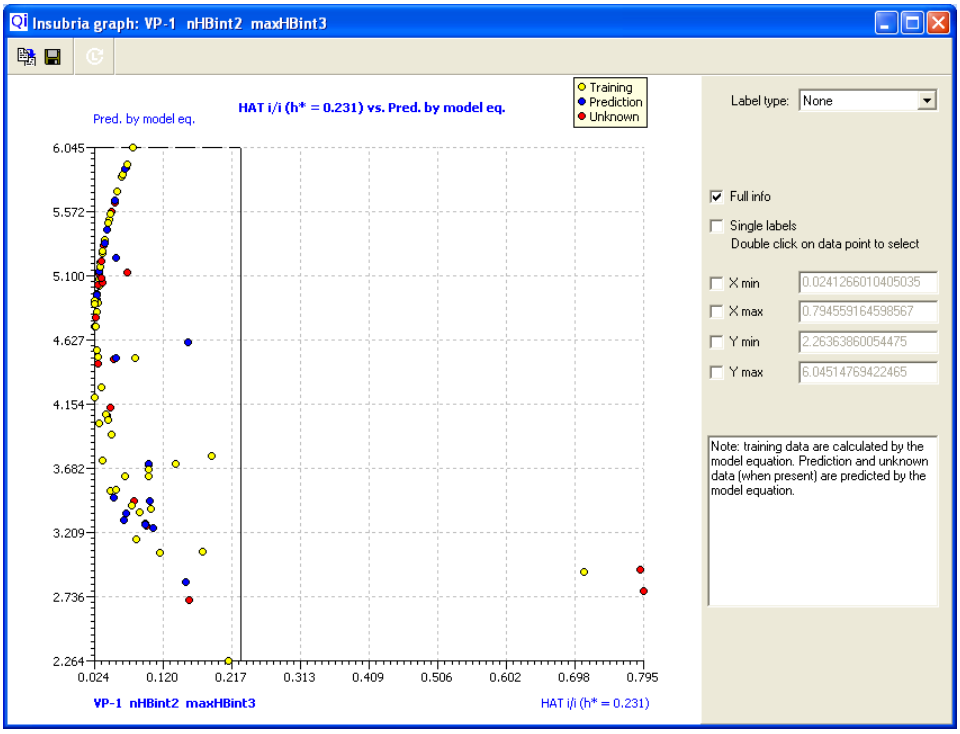


Figure 33 *Example of Insubria graph (press*  *icon)*

The Insubria graph must be applied to verify the applicability domain of a model to chemicals without experimental data. In this graph, 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 of the model. 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 34)
	Y-Scramble plot (Kxy vs. R^2 and Q^2) (Figure 35)
	Random descriptors plot (Kxy vs. R^2 and Q^2)
	Random responses plot (Kxy vs. R^2 and Q^2)

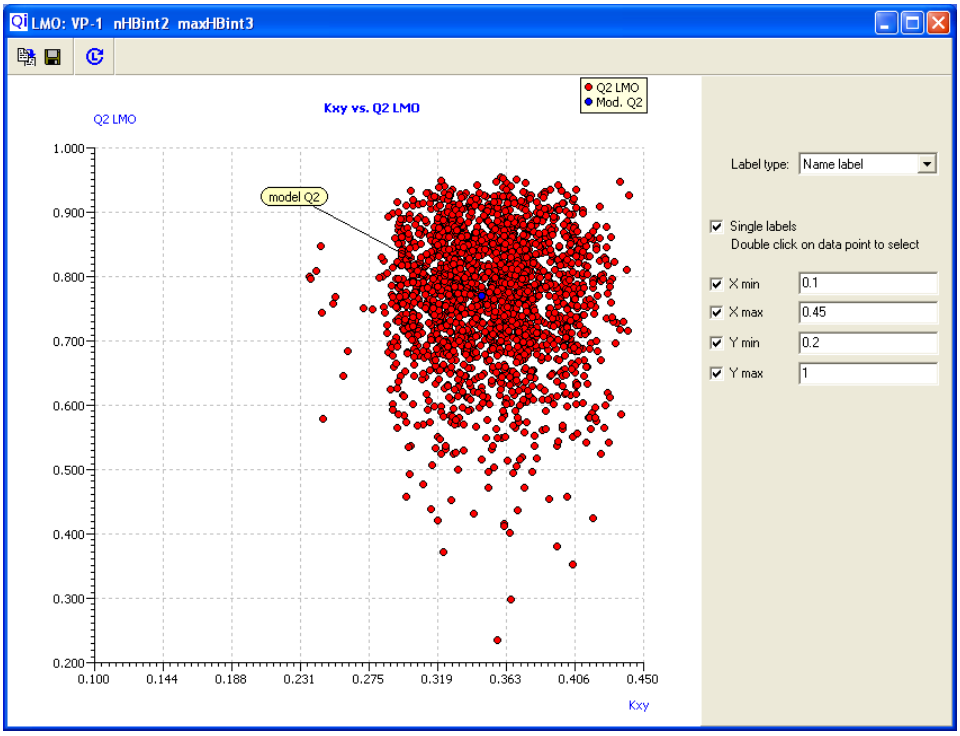
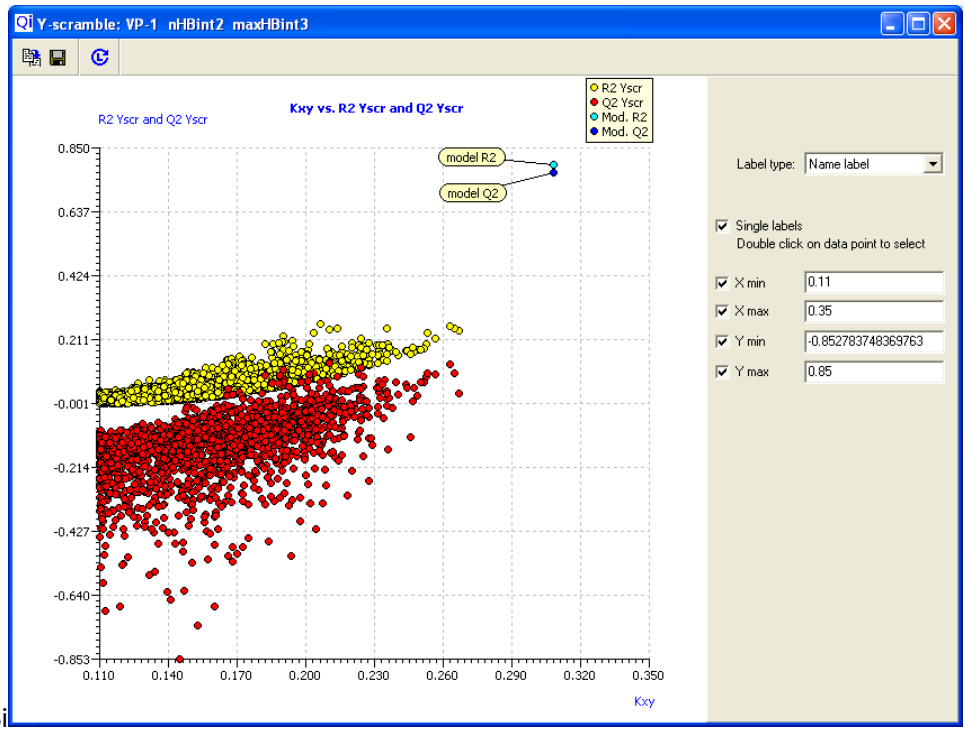


***Figure 34**. Scatter plot of LMO models compared to the QSAR model (press*  *icon)*

On the abscissa the correlations among the block of the descriptors and the experimental data (Kxy) 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.



Si

***Figure 35**. Scatter plot of Y-scrambled models compared to the original 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.
This step allows excluding the possibility that the descriptors in the model are not correlated by chance with the response. This is an internal validation of the model and must not be confused with the randomization process during variable selection described in Section 8. “Check of probability of chance correlation in models using variable selection from large pools of descriptors”

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.

11. Combined modeling

In QSARINS it is possible to apply the combined modeling (named consensus in 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 ). It is here recommended to perform a combined modeling possibly only selecting the models of the same size of variables, in order to avoid wrong calculations in the standardised residuals.

In the “View and select models” dialog (Section 6),  calculates the PCA based on residuals of the selected models, as in the following example (Figure 36):

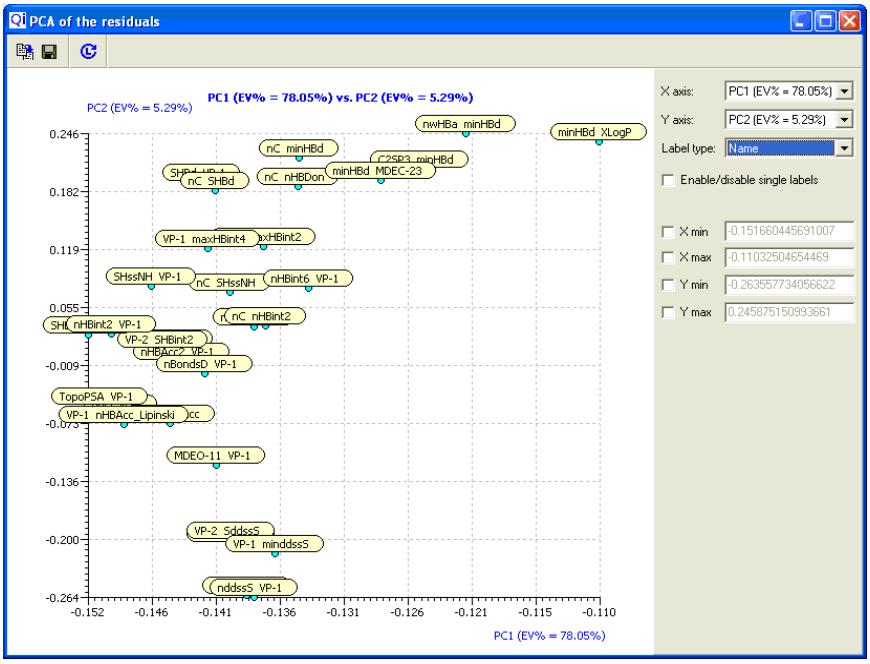


Figure 36 *PCA based on residuals of the selected models*

 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 (Figure 37). (Note: the models must have the same number of descriptors)

