



QSAR-ME Profiler beta version

Manual



QSAR Research Unit in Environmental Chemistry and Ecotoxicology
Department of Theoretical and Applied Sciences (DiSTA)
University of Insubria

Information

The software and models were developed by:

QSAR Research Unit in Environmental Chemistry and Ecotoxicology
Department of Theoretical and Applied Sciences (DiSTA)
University of Insubria, Varese, Italy
<https://dunant.dista.uninsubria.it/qsar/>

Funding information: University of Insubria, Post Doc grant (2021-2022) "In silico solutions for the assessment of biotransformation related endpoints of organic chemicals in multiple organisms."

Contacts

Prof. Ester Papa - e-mail: ester.papa@uninsubria.it (project responsible)
Dr. Nicola Chirico, PhD - e-mail: nicola.chirico@uninsubria.it (software development)

How to cite:

Please cite QSAR-ME Profiler beta version in your publications as:

Chirico N., Bertato L., Papa E., QSAR Multiple Endpoint Profiler (QSAR-ME Profiler) beta version, 2022, freely available at: <https://dunant.dista.uninsubria.it/qsar/>

Limitations of liability and disclaimer of warranty

QSAR-ME Profiler beta version and the accompanying materials and manuals are provided "as they are" without warranty of any kind. The authors do not warrant, guarantee, or make any representations, either expressed or implied, regarding the use, or the results from the use of QSAR-ME Profiler beta version, the accompanying materials and manuals, in terms of correctness, accuracy, reliability, currentness, or otherwise.

You assume the entire risk as to the result and performance of QSAR-ME Profiler beta version.

In no event shall the authors be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with QSAR-ME Profiler beta version or the use or other dealings in QSAR-ME Profiler beta version, even if the authors have been advised of the possibility of such damages.

The QSAR-ME Profiler beta version software, the accompanying materials and manuals are protected by copyright: 2022, University of Insubria, <http://www.uninsubria.it> - Varese, Italy.

INDEX

1	Introduction.....	3
2	System requirements and installation.....	3
3	Running QSAR-ME Profiler	3
3.1	Selecting a QSAR.....	5
3.2	Selecting training chemicals and neighbors	7
3.3	Predicting the endpoint of new chemicals	9
3.4	Selecting user-entered chemicals and neighbors.....	11
3.5	Customizing the main window	13
3.6	Available QSAR diagnostic charts	14
3.6.1	MLR QSAR graphical diagnostics	14
3.6.2	LDA QSAR graphical diagnostics	17
3.7	Endpoint prediction reports of user entered chemicals	18
3.7.1	MLR QSAR reports	18
3.7.2	MLR models for the prediction of biotransformation <i>in vitro</i>	19
3.7.3	LDA QSAR Reports	19
4	How to customize QSAR-ME Profiler available QSARs	20
4.1	QSAR categories	20
4.1.1	How to add or delete a QSAR category	21
4.1.2	How to create a QSAR category	21
4.2	How to create QSAR files.....	22
4.2.1	How to create MLR QSAR XML and CSV files	22
4.2.2	How to create Toxtree derived MLR QSAR XML and CSV files	26
4.2.3	How to create LDA QSAR XML and CSV files	26
4.3	How to create cache data.....	28
5	Models included in QSAR-ME Profiler	30
6	Acknowledgments	30

1 Introduction

QSAR-ME Profiler stands for "Quantitative Structure-Activity Relationship Multiple Endpoint Profiler", which is a multi-platform GUI driven user-friendly software allowing the application of QSARs for the prediction of the activity of new chemicals, supported by structural comparison.

QSAR-ME Profiler comes shipped with more than 100 QSARs¹ for the assessment of the potential hazard and risk of heterogeneous organic chemicals, developed in the last 20 years by the QSAR Research Unit in Environmental Chemistry and Ecotoxicology of the University of Insubria, covering physical-chemical properties, global indexes, aquatic toxicity, *in vivo* and *in vitro* mammalian biotransformation. These QSARs were developed according to the OECD Principles for regulatory purposes of (Quantitative) Structure-Activity Relationship models and are accompanied by the corresponding QMRF (QSAR Model Reporting Format) documents to help decisions in the regulatory context.

QSARs are applied in batch to new chemicals by QSAR-ME Profiler whose predictions, both as single and combined, are organized as browsable multiple tables. Predictions can also be checked, in the context of the applied QSARs, both in tabular and graphical forms. QSAR-ME Profiler calculates structural distances among the chemicals to further check the coherence of the prediction of the new chemicals, both among them and in the context of the applied QSARs. Chemical structures are also automatically depicted while browsing the chemicals, to further help the user in assessing the predictions.

Supporting software like PaDEL-Descriptor and Toxtree², for the descriptors calculation and metabolic reaction detection, are also automatically called and managed by QSAR-ME Profiler.

2 System requirements and installation

QSAR-ME Profiler runs with Java™ SE 17 (or more recent) runtime environment installed. To use QSAR-ME Profiler you need to unzip the compressed file in a folder of your choice, then open the folder and locate the QSAR-ME Profiler executable, named **QSAR-ME-Profiler.jar**.

3 Running QSAR-ME Profiler

To run QSAR-ME-Profiler.jar, double click on its icon or, if your system is not configured for running Java™ executables by double clicking, open a terminal where the file is located and type in **java -jar QSAR-ME-Profiler.jar**³. The first time you run QSAR-ME Profiler beta version, you will be asked to read and accept the Licence agreement, as shown in Figure 1.

¹ See "Models included in QSAR-ME Profiler" for further details. The set of available QSARs is fully customizable also allowing the addition of models developed by the user, see "How to customize QSAR-ME Profiler available QSARs" for further details.

² The user is allowed to perform the descriptors calculation and the reaction detection by other means, if preferred.

³ If it is not working, it usually depends on the configuration of the Java™ environment of your machine.

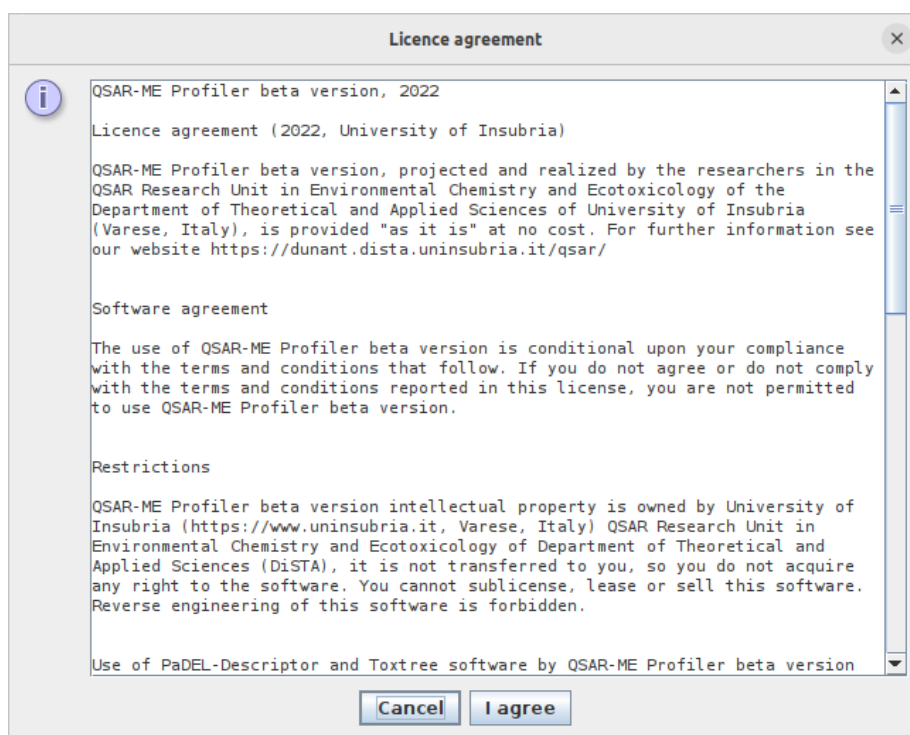


Figure 1. QSAR-ME Profiler licence agreement

Once accepted⁴ the main window of QSAR-ME Profiler will be displayed, like the one reported⁵ in Figure 2, which can be divided broadly into logical sections.

⁴ On subsequent running of QSAR-ME Profiler you will be not asked again unless you did not accept the licence.

⁵ Appearance may differ depending on your O.S. and/or desktop theme settings.