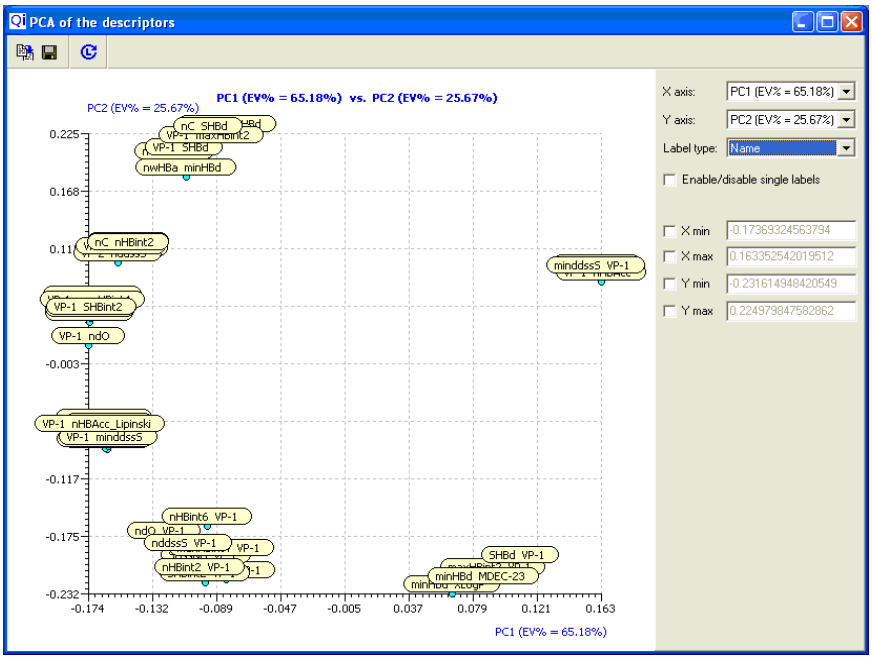


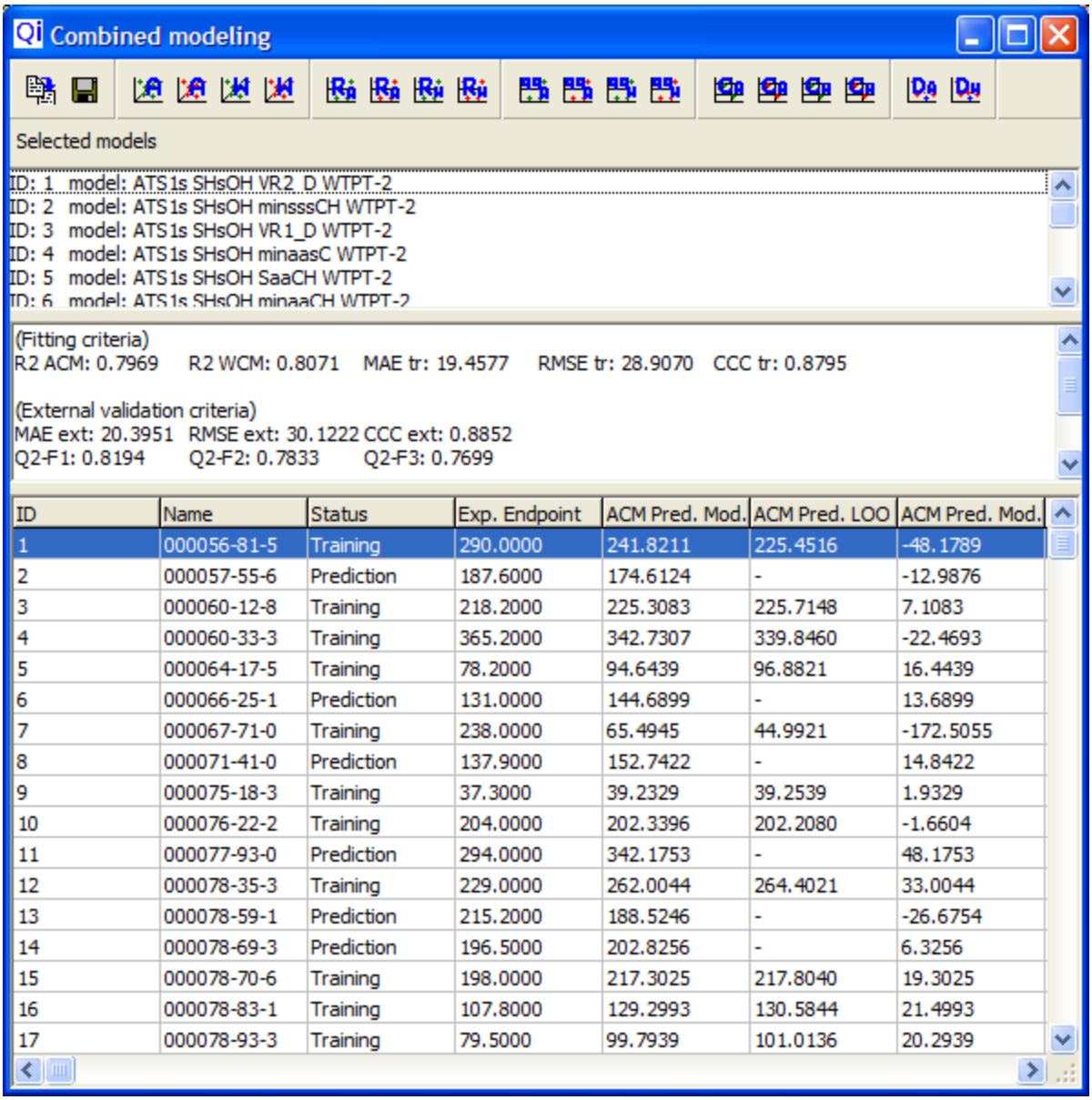
***Figure 37**. PCA based on the descriptors of the selected models*

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 combined 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 combined modeling, clicking on the icon Co or selecting the corresponding item “Combined modeling” from the pop-up menu of the “View and selects models” (Section 6) window.

When the user clicks the Co icon, all the selected models (status “S”) will be used to calculate the combined model. Combined 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 Combined model, QSARINS provides the following information (Figure 38):



ID	Name	Status	Exp. Endpoint	ACM Pred. Mod.	ACM Pred. LOO	ACM Pred. Mod.
1	000056-81-5	Training	290.0000	241.8211	225.4516	-48.1789
2	000057-55-6	Prediction	187.6000	174.6124	-	-12.9876
3	000060-12-8	Training	218.2000	225.3083	225.7148	7.1083
4	000060-33-3	Training	365.2000	342.7307	339.8460	-22.4693
5	000064-17-5	Training	78.2000	94.6439	96.8821	16.4439
6	000066-25-1	Prediction	131.0000	144.6899	-	13.6899
7	000067-71-0	Training	238.0000	65.4945	44.9921	-172.5055
8	000071-41-0	Prediction	137.9000	152.7422	-	14.8422
9	000075-18-3	Training	37.3000	39.2329	39.2539	1.9329
10	000076-22-2	Training	204.0000	202.3396	202.2080	-1.6604
11	000077-93-0	Prediction	294.0000	342.1753	-	48.1753
12	000078-35-3	Training	229.0000	262.0044	264.4021	33.0044
13	000078-59-1	Prediction	215.2000	188.5246	-	-26.6754
14	000078-69-3	Prediction	196.5000	202.8256	-	6.3256
15	000078-70-6	Training	198.0000	217.3025	217.8040	19.3025
16	000078-83-1	Training	107.8000	129.2993	130.5844	21.4993
17	000078-93-3	Training	79.5000	99.7939	101.0136	20.2939

***Figure 38**. Combined model dialog box*

The first box in Figure 38 reports the list of selected models used for Combined (Model ID and molecular descriptors). The second box reports the statistical parameters for the Combined, described below (the names in squared brackets are the ones reported by QSARINS).

Fitting criteria: R^2_{ACM} [R2 ACM], R^2_{WCM} [R2 WCM], MAE tr, RMSE tr, CCC tr.

External validation criteria: MAE ext, RMSE ext, CCC ext, Q^2_{F1} [Q2-F1], Q^2_{F2} [Q2-F2], Q^2_{F3} [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 combined prediction ("ACM Pred.", both predicted by the model eq. and by LOO), weighted combined prediction ("WCM Pred." both predicted by the model eq. and by LOO, calculated averaging the predictions of the molecules separately from every model), residuals of the average combined predictions (ACM Pred. Res., both predicted by the model eq. and by LOO), residuals of the weighted combined predictions (WCM Pred. Res., both predicted by the model eq. and by LOO) 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"), average leverage value ("Av.HAT i/i"), standardised residuals of the average combined predictions (Std. ACM Pred. Res., both predicted by the model eq. and by LOO) and standardised residuals of the weighted combined predictions (Std. WCM Pred. Res., both predicted by the model eq. and by LOO).

The icons under the dialog title in Figure 38 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 combined model performances, as exemplified in the table:

	Scatter plot of experimental endpoint vs. ACM predicted by model eq.
	Scatter plot of experimental endpoint vs. ACM predicted by LOO
	Scatter plot of experimental endpoint vs. WCM predicted by model eq.
	Scatter plot of experimental endpoint vs. WCM predicted by LOO
	Residual plot of experimental endpoint vs. residuals from the ACM predictions by model eq.
	Residual plot of experimental endpoint vs. residuals from the ACM predictions by LOO
	Residual plot of experimental endpoint vs. residuals from the WCM predictions by model eq.
	Residual plot of experimental endpoint vs. residuals from the WCM predictions by LOO

 q-q plot of experimental endpoint vs. residuals from the ACM predictions by model eq.

 q-q plot of experimental endpoint vs. residuals from the ACM predictions by LOO

 q-q plot of experimental endpoint vs. residuals from the WCM predictions by model eq.

 q-q plot of experimental endpoint vs. residuals from the WCM predictions by LOO

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

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

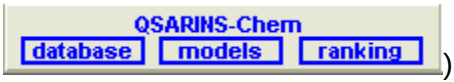
 Williams plot of average hat diagonal elements vs. WCM. predicted by model eq. residuals

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

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

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

12. 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 44 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.