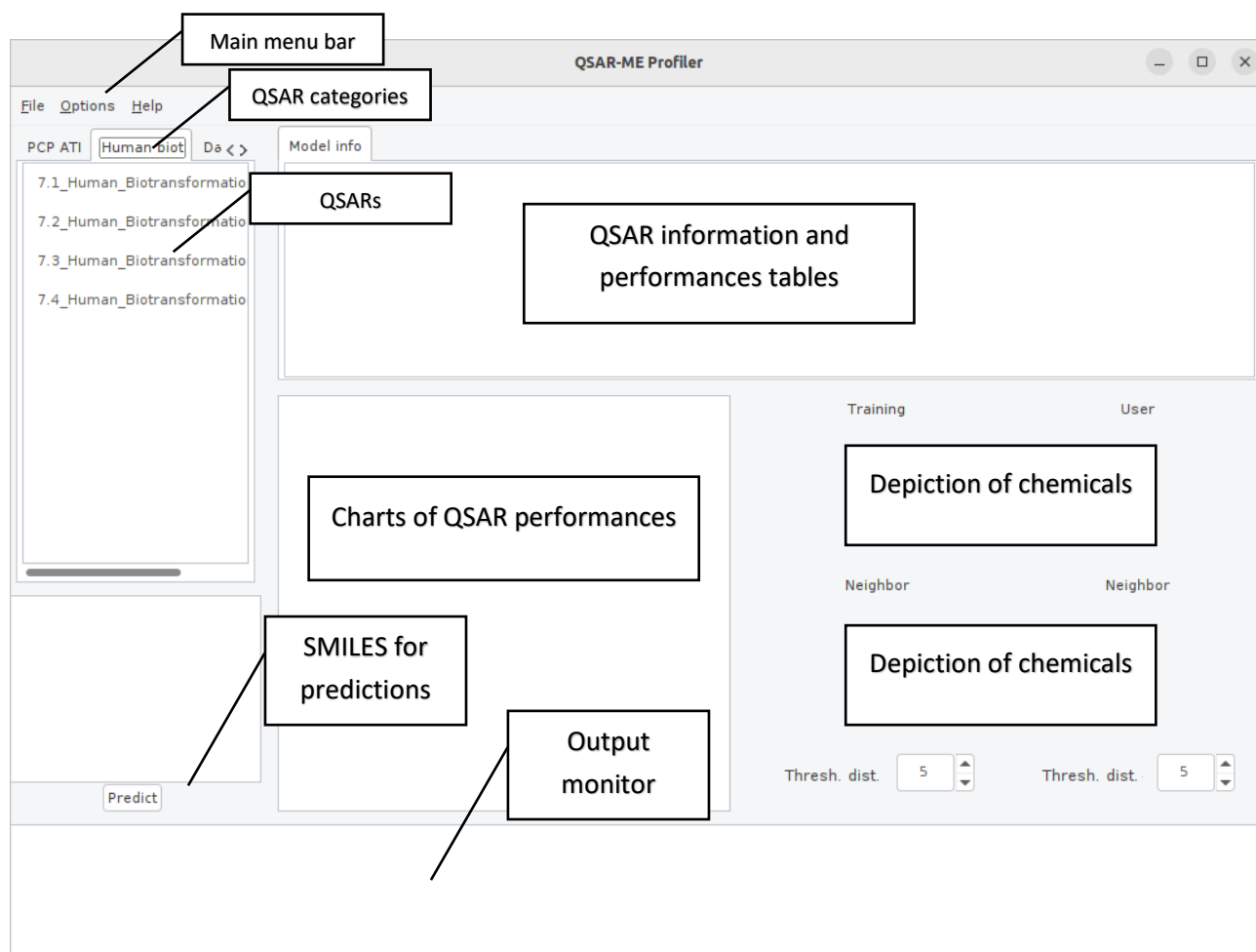


Figure 2. QSAR-ME Profiler main window logical sections

3.1 Selecting a QSAR

QSARs in QSAR-ME Profiler can be explored singularly (also) before their application for prediction of new chemicals endpoints. QSARs are organized in coherent categories like, for example, biotransformation in human hepatocytes or Toxicity in fish. To scroll the QSAR categories, press on the left or right arrows as indicated in Figure 3, then select the category by clicking on the corresponding tab and finally select the QSAR of interest from the list.

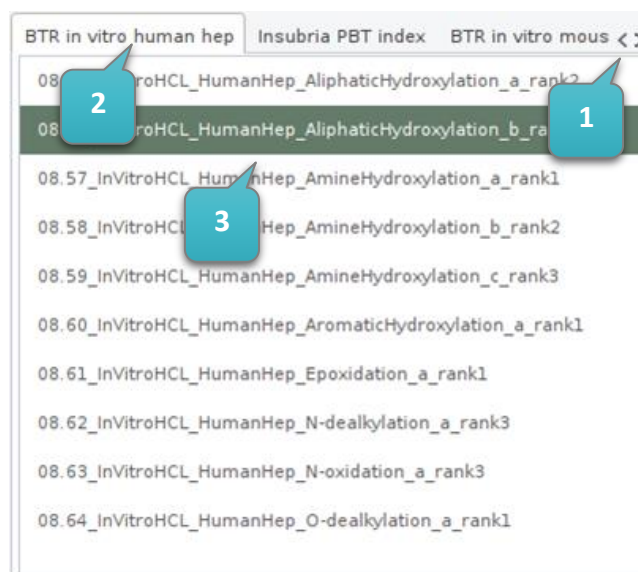


Figure 3. Selection of a QSAR. Arrows on the top right (1) allow scrolling the QSAR groups. For the selection of a QSAR you need to click on one of the QSAR group's tab (2) and then click on the QSAR of interest from the list (3)

Once selected, information, charts, and performances concerning the QSAR will be shown in the main window as in Figure 4.

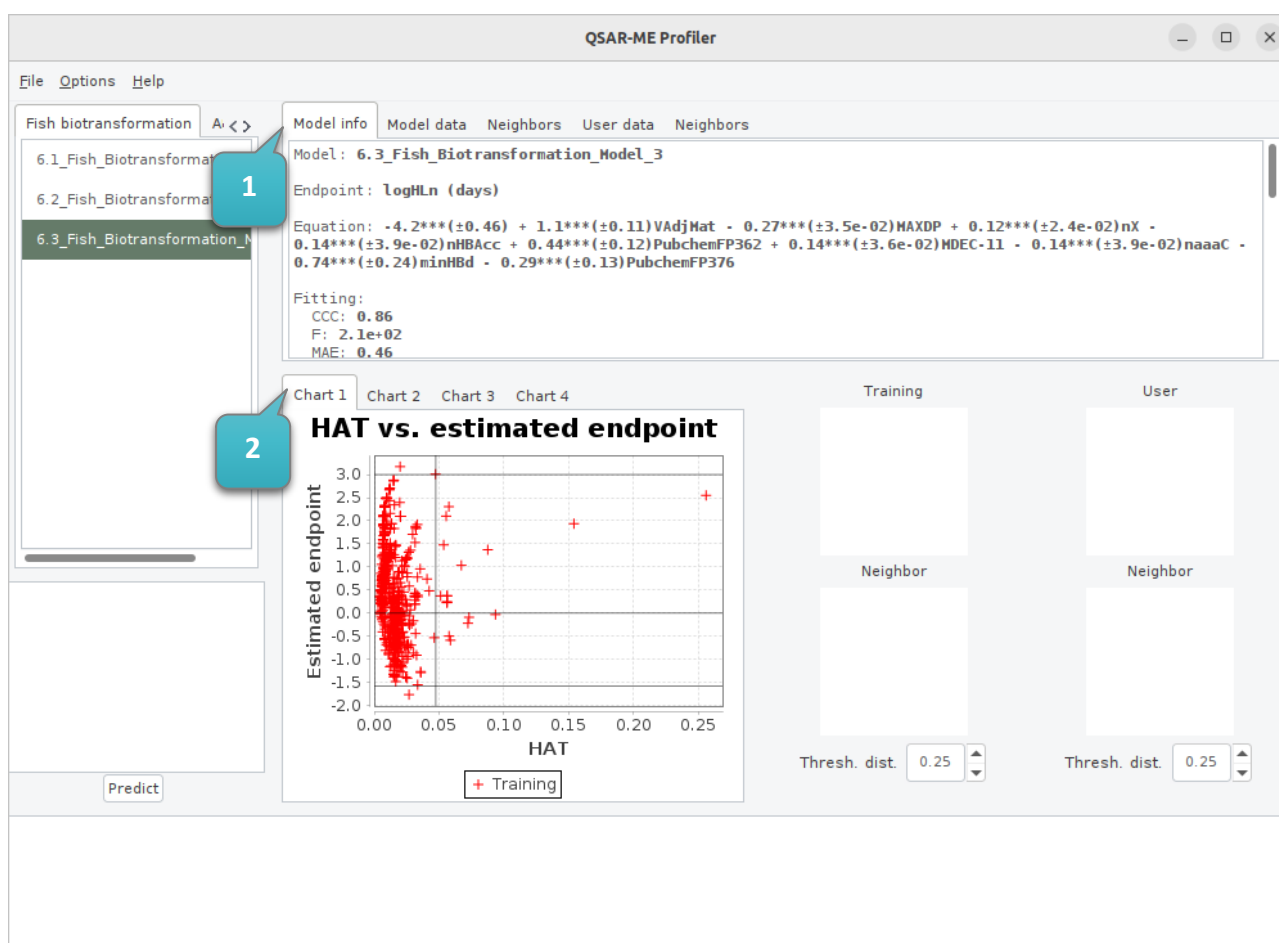


Figure 4. QSARs information tab (1) and charts (2)

The model info tab contains QSAR's information like the name, the equation and performances, while charts are displayed below. Four charts are available for MLR (Multiple Linear Regression) QSARs: HAT vs. estimated endpoint, experimental vs. predicted endpoint, endpoint residual and HAT vs. standardized residuals. ROC (Receiver Operating Characteristic) charts are displayed for LDA (Linear Discriminant Analysis) based QSARs.

3.2 Selecting training chemicals and neighbors

Once a QSAR has been selected, if you browse the training set chemicals, the most structurally similar chemicals are automatically detected and depicted. Chemicals can be selected in two ways (see also Figure 5): A) Select the "Model data" tab in the main window and then select a chemical from the list, B) for MLR QSARs only, hover the mouse cursor on a data point and then left click (note: when the mouse pointer is correctly located over a data point, a hint reporting the chemical's name is shown). Once selected, the data point is targeted with two red⁶ crossing lines to simplify its localization, while the structure of the chemical is shown in the "Training" depiction space, as shown in Figure 5.

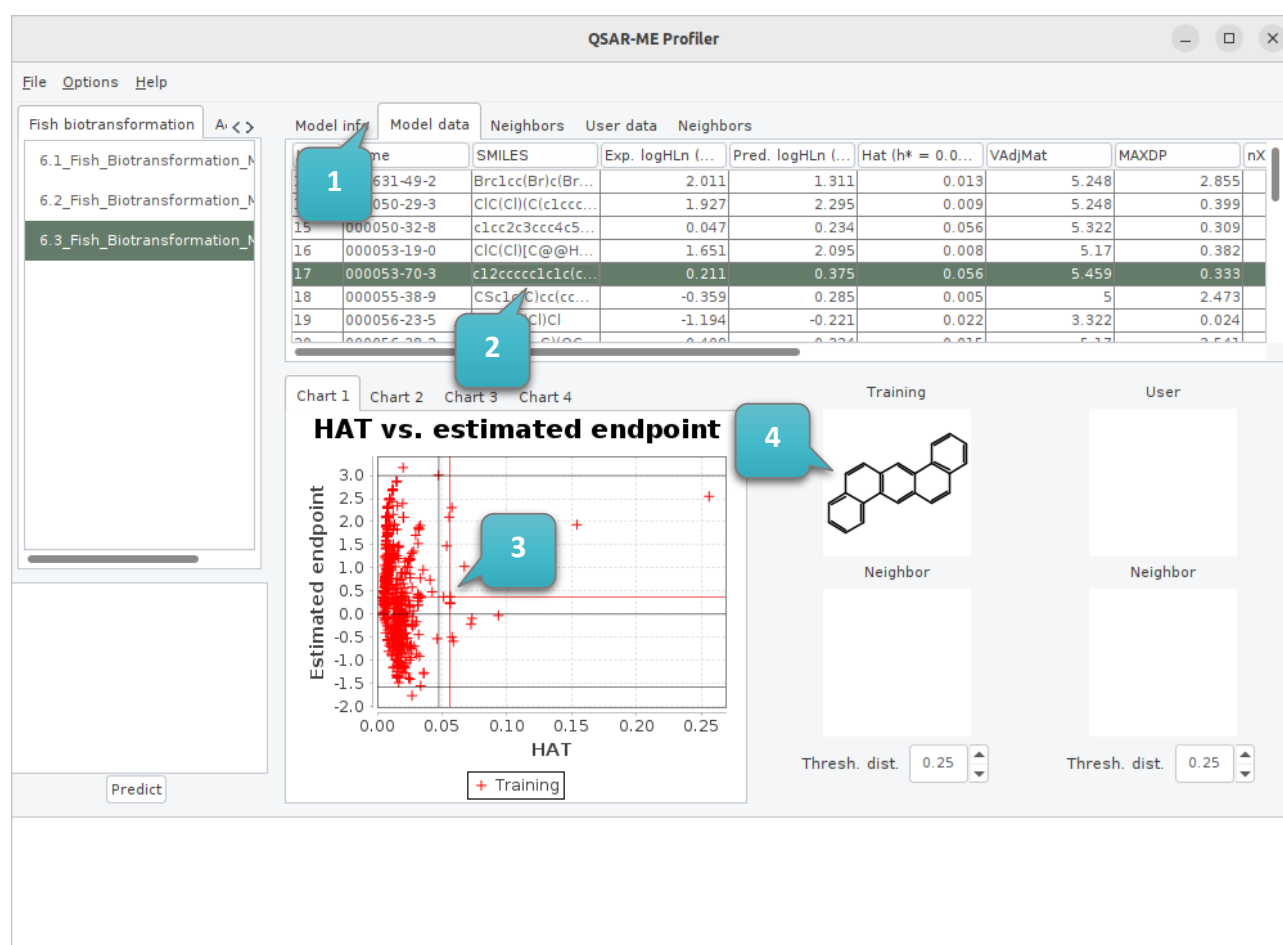


Figure 5. Selection of training set chemicals. Model data tab (1) contains the chemicals' list (2) from which the chemical of interest can be selected. The corresponding chemical will then be targeted in the chart (3) and the structure depicted (4)

Once a chemical is selected, neighbors are automatically detected according to the fingerprint type and the distance measure chosen by the user. Five fingerprints (Pubchem, E-State, Klekota and Roth, Klekota and Roth count, Substructure, Substructure count) and three measures of distance (Tanimoto, Cosine and Dice) are available, and can be selected by the user using menu Options → Similarity → Fingerprint, and Options

⁶ Color can be customized via menu Options → Color → Training

→ Similarity → Distance. Neighbors are likely to change accordingly to these settings and is left to the user to select the most appropriate.

The number of neighbors to be detected can be chosen according to a number or a distance threshold. The type of threshold (number or distance) can be selected by menu Options → Similarity → Threshold while the threshold values can be set using the spinner buttons below the depiction of neighbors, as shown in Figure 6. Starting from the selected chemical, neighbors are sorted from the most to the least similar. In case two or more chemicals share the same value of similarity index, these chemicals count as one.