12.1 Database

Pressing the database icon (or, alternatively, selecting Tools->Query database menu item) in the main screen, the following dialog will appear (Figure 39), where it is possible to query the compounds included in the database:

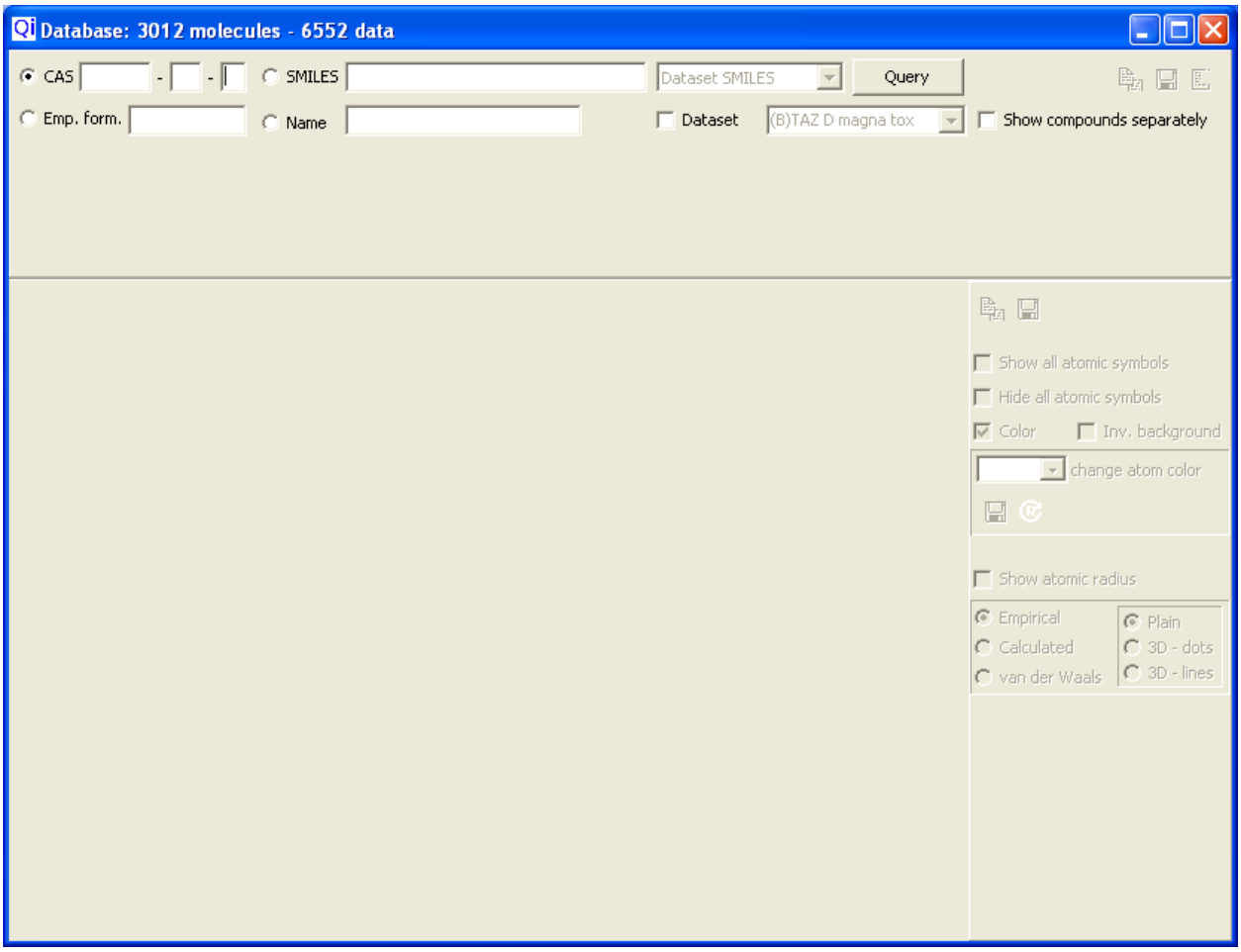


***Figure 39**. Database dialog at its first appearance*

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).

12.1.1 Querying the database

Pressing the Query button leaving all the other fields blank, as shown in Figure 39, 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 40:

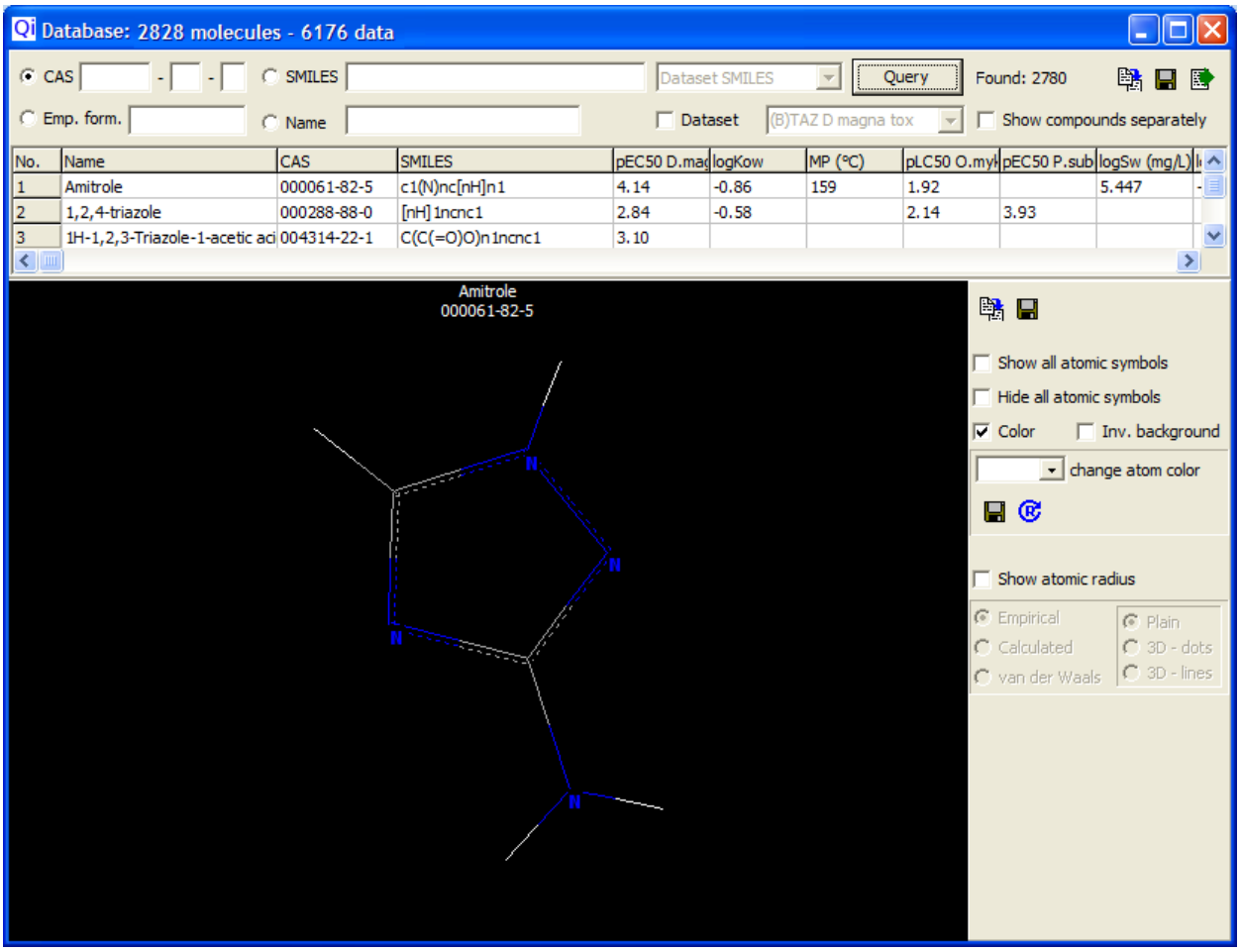
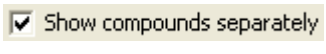


***Figure 40**. Database dialog when molecules are queried*

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 40, 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 12.1.3 "Preparing a user-defined dataset"), the option 

must be checked.

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

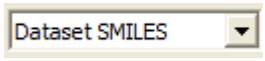
To query the compounds by CAS, the corresponding radio button must be pressed (

). 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 

will extract all the molecules that share 000050-00 whatever the last box value is, 

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 

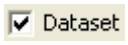
radio button must be selected: as a consequence the following 

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 (

) and by 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 

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 40, except that a new column containing the compound ID given in the original dataset (ID dataset column) will be shown, as in Figure 41:.