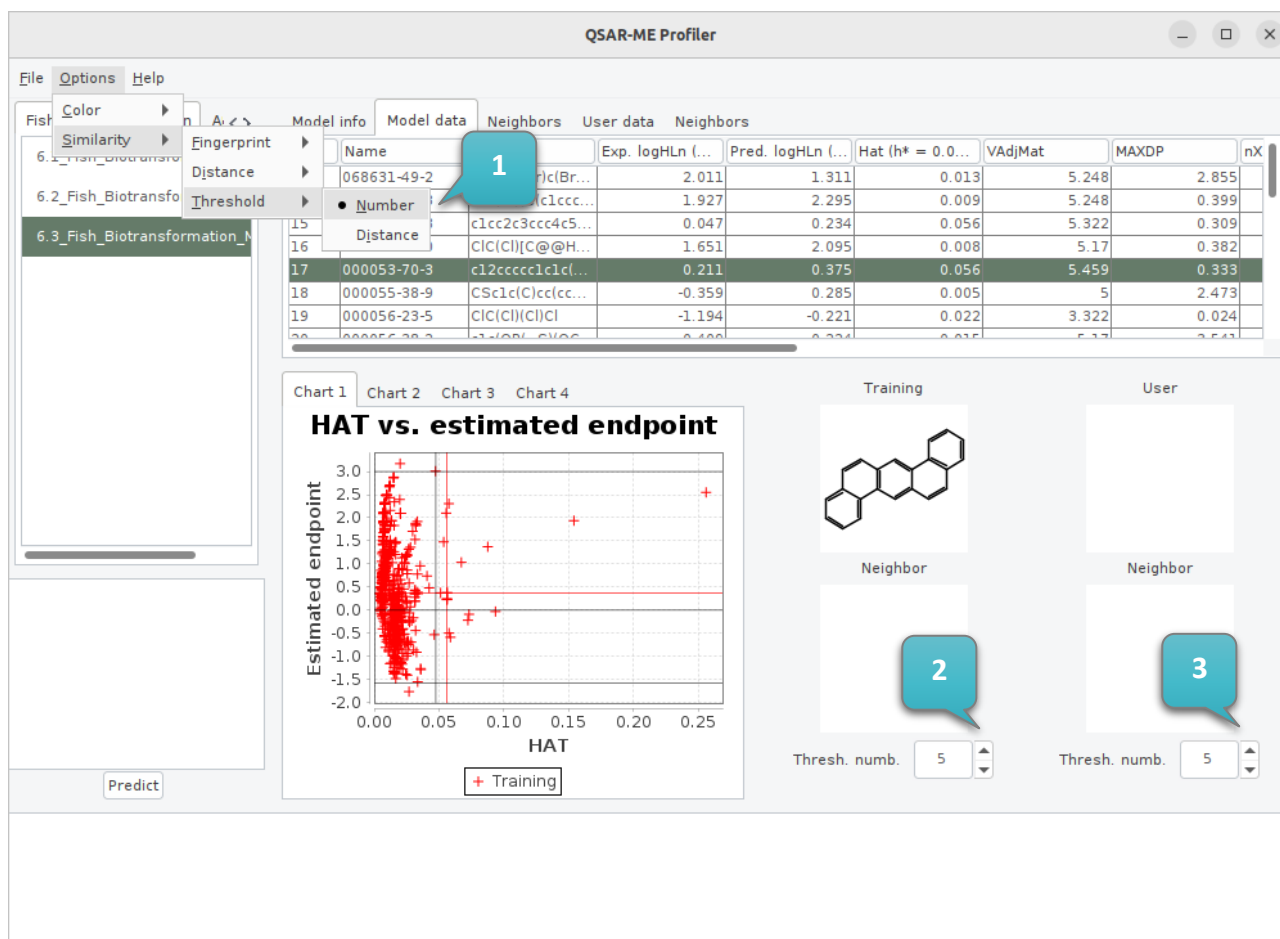


Figure 6. Selection of similarity thresholds. The type of threshold can be selected by the menu option indicated by (1) while the threshold can be set by the spinner arrows indicated by (2) (training set) and the (3)(user entered chemicals)

Neighbors can be selected from the Neighbors tab and then by clicking on one chemical from the list. The corresponding data point on the charts is then targeted⁷ with magenta⁸ crossing lines while the structure of the chemical is shown in the “Neighbor” depiction space under “Training”. In the example of Figure 7 the estimated endpoint is very similar for both the selected training chemical and the selected neighbor, so the red and magenta horizontal lines overlap, while a structural difference has been detected in terms of HAT values, since the vertical lines are separated.

⁷ Targeting on the charts is possible for MLR QSARs charts only.

⁸ Color can be customized via menu Options → Color → Training neighbors

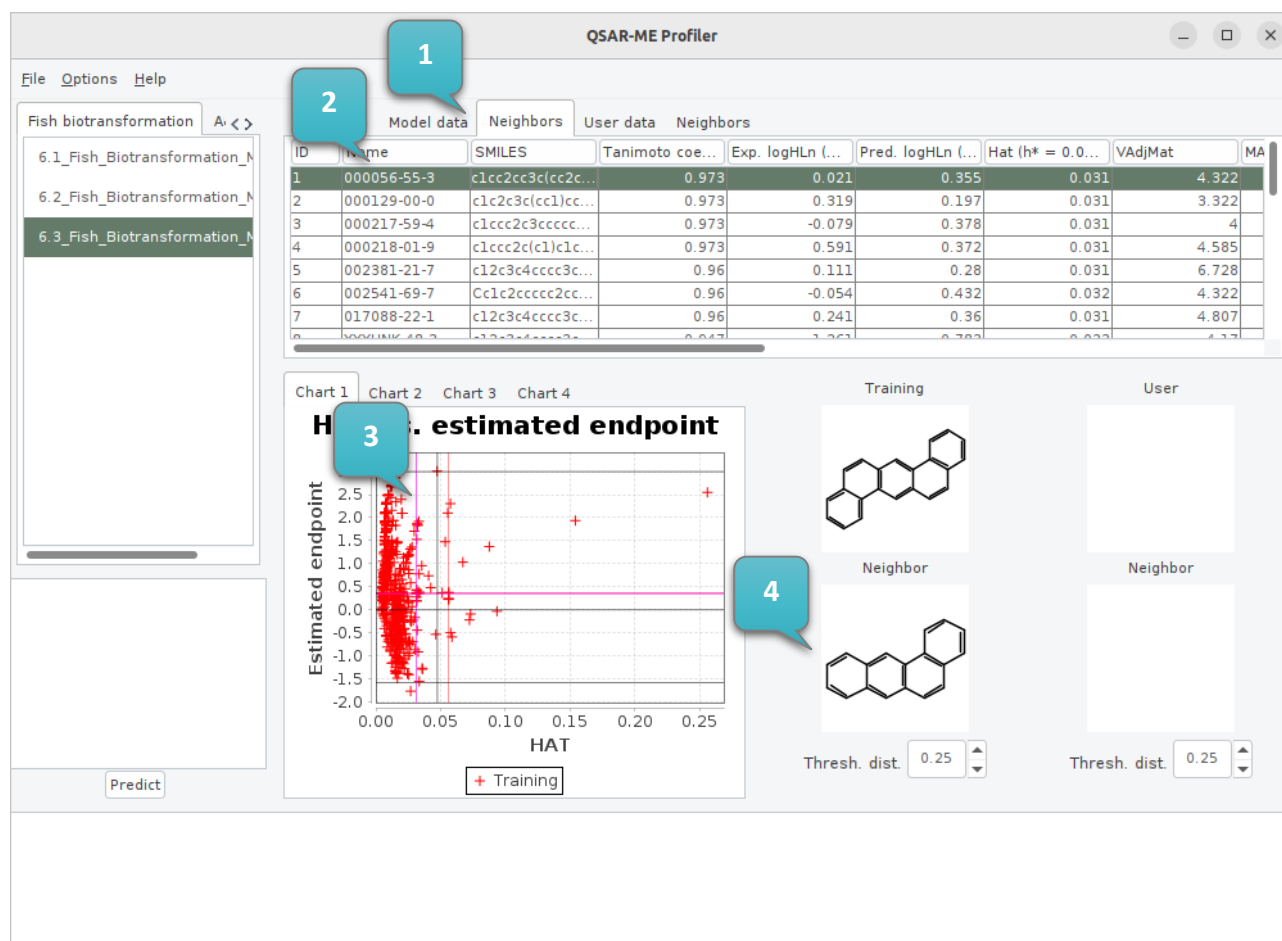


Figure 7. Selection and depiction of chemicals similar (neighbors) to the one selected in the training set. By clicking on the Neighbors tab (1) and then clicking on the chemical of interest (2) the chemical is then targeted on the charts (3)(magenta lines) and the structure depicted (4)

3.3 Predicting the endpoint of new chemicals

For prediction purposes, new chemicals must be entered as SMILES in the text area (see Figure 8) above the “Predict” button.

```

OC(=O)COC1=C(C1)C=C(C1)C=C1      MOL_01
CC1=CC(OCCCC(C)(C)C(=O)=O)=C(C)C=C1      MOL_02
OC1=CC=C(C=C1)C1(OC(=O)C2=C1C=CC=C2)C1=CC=C(O)C=C1      MOL_03
  
```

Predict

Figure 8. Text area for entering SMILES of the chemicals whose endpoint must be predicted

SMILES must be followed by the name of the chemical (in this example MOL_1, MOL_2 and MOL_3), separated by a space or a tabulation. Once entered, you will be prompted by a menu asking whether descriptors must be calculated by the PaDEL-Descriptor software or entered manually, as shown in Figure 9.

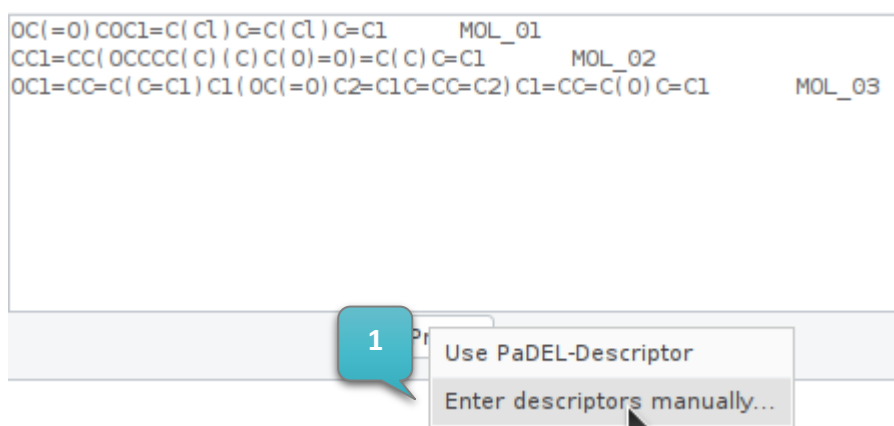


Figure 9. By pressing on the Predict button a menu (1) asks whether descriptors must be calculated automatically by PaDEL-Descriptor or entered manually

In the first case, descriptors will be calculated automatically while in the second you will be asked⁹ to copy and paste or type in them manually. A dialog for entering descriptors is then opened, where molecules names are automatically detected from the prediction text area, while the number of descriptors must be selected manually by using the dedicated spinner arrows, as shown in Figure 10.

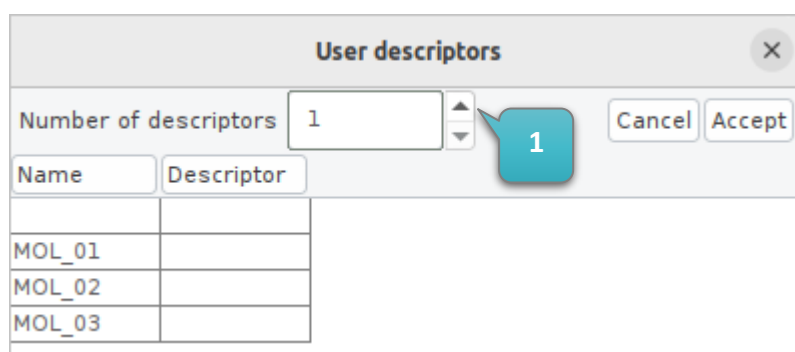


Figure 10. The number of descriptors must be set using the spinners (1)

Available models are automatically selected according to the descriptors provided. For example, if the equation of a QSAR contains the descriptors SsOH and JGI3 and another QSAR equation contains the descriptors SsOH and VE1_D, you must provide all these descriptors, as in Figure 11.

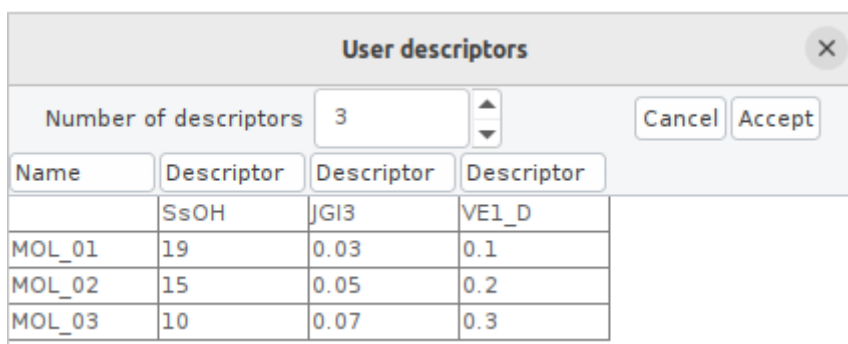


Figure 11. The descriptors of all the QSARs that must be applied must be entered. See text for further details

⁹ This option is for users willing to apply models whose descriptors were calculated using software different from PaDEL-Descriptor.

Once accepted, you will be asked¹⁰ to select the most probable *in vitro* CYP P-450 reaction/s for the entered chemicals. QSAR-ME Profiler then uses this information to select the appropriate models that were developed by creating the training set according to these reactions. As shown in Figure 12, available options are: A) reactions can be selected manually, to be then accepted by pressing “Apply selected”, B) let Toxtree detects the most probable reaction (in this case the manually selected reactions, if any, are not considered).

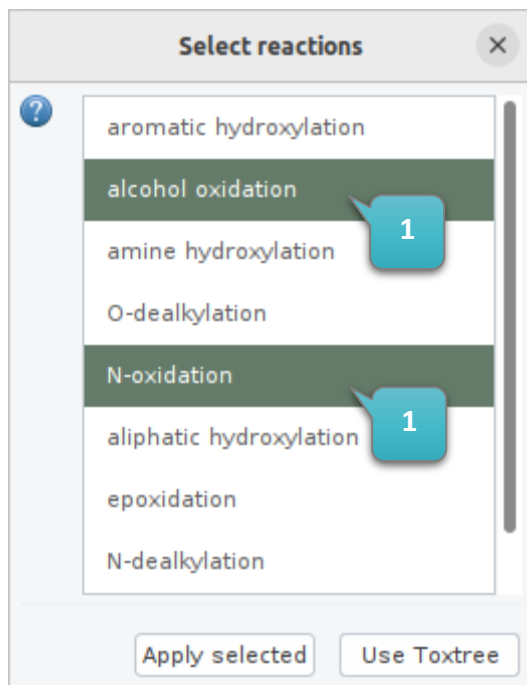


Figure 12. Dialog for the selection (1) of the most probable *in vitro* CYP P-450 reaction/s. “Apply selected” is for using the selected reactions, while “Use Toxtree” is for automatic detection (selected reactions are thus not considered)

3.4 Selecting user-entered chemicals and neighbors

Once the endpoint of the user-entered chemicals is predicted, the “User data” tab is filled with the corresponding chemicals’ details. User-entered chemicals is shown in blue¹¹ (while neighbors are dark cyan¹²) in the HAT vs. Predicted endpoint MLR QSAR chart. Chemicals and neighbors can be selected the same way as in section “Selecting training chemicals and neighbors”, except that the “User data” tab must be selected instead of the “Model data” tab, and the “Neighbor” tab is the one on the right of the “User data” tab. It is here recalled that neighbors are selected from the training set. Figure 13 shows an example of a user-selected chemical and the corresponding neighbor.

¹⁰ Note: if you exclude all models developed by the *in vitro* CYP P-450 by customization of the list of available QSARs (see “How to add or delete a QSAR category” for further details), you will not be asked anymore.

¹¹ Color can be customized via menu Options → Color → User

¹² Color can be customized via menu Options → Color → User neighbors

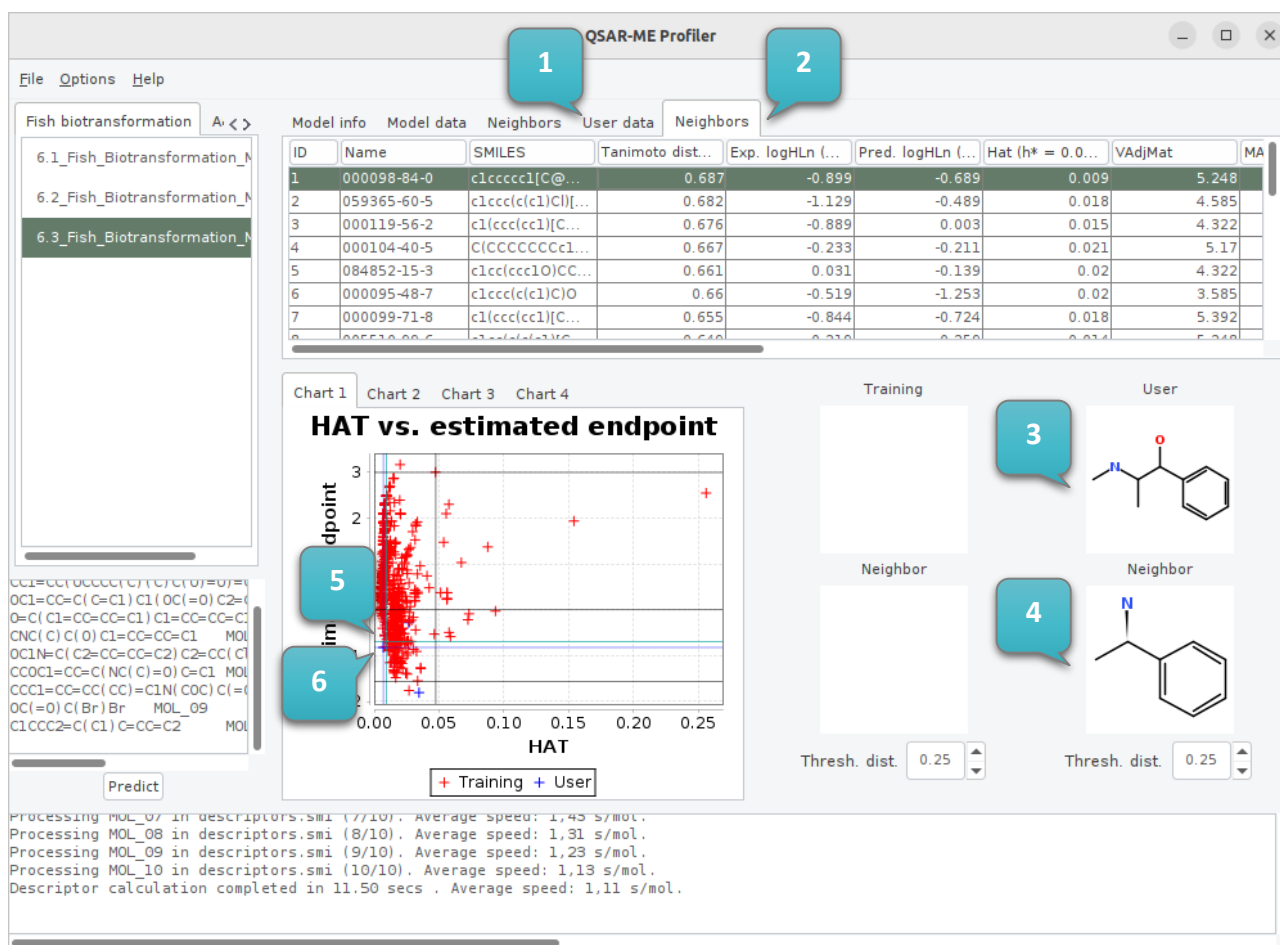


Figure 13. User-entered chemicals can be selected from the “User data” tab (1) while the neighbors from the “Neighbors” tab on its right (2). Corresponding chemicals are depicted (3)(4) and targeted in the graphs (5)(6)

Both training and user-entered chemical, including neighbors, can be displayed simultaneously as shown in Figure 14.