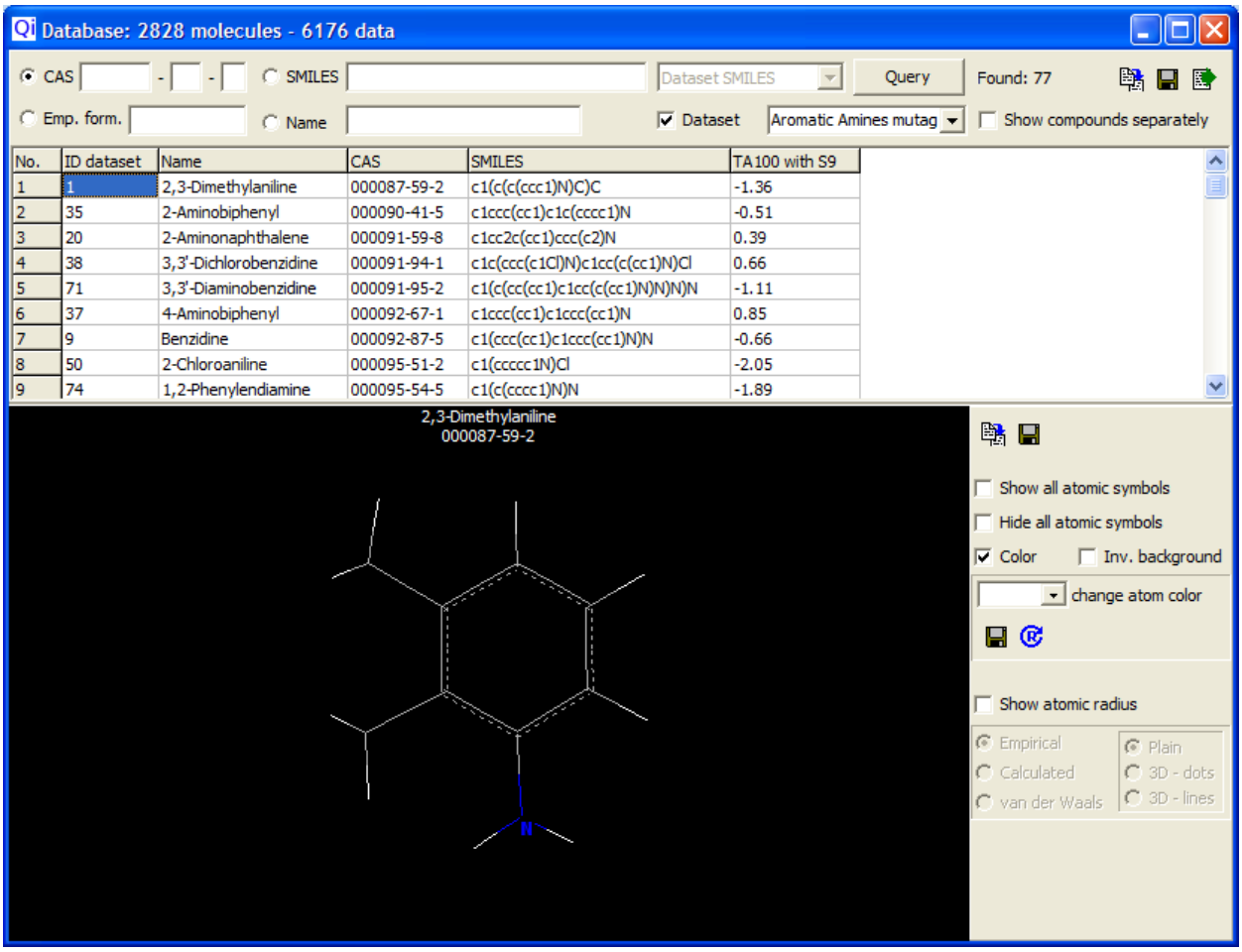
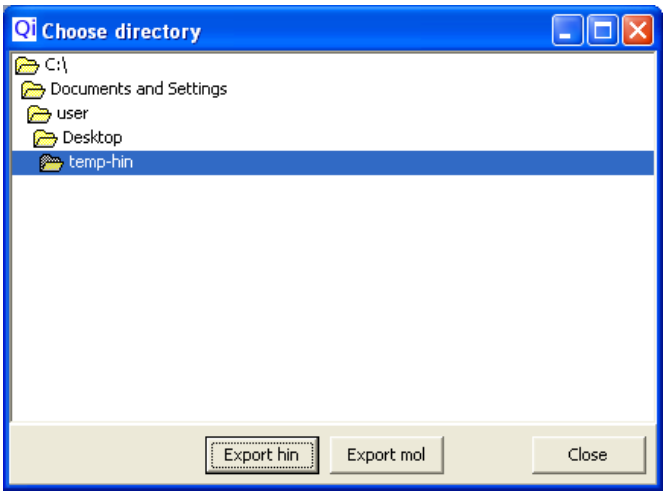


***Figure 41**. Database dialog using the Dataset option*

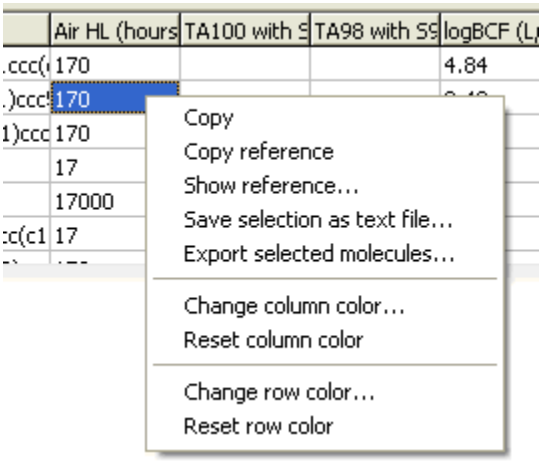
After querying the dataset, if compounds are found, some supporting tools will be activated ( ).

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:



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.

12.1.2 Visualization of molecules structure

In QSARINS it is also possible to visualize the 3D structure of the molecules as in the following example (Figure 42):

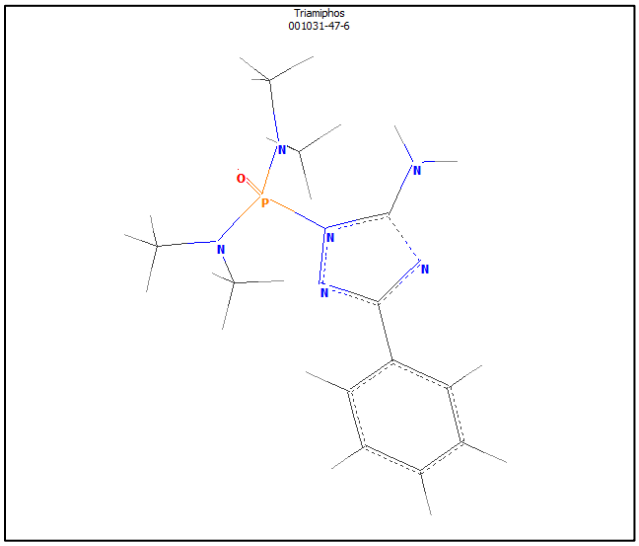


***Figure 42**. Example of a 3D view of a molecule*

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 43):

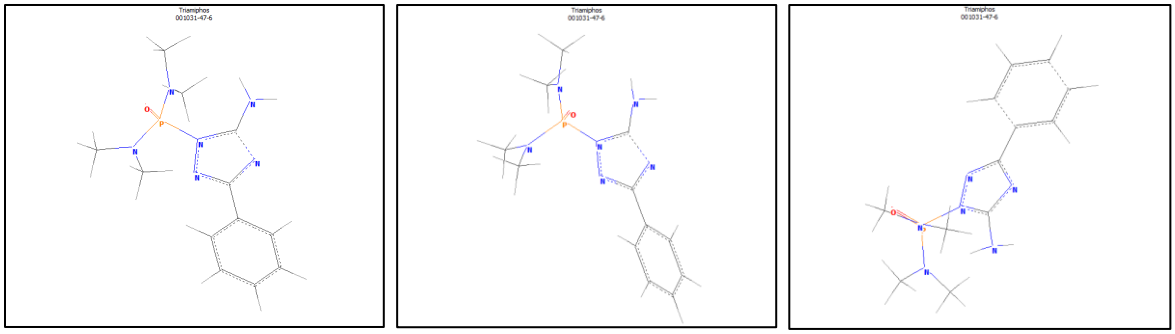


***Figure 43**. Rotating the 3D view of a molecule*

Pressing the left mouse key holding the Ctrl key down on the keyboard it can be scaled (Figure 44):

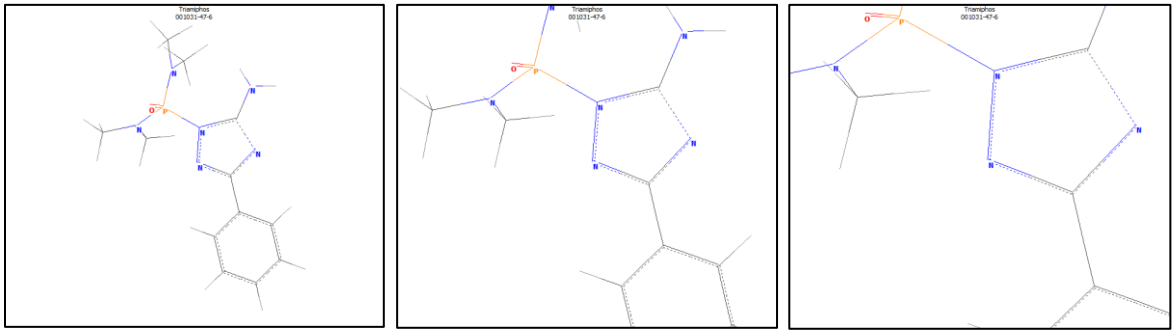


***Figure 44**. Zooming the 3D view of a molecule*

Pressing the left mouse key while keeping the shift key down the structure can be translated (Figure 45):

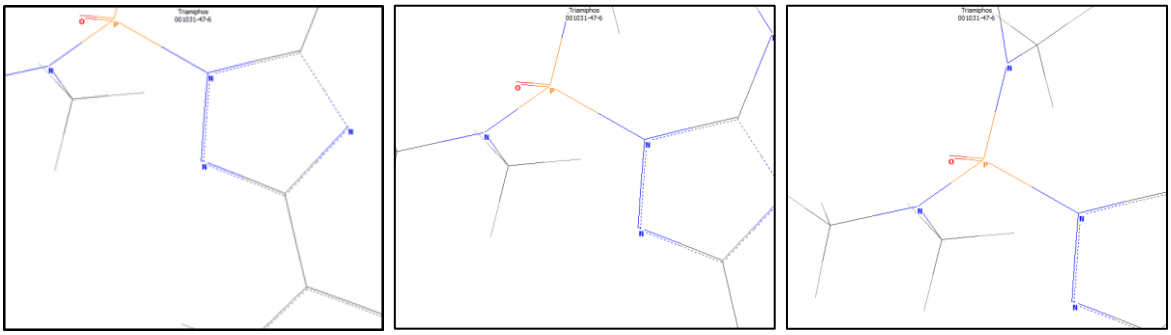


***Figure 45**. Translating the 3D view of a molecule*

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 40-41). 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 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 46):