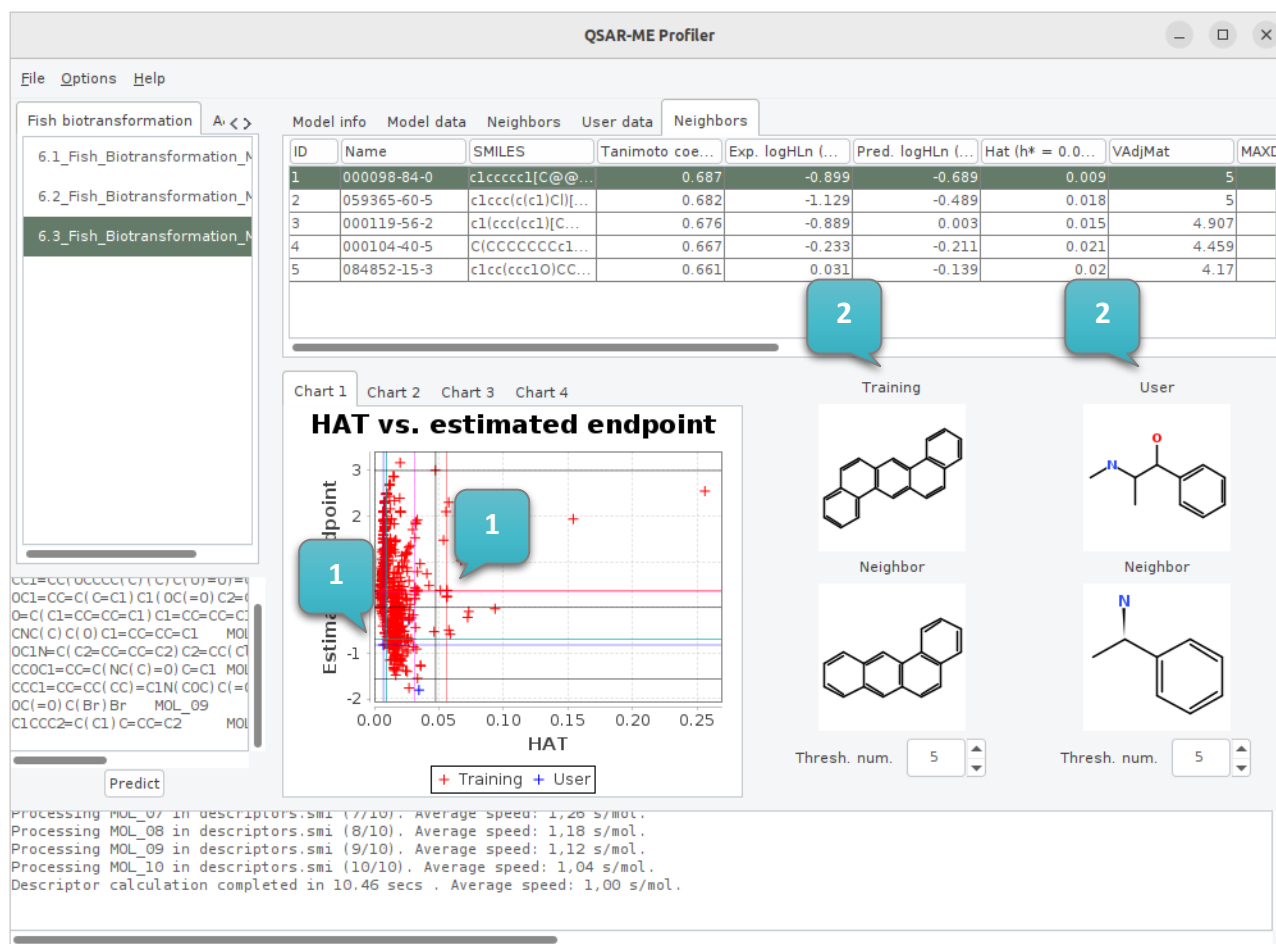


Figure 14. Training and user-entered chemicals, and corresponding neighbors, can be targeted (1) and depicted (2) simultaneously

3.5 Customizing the main window

Size of logical sections of the QSAR-ME Profiler main window can be customized according to your need. While hovering the mouse pointer between boundaries, it turns to a double arrow indicating that the boundary can be moved, as shown in Figure 15.

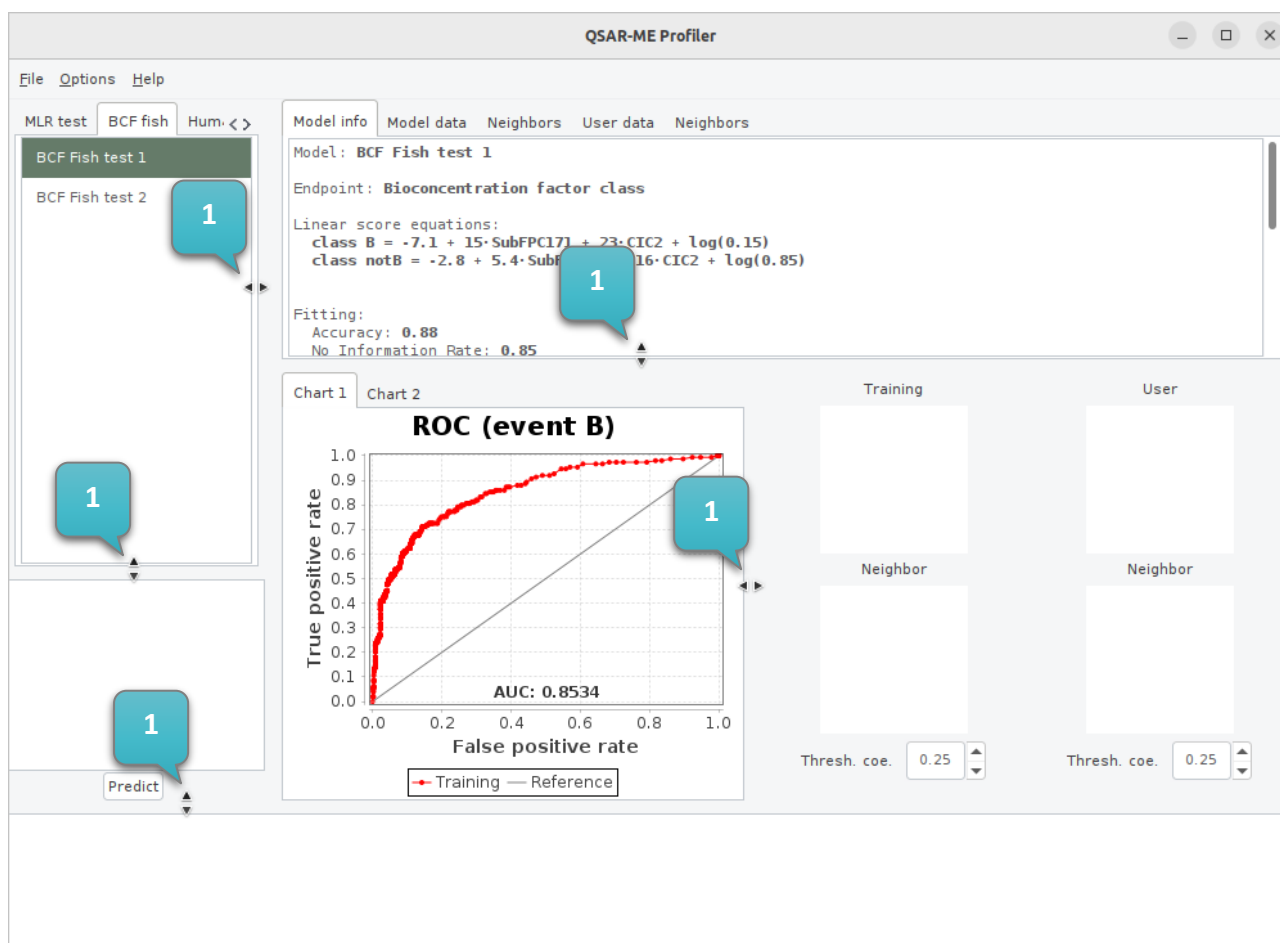


Figure 15. Boundaries (1) can be moved to resize the main window sections

3.6 Available QSAR diagnostic charts

Performances of QSARs and corresponding predictions of user-entered chemicals can be further evaluated by using performances charts, which differ according to the QSAR type.