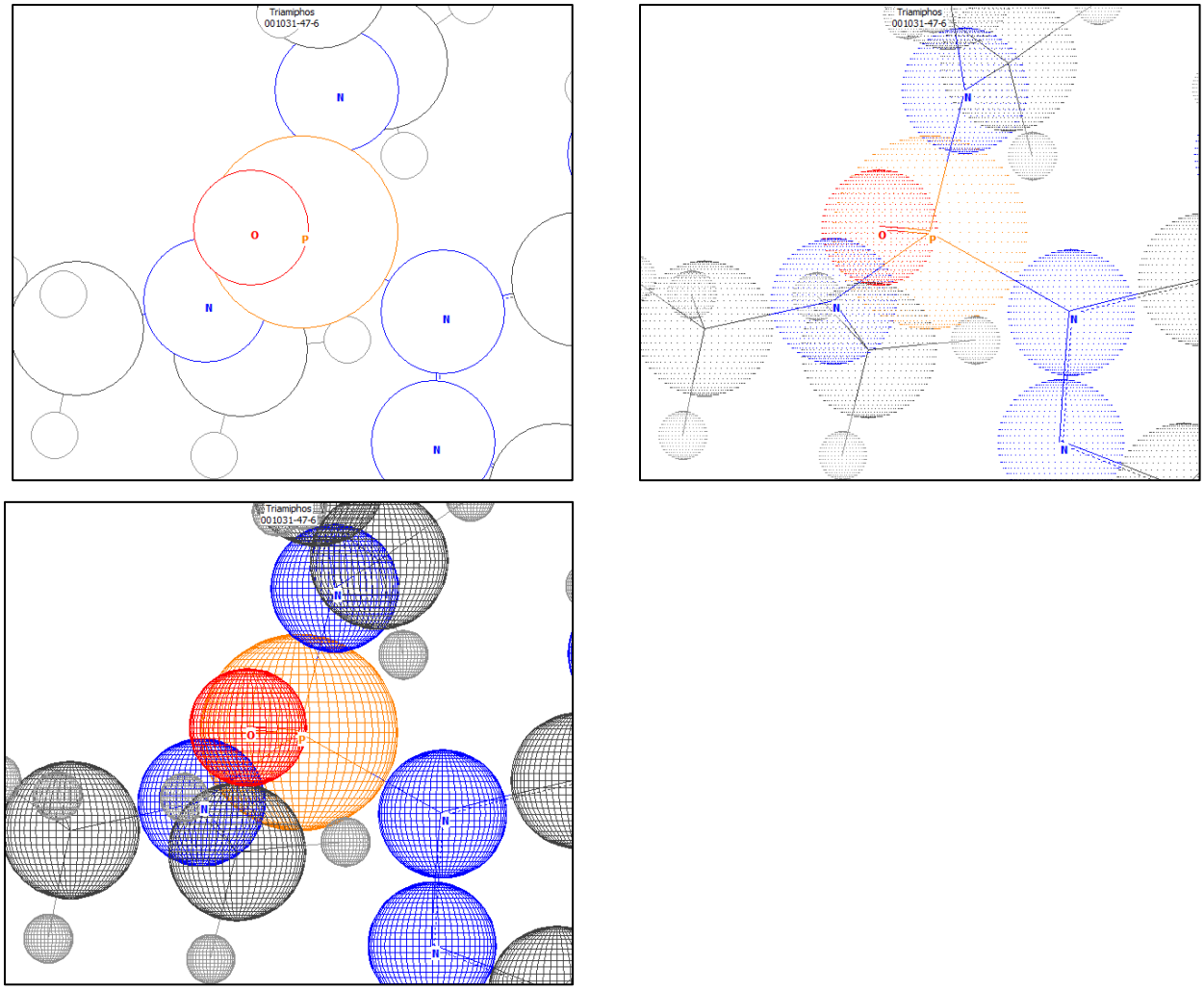


***Figure 46**. Display of the atomic radius in the 3D view of a molecule. Upper left: plain. Upper right: 3D-Dots. Lower: 3D-Lines.*

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

12.1.3 Preparing 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 (Figure 47):

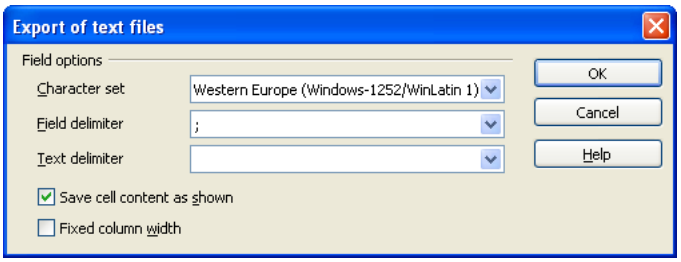
	A	B	C	D	E	F
1	ID	Name	CAS	Smiles	logH (Pa m3/mol, 25 °C)	QSAR & Comb.Sci. 28, 2009, pp 790-796
2	28	2,4,4'-triBDE	041318-75-6	c1(cc(ccc1Oc1ccc(cc1)Br)Br)Br	0.684	
3	47	2,2',4,4'-tetraBDE	005436-43-1	c1(cc(ccc1Oc1ccc(cc1Br)Br)Br)Br	-0.071	
4	99	2,2',4,4',5-pentaBDE	060348-60-9	c1(cc(c(cc1Oc1ccc(cc1Br)Br)Br)Br	-0.222	
5	100	2,2',4,4',6-pentaBDE	189084-64-8	c1(cc(cc(c1Oc1ccc(cc1Br)Br)Br)Br	-0.62	
6	153	2,2',4,4',5,5'-hexaBDE	068631-49-2	c1(cc(c(cc1Oc1cc(c(cc1Br)Br)Br)B	-0.585	
7	154	2,2',4,4',5,6'-hexaBDE	207122-15-4	c1(cc(c(cc1Oc1c(cc(cc1Br)Br)Br)B	-1.097	
8	209	2,2',3,3',4,4',5,5',6,6'-decaBDE	001163-19-5	c1(c(c(c(c(c1Oc1c(c(c(c(c1Br)Br)B	-1.398	

Figure 47 *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 10.1.1), 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 48) 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: 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 48):

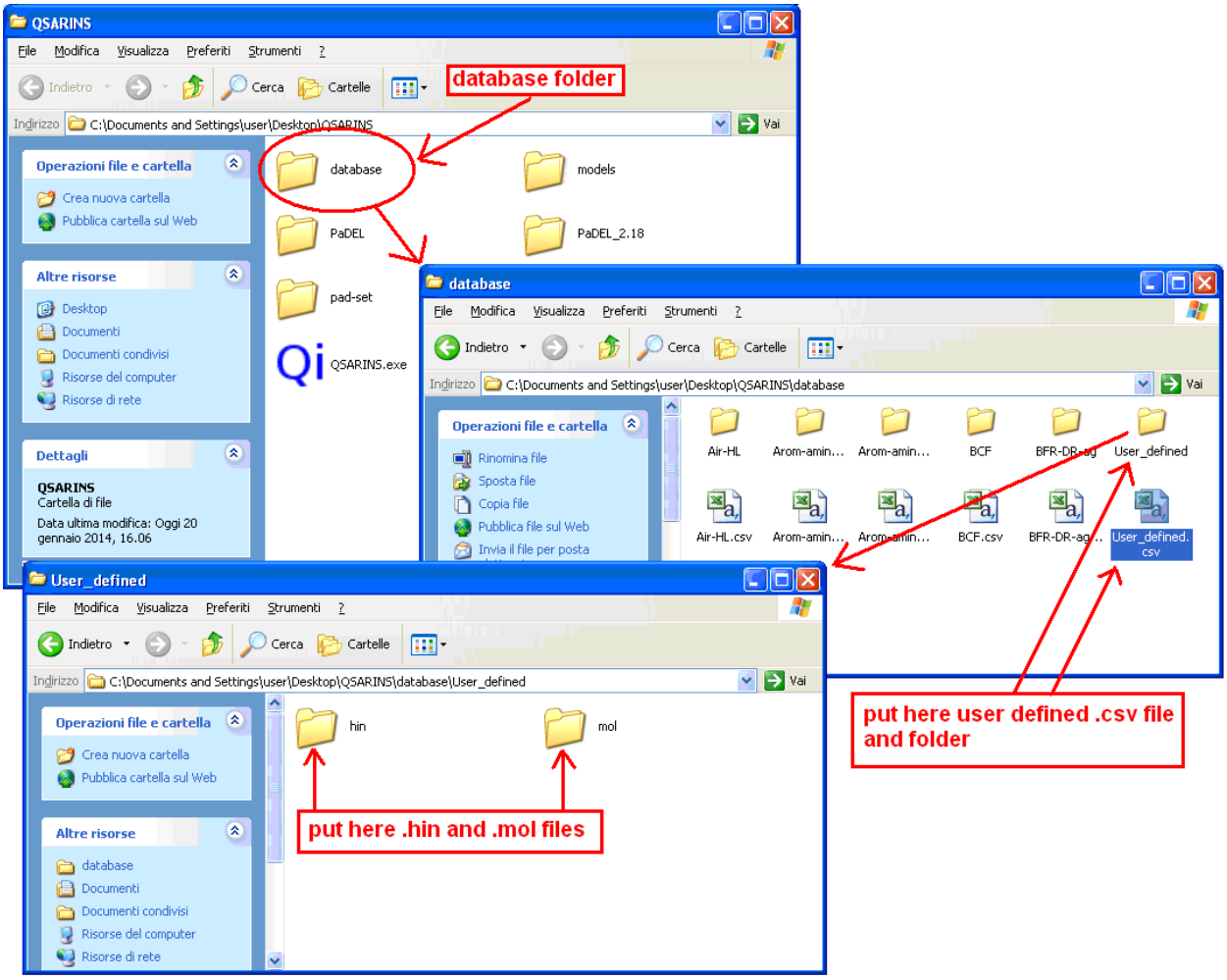


Figure 48. *Dataset files and folders location*

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 47) 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.

12.2 Apply developed model for new chemical prediction

All the stored models (see Section 10, “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 49):