3.6.1 MLR QSAR graphical diagnostics

To evaluate performances of MLR QSARs, four charts are available in QSAR-ME Profiler. The first concerns HAT vs. estimated endpoint, like the example reported in Figure 17.

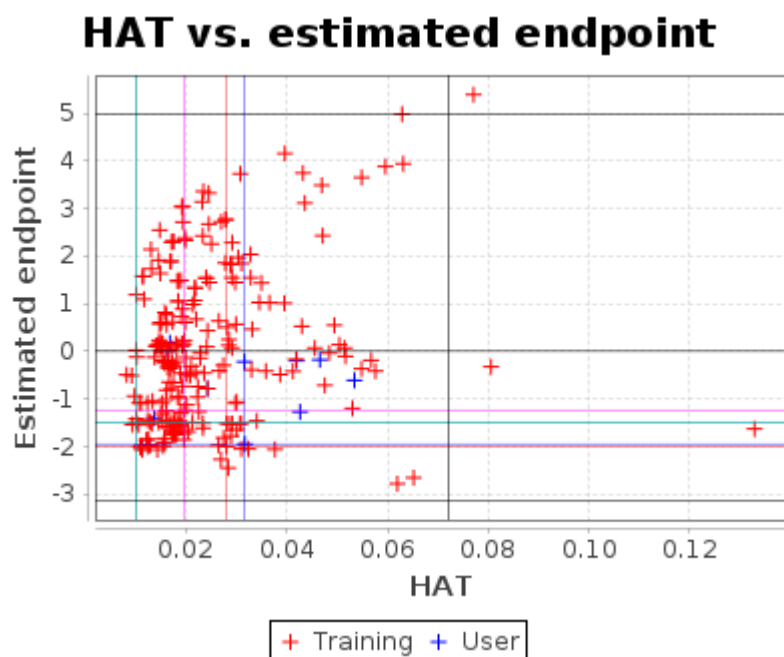


Figure 17. HAT vs. estimated endpoint allows to graphically evaluate whether the user-entered chemicals are outside the applicability domain of the QSARs. Colored lines correspond to the selected chemicals, including neighbors, both for training and user-entered ones

This chart allows to visually check whether the user entered chemicals (here as blue crosses) are within the structural and model endpoint applicability domain. The vertical dark grey line is the threshold HAT value (calculated as $3p' / n$, where p' is the number of the descriptors + 1 and n the number of compounds). User-entered chemicals on the right of this line should be considered as out of the structural domain. Horizontal dark grey lines (above and below 0) correspond to the minimum and maximum training experimental endpoint values. User-entered chemicals with a predicted endpoint value above or below these lines should be considered as extrapolated by the model.

Red and magenta colored crossing lines correspond respectively to a selected training chemical and one of its neighbors (see section “Selecting training chemicals and neighbors” for further details), while blue and dark cyan lines correspond to a selected user-entered chemical and one of its neighbors (see section “Selecting user-entered chemicals and neighbors” for further details).

The following charts allows further evaluation of the selected training chemicals and user-entered chemical neighbors (user-entered chemical themselves cannot be shown because the experimental response is not available).

The Experimental vs. predicted graph allow evaluating the difference between the experimental and the predicted endpoint of the training set chemicals, as shown in Figure 18.

Experimental vs. predicted endpoint (training set)

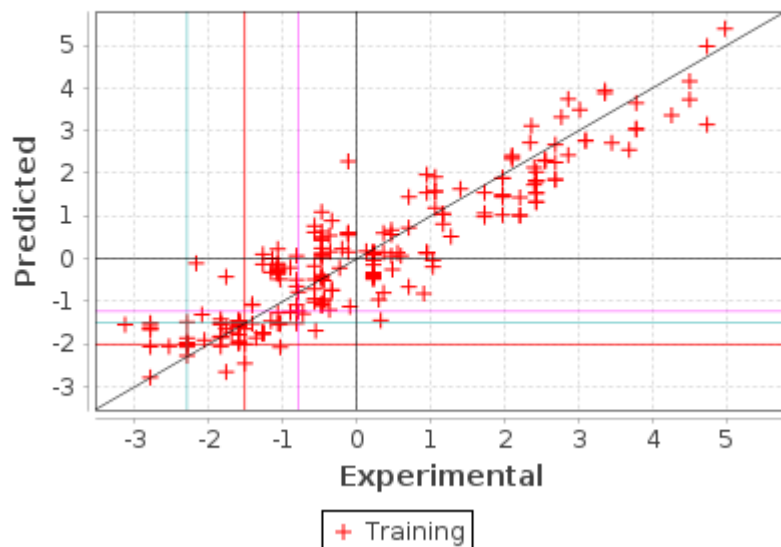


Figure 18. Experimental vs. predicted endpoint (training set) chart example. Colored lines are the selected training chemical and its neighbor, in addition to the neighbor of the chemical selected by the user (the latter cannot be shown here since the experimental value is not available)

The residuals chart allows a similar evaluation but based on the difference between the experimental and the predicted endpoint, as shown in Figure 20.

Endpoint residual (training set)

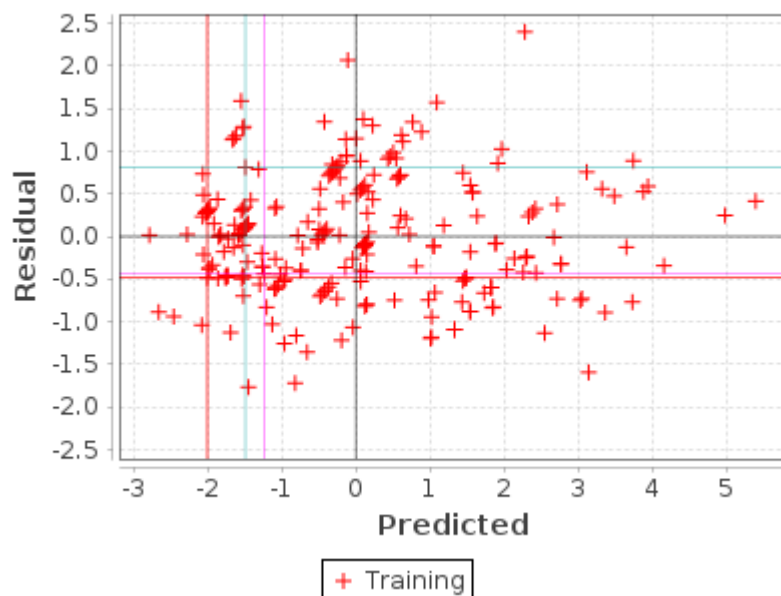


Figure 20. Endpoint residual (training set) chart example. Colored lines are the selected training chemical and its neighbor, in addition to the neighbor of the chemical selected by the user (the latter cannot be shown here since the experimental value is not available)

The latter chart, HAT vs. Standardized residuals (training set), works similarly as HAT vs. estimated endpoint, but uses standardized residuals on the y-axis instead of the predicted endpoint. One example is depicted in Figure 21. Standardized residual thresholds are arbitrarily set to 2.5 (horizontal black lines) so chemicals falling outside should be considered as outliers.

HAT vs. standardized residuals (training set)

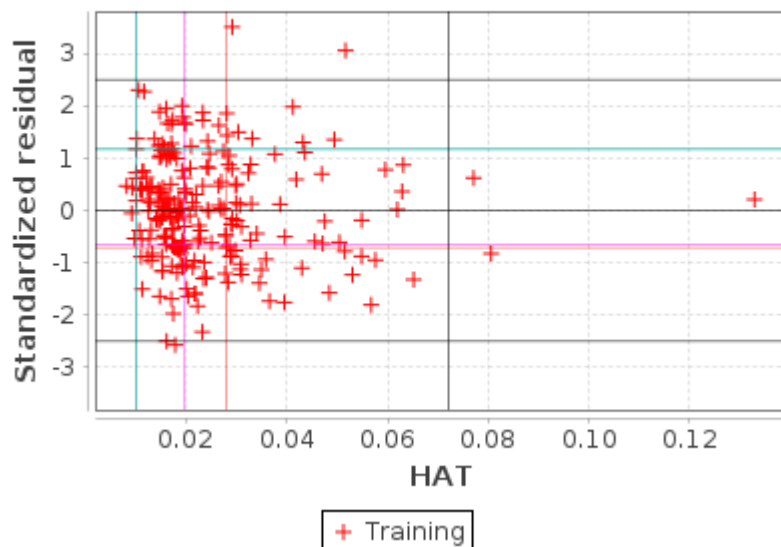


Figure 21. HAT vs. standardized residual (training set) chart example. Colored lines are the selected training chemical and its neighbor, in addition to the neighbor of the chemical selected by the user (the latter cannot be shown here since the experimental value is not available)

3.6.2 LDA QSAR graphical diagnostics

ROC (Receiver Operating Characteristic) charts are available for the evaluation of classification models, like the example shown below in Figure 22 for two classes.