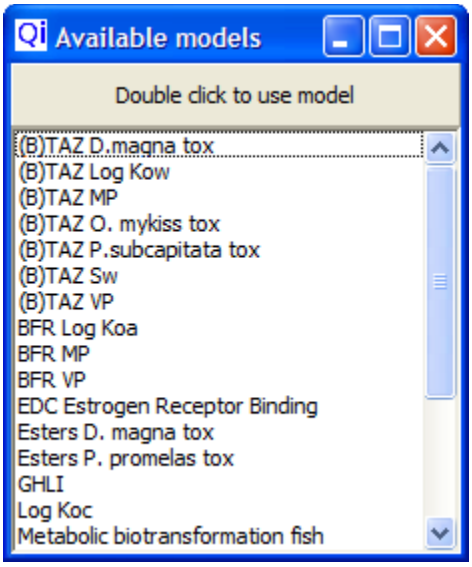


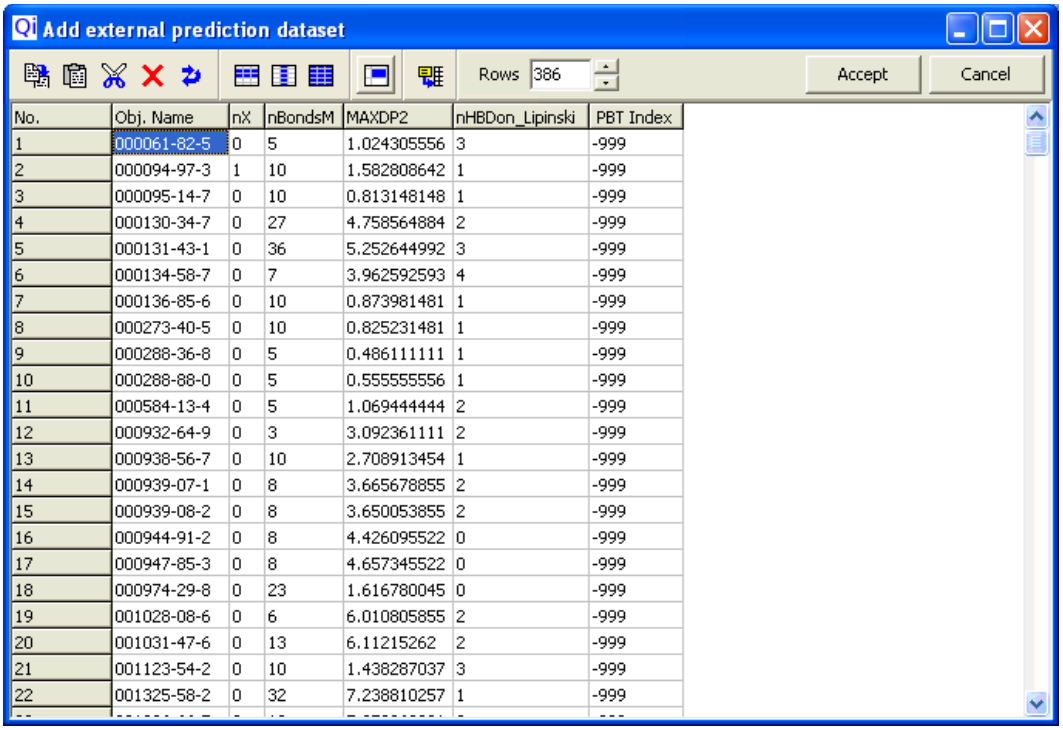
Figure 49*. List of available models*

Double-clicking the preferred model in the list, another dialog appears, which allows setting the input of new chemicals for their predictions (Figure 50). Now, in QSARINS it is possible to

calculate the modeling descriptors from PaDEL-Descriptor (icon, see below) and automatically fill the fields reported in Figure 50.

Otherwise, if the user wants to apply its personal MLR model developed with its own descriptors, 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.

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).



No.	Obj. Name	nX	nBondsM	MAXDP2	nHBDon_Lipinski	PBT Index
1	000061-82-5	0	5	1.024305556	3	-999
2	000094-97-3	1	10	1.582808642	1	-999
3	000095-14-7	0	10	0.813148148	1	-999
4	000130-34-7	0	27	4.758564884	2	-999
5	000131-43-1	0	36	5.252644992	3	-999
6	000134-58-7	0	7	3.962592593	4	-999
7	000136-85-6	0	10	0.873981481	1	-999
8	000273-40-5	0	10	0.825231481	1	-999
9	000288-36-8	0	5	0.486111111	1	-999
10	000288-88-0	0	5	0.555555556	1	-999
11	000584-13-4	0	5	1.069444444	2	-999
12	000932-64-9	0	3	3.092361111	2	-999
13	000938-56-7	0	10	2.708913454	1	-999
14	000939-07-1	0	8	3.665678855	2	-999
15	000939-08-2	0	8	3.650053855	2	-999
16	000944-91-2	0	8	4.426095522	0	-999
17	000947-85-3	0	8	4.657345522	0	-999
18	000974-29-8	0	23	1.616780045	0	-999
19	001028-08-6	0	6	6.010805855	2	-999
20	001031-47-6	0	13	6.11215262	2	-999
21	001123-54-2	0	10	1.438287037	3	-999
22	001325-58-2	0	32	7.238810257	1	-999

Figure 50*. Dialog for adding external chemicals to existing model*

The first group of icons in Figure 50 ( ) 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

The second group of icons in Figure 50 () simplify the selection of multiple cells:  extends the selected cells to the corresponding rows,  extends the selection to the corresponding columns and  selects all data.

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.

As said before, it is also possible to calculate the descriptors from external sources (i.e. PaDEL-Descriptor software) pressing the  icon. This will open a dialog box asking to select the folder containing the structural files: once selected, PaDEL-Descriptor will calculate the descriptors and the output will be automatically loaded in the data grid.

The descriptors of the models provided by default with QSARINS were calculated using the PaDEL-Descriptor version 2.18, so this version is used, for consistency, when the new descriptors are calculated.

The descriptors of new models generated by the user (see below) are calculated using the latest version of PaDEL-Descriptor configured in QSARINS (currently the 2.21, or a new version updated as described in section 2.1 “Updating PaDEL-Descriptor software”).

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

“Rows” option: it must match the number of chemicals to be tested if descriptors values are copied from external sources or manually written. The correct value is automatically set if descriptors are calculated using the “Calculate descriptors” option ( ).

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 10).

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 44 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 26).

12.2.1 Configuring user-defined models for descriptors calculation

To add the possibility of calculate automatically the descriptors for user defined models (see Section 10, “Analysis of single models”, to see how to create one) the user must go to the main window of QSARINS (see Figure 1) and run PaDEL-Descriptor from the icon ( ) or the corresponding menu “Tools > Run PaDEL-Descriptor”. When loaded PaDEL-Descriptor, the aim is to create the configuration files for descriptors calculation. The first step is to set the PaDEL-Descriptor options as needed and then the configuration files must be saved in the correct place with the correct names. For example, let us assume the user model is named “user-model.smd”: from the PaDEL menu “File > Save configuration”, save a file named “user-model.cfg” in the pad-set folder, located in QSARINS folder, as showed in the figure below (Figure 51). Then from the menu “File > Save descriptor types” save a file called “user-model.xml” in the pad-set folder too.

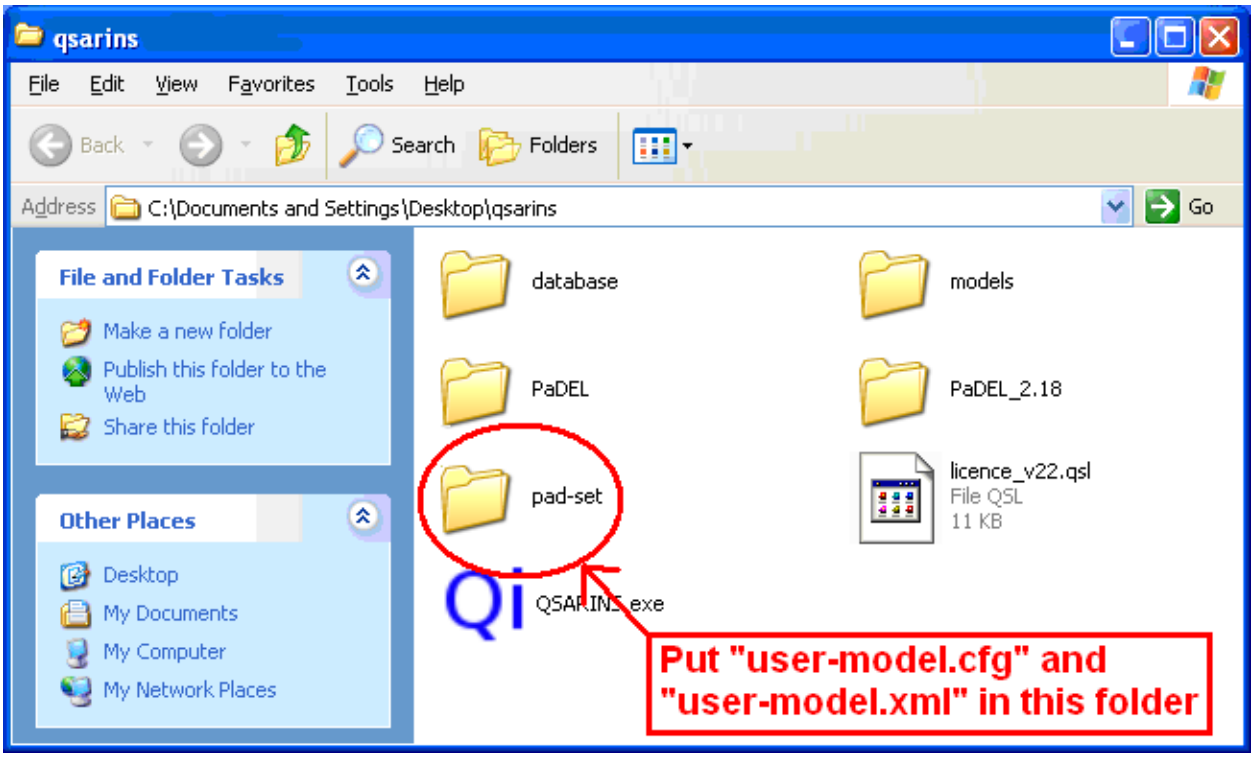


***Figure 51**. Folder where to put descriptors configuration files*

It is essential that the configuration file names (i.e. “user-model”, in the aforementioned example) are the same (beware of blank spaces!), and the extensions are the ones required (“.cfg” and “.xml”), otherwise the descriptors will be not calculated.

12.2.2 Apply developed model: PBT Index example

The PBT Index, included in the stored models available in QSARINS (Figure 49), is also achievable from the main software window (Figure 1) , where the **PBT index** button (or Tools->Apply PBT index model) applies the model developed and proposed 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:

PBT Index = -1.44 (±0.10) + 0.65 (±0.03) nX + 0.22 (±0.01) nBM - 0.39 (±0.06) nHDon - 0.07 (±0.03) MAXDP

n = 180, R^2 = 88.40%; Q^2_{LOO} = 87.72%; Q^2_{LMO} = 87.73%; R^2_{YS} = 0.02; 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^2ext $_{(F1)}$ = 80.72%; R^2_{ext} = 89.27%; $RMSE_P$ = 0.72)

The model implemented in QSARINS software is the following, redeveloped and based on freely available software for descriptors calculation PaDEL-Descriptor (Yap, 2011), published in Gramatica et al . 2013.

PBT Index = -1.46 (±0.19) + 0.64 (±0.05) nX + 0.22 (±0.02) nBondsM - 0.39 (±0.13) nHBDon_Lipinksi - 0.06 (±0.06) MAXDP2

n = 180, R^2 = 88.89%; Q^2_{LOO} = 88.25%; Q^2_{LMO} = 88.28%;R^2_{YS} = 0.02; $RMSE_{TR}$ = 0.51;
Q^2ext $_{(F1)}$* = 89%; $RMSE_P$ *= 0.49.
*: split model, with 92 compounds in Training set and 88 compounds in 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, calculating the four molecular descriptors directly in QSARINS (from PaDEL-Descriptor software).

1) Choose and enter in PBT Index model of QSARINS, pressing on 

button or selecting it from the 

icon;

2) Clicking on “Calculate descriptors” icon () (see Figure 50);

3) Selecting the folder that contains the input file (MDL MOL, .hin, SMILES and so on) for descriptors calculations, and clicking on “Accept”;

4) In few seconds, the rows and columns will be automatically filled with the calculated descriptors (in this case, nX, nBondsM, MAXDP2 and nHBDon_Lipinski) of any studied chemical;

5) Apply the PBT Index model simply pushing the button “Accept” (see Figure 50);

6) 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);

7) 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”) respectively For the explanation of the graphs, see the Section 10 ”Analysis of single models”;

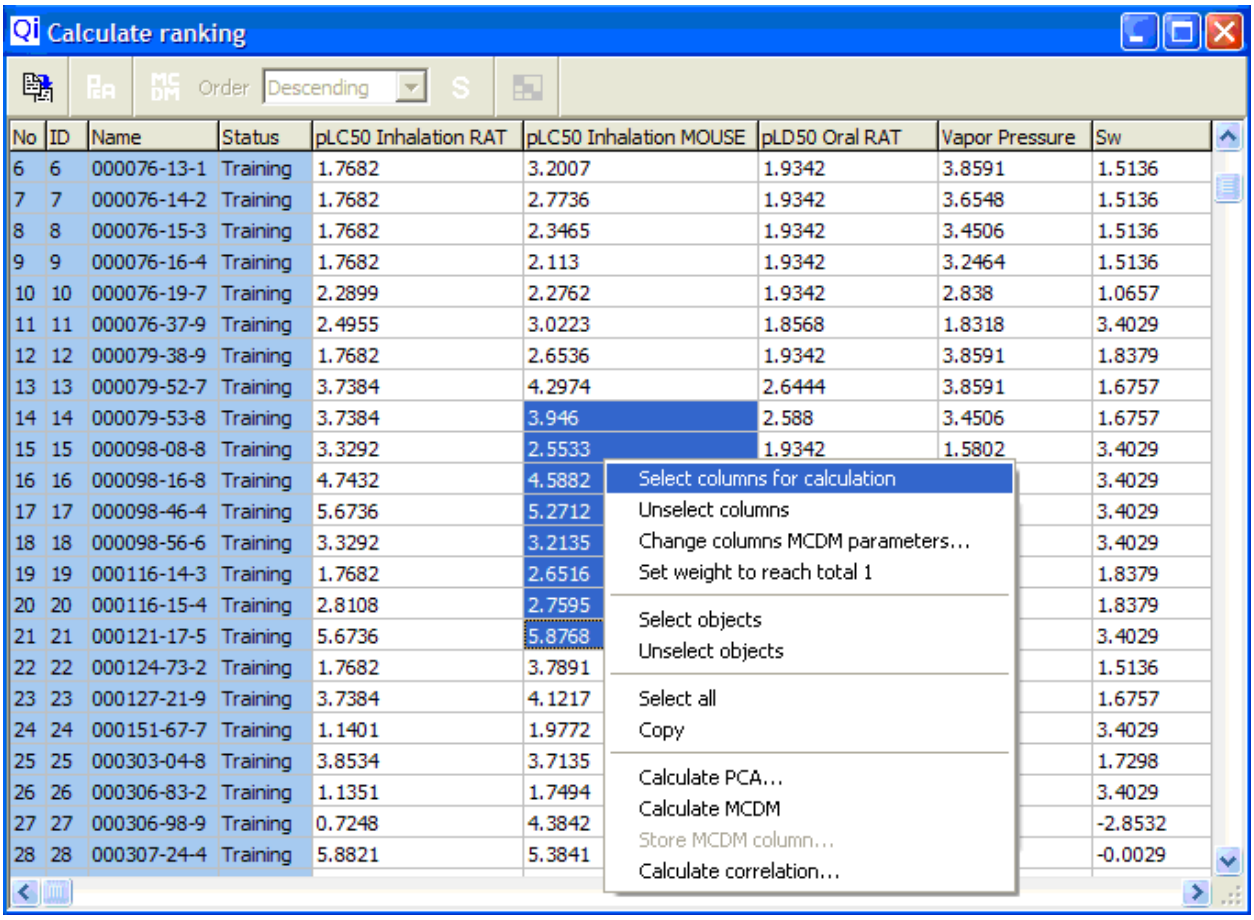
8) The user can save and/or copy the results.

Any other proprietary single model can be used by QSARINS if it is saved in the appropriate models folder (see section 10, “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 44 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.

12.3. Ranking

Once imported a dataset, pressing the ranking icon (or, alternatively, selecting Tools->Calculate ranking menu item) in the main screen, the following dialog (Figure 52) 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):



No	ID	Name	Status	pLC50 Inhalation RAT	pLC50 Inhalation MOUSE	pLD50 Oral RAT	Vapor Pressure	Sw
6	6	000076-13-1	Training	1.7682	3.2007	1.9342	3.8591	1.5136
7	7	000076-14-2	Training	1.7682	2.7736	1.9342	3.6548	1.5136
8	8	000076-15-3	Training	1.7682	2.3465	1.9342	3.4506	1.5136
9	9	000076-16-4	Training	1.7682	2.113	1.9342	3.2464	1.5136
10	10	000076-19-7	Training	2.2899	2.2762	1.9342	2.838	1.0657
11	11	000076-37-9	Training	2.4955	3.0223	1.8568	1.8318	3.4029
12	12	000079-38-9	Training	1.7682	2.6536	1.9342	3.8591	1.8379
13	13	000079-52-7	Training	3.7384	4.2974	2.6444	3.8591	1.6757
14	14	000079-53-8	Training	3.7384	3.946	2.588	3.4506	1.6757
15	15	000098-08-8	Training	3.3292	2.5533	1.9342	1.5802	3.4029
16	16	000098-16-8	Training	4.7432	4.5882			3.4029
17	17	000098-46-4	Training	5.6736	5.2712			3.4029
18	18	000098-56-6	Training	3.3292	3.2135			3.4029
19	19	000116-14-3	Training	1.7682	2.6516			1.8379
20	20	000116-15-4	Training	2.8108	2.7595			1.8379
21	21	000121-17-5	Training	5.6736	5.8768			3.4029
22	22	000124-73-2	Training	1.7682	3.7891			1.5136
23	23	000127-21-9	Training	3.7384	4.1217			1.6757
24	24	000151-67-7	Training	1.1401	1.9772			3.4029
25	25	000303-04-8	Training	3.8534	3.7135			1.7298
26	26	000306-83-2	Training	1.1351	1.7494			3.4029
27	27	000306-98-9	Training	0.7248	4.3842			-2.8532
28	28	000307-24-4	Training	5.8821	5.3841			-0.0029

***Figure 52**. 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 52). To select all data and copy to the clipboard, press “Select all” and then “Copy” options.

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 columns 0.34 (this is a trivial example, but in practice it can became 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 trimmed to reach a total of 1.

Once the desired columns are selected and the weights/optimality are assigned according to the user needs, pressing the MCDM 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 MCDM button the objects scores will be sorted accordingly. Once the scores are calculated, the S 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.

Note that in the PCA calculated here for ranking purposes, there is also the possibility to visualize within the graph (Score plot) the obtained MCDM order value for every object (chemicals, in our case), selecting the “MCDM order” option in the “Label type” drop box (Figure 53). 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 52, identified as “No”). If the PCA Score plot is visualized before the calculation of MCDM, when asking for the score order, every object (chemical) is marked with n.a., since no MCDM score is available yet (Figure 53 left); after the calculation of MCDM scores, the MCDM order will be available in the PCA score plot (Figure 53,right).

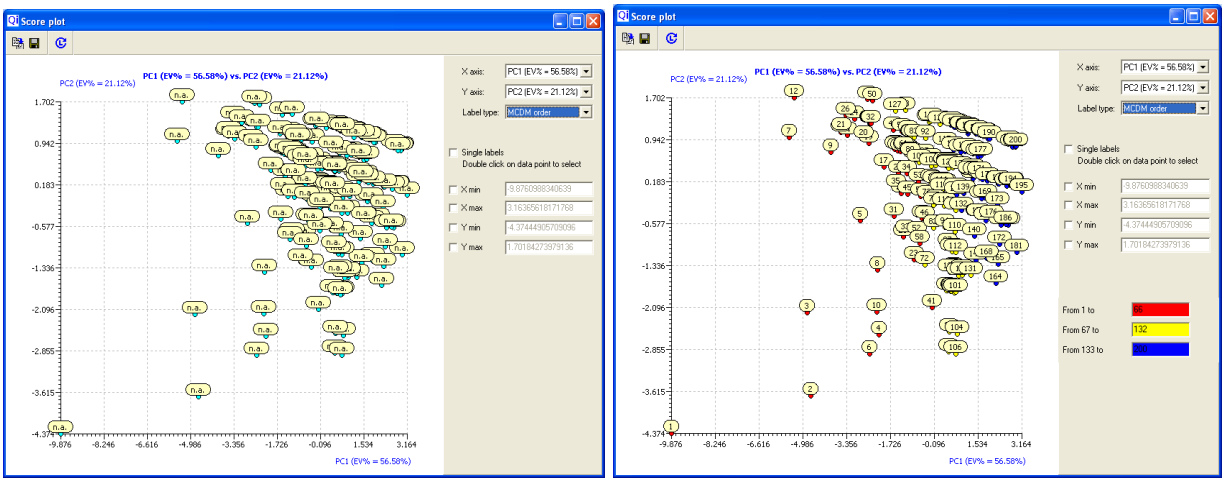


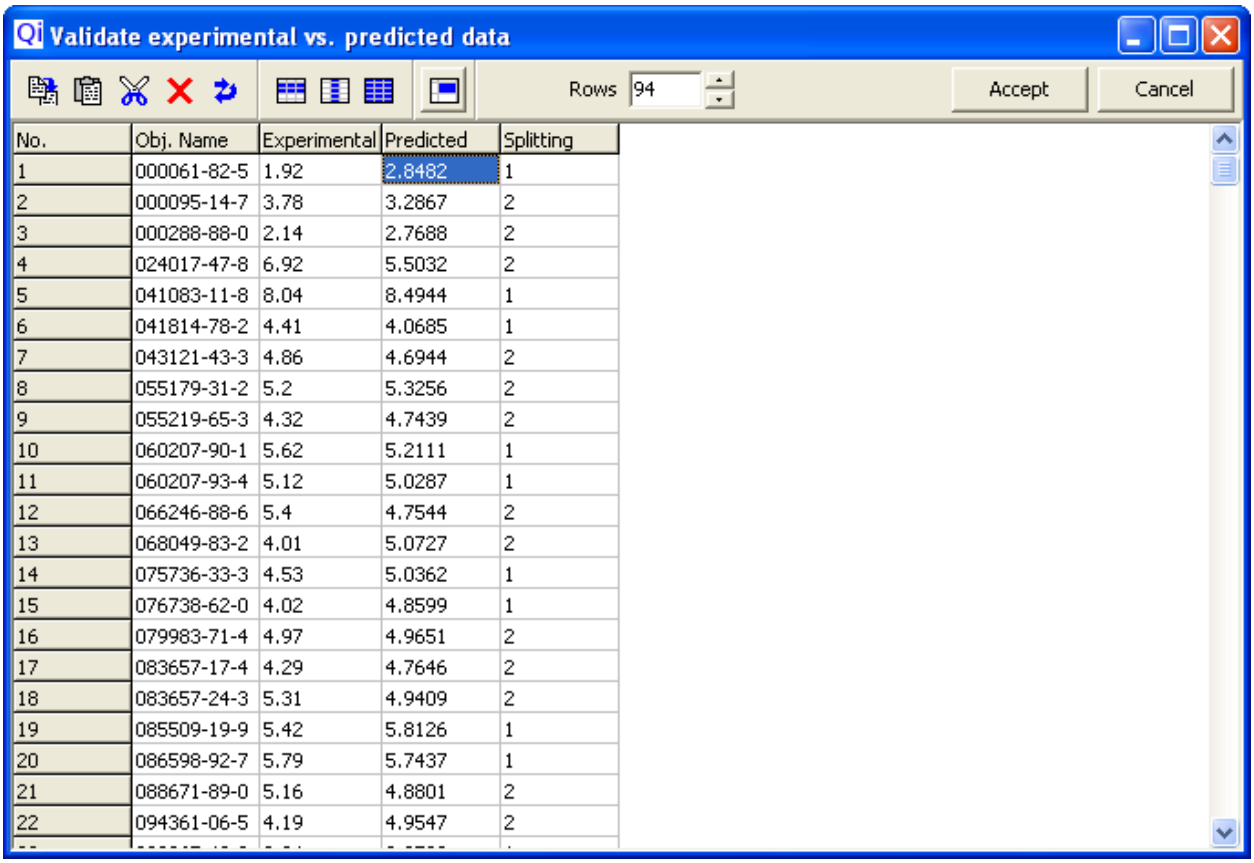
Figure 53 *PCA Score plots where is possible to visualize the obtained MCDM order for every object.*

When the MCDM score has been calculated, on visualizing the score plot the MCDM order labels become available and three colours corresponding to three different ranking ranges will be activated. By default the ranges equally divides the whole range of the MCDM scores order. For example, in the right Figure 53, 200 molecules are used for ranking: the red color MCDM ordered scores range from 1 to 66, the yellow ones range from 67 to 132 and finally the blue ones from 133 to 200. 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.

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

13. 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 (Figure 54) appears, allowing the user entering experimental and predicted data and, optionally, a splitting column (1 for Training Set, 2 for Prediction Set; if not used, it must be left blank).



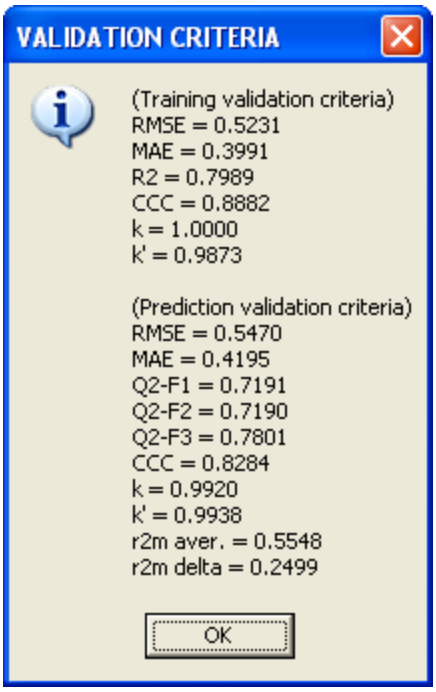
No.	Obj. Name	Experimental	Predicted	Splitting
1	000061-82-5	1.92	2.8482	1
2	000095-14-7	3.78	3.2867	2
3	000288-88-0	2.14	2.7688	2
4	024017-47-8	6.92	5.5032	2
5	041083-11-8	8.04	8.4944	1
6	041814-78-2	4.41	4.0685	1
7	043121-43-3	4.86	4.6944	2
8	055179-31-2	5.2	5.3256	2
9	055219-65-3	4.32	4.7439	2
10	060207-90-1	5.62	5.2111	1
11	060207-93-4	5.12	5.0287	1
12	066246-88-6	5.4	4.7544	2
13	068049-83-2	4.01	5.0727	2
14	075736-33-3	4.53	5.0362	1
15	076738-62-0	4.02	4.8599	1
16	079983-71-4	4.97	4.9651	2
17	083657-17-4	4.29	4.7646	2
18	083657-24-3	5.31	4.9409	2
19	085509-19-9	5.42	5.8126	1
20	086598-92-7	5.79	5.7437	1
21	088671-89-0	5.16	4.8801	2
22	094361-06-5	4.19	4.9547	2

***Figure 54**. Dialog for adding experimental and predicted data for validation.*

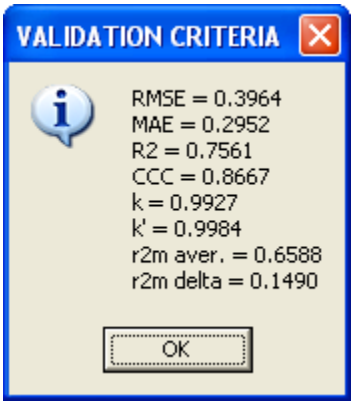
For entering the data, the dialog works in the same manner as in section 12.2 “Apply developed model”, so it is suggested to read it as a reference.

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.



After the validation criteria dialog box, the plotting of experimental vs. predicted and the plotting of the residuals will appear, as exemplified in Figure 56.

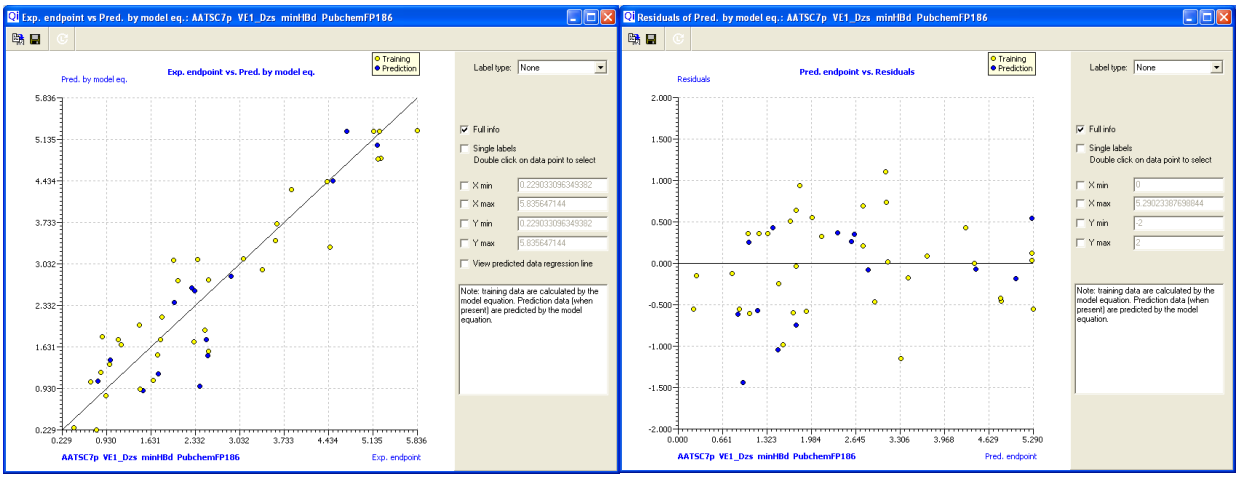


Figure 55 *Graph of experimental vs. predicted data and residuals.*

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 Making (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 3014 chemicals, studied and modeled by the Insubria group, are available with their 3D structure and experimental responses. In addition, 44 QSAR/QSPR/QAAR models of environmental end-points, based on free software for molecular descriptors (PaDEL-Descriptor 2.18/2.21) are available. These models, supported by their QMRF, can be applied for any new chemical, belonging to the applicability domain, verified by the Insubria graph (Figure 35)

It is important to note that any user can also 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

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.

Eriksson, L., Joanna Jaworska, J., Worth, A., Mark Cronin, M., McDowell, R.M., Gramatica P. 2003. Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs *Environ. Health Perspectives 111 (10),* 1361-1375.

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

Golbraikh, A., Tropsha, A. 2002. Beware of q2. *J. Mol. Graphics Model.*, 20, 269-276.

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., 2014. External Evaluation of QSAR Models, in Addition to Cross-Validation: Verification of Predictive Capability on Totally New Chemicals, *Mol. Inf*. 33, 311-314.

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,* 2014, *35*, 1036–1044.

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.

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.

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., Sangion, A. 2016. A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology *J Chem Inf Mod* 56 (6), 1127-31.

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

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

Katritzky, A.R., Dobchev, D.A., Slavov, S., Karelson, M. 2008. Legitimate Utilization of Large Descriptor Pools for QSPR/QSAR Models. J. Chem. Inf. Model. 48, 2207–2213.

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 r2m 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.

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

Rücker, C., Rücker, G., Meringer, M. 2007. y-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.

Tropsha, A., Gramatica, P., Gombar, V.K. 2003. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR &Comb. Sci.*, 22, 69-77

Wehrens, R., Putter, H., Buydens, L.M.C. 2000. The boostrap: 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. Availabe online at http://padel.nus.edu.sg/software/padeldescriptor/index.html