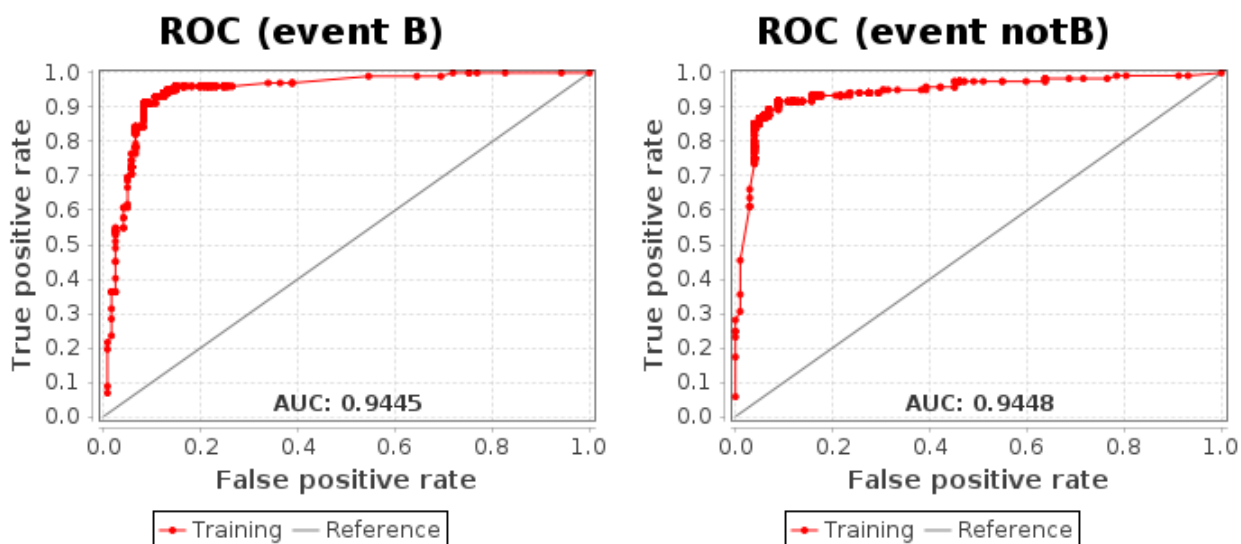


Figure 22. Example of ROC (Receiver Operating Characteristic) charts available in QSAR-ME Profiler

Every class can be considered an event, depending on the users' preference, so a ROC chart is plotted for each. In the example above, a chemical classified as bioaccumulative (B) could be considered as the event of choice but could also be the opposite (not bioaccumulative i.e., notB). The area under the curve (AUC) is a gross measure (between 0 and 1) of the discriminant power of the classification model: the higher the AUC, the better the classification accuracy of the QSAR. In case of more than two classes, ROC charts are based on

a one versus the rest approach, where the class under consideration is the event while the remaining together are considered as the no-event.

Hovering over the dots, hints display the sensitivity and specificity according to the probability threshold compared to the post probability.

3.7 Endpoint prediction reports of user entered chemicals

After pressing the “Predict” button, once calculation is completed, a window containing the predictions of the automatically selected QSARs by QSAR-ME Profiler is reported, like the example in Figure 23.

ID	Name	SMILES	Endpoint	Consensus w...	Consensus w...	4.3_PCP_Fi...	3.1_Fish_A...	4.4_PCP_Fi...
1	MOL_01	OC(=O)COC1=...	pLC50 (mol/L)	4.1±0.81	4.1±0.81	3.7±1.4	4.1±1.3	4.1±1.3
2	MOL_02	CC1=CC(OCCC...	pLC50 (mol/L)	5.0±0.81	5.0±0.81	4.5±1.4	4.8±1.4	4.8±1.4
3	MOL_03	OC1=CC=C(C...	pLC50 (mol/L)	5.2±0.82	5.2±0.82	5.1±1.5	4.4±1.4	4.4±1.4
4	MOL_04	O=C(C1=CC=...	pLC50 (mol/L)	4.9±0.81	4.9±0.81	5.4±1.5	4.6±1.4	4.6±1.3
5	MOL_05	CNC(C)(O)C1...	pLC50 (mol/L)	3.8±0.81	3.8±0.81	4.0±1.4	3.3±1.4	4.0±1.3
6	MOL_06	OC1N=C(C2=C...	pLC50 (mol/L)	4.4±0.81	4.4±0.81	4.7±1.4	4.5±1.4	4.2±1.3
7	MOL_07	CCOC1=CC=C...	pLC50 (mol/L)	4.1±0.81	4.1±0.81	4.0±1.5	3.7±1.4	4.5±1.3
8	MOL_08	CCC1=CC=CC(...	pLC50 (mol/L)	4.7±0.81	4.7±0.81	4.6±1.4	5.4±1.4	4.1±1.3
9	MOL_09	OC(=O)C(Br)Br	pLC50 (mol/L)	2.6±0.82	2.6±0.82	2.2±1.5	3.4±1.4	2.2±1.4

Figure 23. Example of a report of predictions. QSAR categories can be scrolled by the top right arrows (1) and selected by clicking on the tab with the category of interest (2). QSARs in the main window and visualization of QMRF can be accessed by clicking on the down pointing arrow (3)

The left and right arrows on the right of the report allow scrolling the QSAR categories, while the corresponding tabs allow their selection. For each category, columns containing the IDs, names, SMILES and consensus statistics for the user-entered chemicals are shown, then follow columns containing the selected models and the corresponding predictions, which can be recognized by a down-directed arrow on the left of the name of the QSAR. By clicking on the arrow, a menu is shown allowing for A) the selection of the corresponding QSAR in the main QSAR-ME Profiler window and B) the visualization of the corresponding QMRF.

3.7.1 MLR QSAR reports

MLR QSAR reports are organized as in the following example of Figure 24.

ID	Name	SMILES	Endpoint	Consensus weighted ALL	Consensus weighted AD	4.3_PCP_Fi...	3.1_Fish_A...	4.4_PCP_Fi...
1	MOL_01	OC(=O)COC1=...	pLC50 (mol/L)	4.1±0.81	4.1±0.81	3.7±1.4	4.6±1.4	4.1±1.3
2	MOL_02	CC1=CC(OCCC...	pLC50 (mol/L)	5.0±0.81	5.0±0.81	4.5±1.4	5.7±1.4	4.8±1.4
3	MOL_03	OC1=CC=C(C...	pLC50 (mol/L)	5.2±0.82	5.2±0.82	5.1±1.5	6.1±1.4	4.4±1.4
4	MOL_04	O=C(C1=CC=...	pLC50 (mol/L)	4.9±0.81	4.9±0.81	5.4±1.5	4.6±1.4	4.6±1.3
5	MOL_05	CNC(C)(O)C1...	pLC50 (mol/L)	3.8±0.81	3.8±0.81	4.0±1.4	3.3±1.4	4.0±1.3
6	MOL_06	OC1N=C(C2=C...	pLC50 (mol/L)	4.4±0.81	4.4±0.81	4.7±1.4	4.5±1.4	4.2±1.3
7	MOL_07	CCOC1=CC=C...	pLC50 (mol/L)	4.1±0.81	4.1±0.81	4.0±1.5	3.7±1.4	4.5±1.3
8	MOL_08	CCC1=CC=CC(...	pLC50 (mol/L)	4.7±0.81	4.7±0.81	4.6±1.4	5.4±1.4	4.1±1.3
9	MOL_09	OC(=O)C(Br)Br	pLC50 (mol/L)	2.6±0.82	2.6±0.82	2.2±1.5	3.4±1.4	2.2±1.4
10	MOL_10	C1CCC2=C(C1...	pLC50 (mol/L)	4.6±0.81	4.6±0.81	5.2±1.5	4.4±1.4	4.3±1.4

Figure 24. Example of a report of predictions for multiple linear regression (MLR) QSAR. See text for further details

The first columns contain the IDs, the names, the SMILES and the name of the endpoint of the user-entered chemicals. The “Consensus weighted” columns contain the combined predictions¹³ of the QSARs, which can be found after the consensus columns. “Consensus weighted ALL” means that all predictions are used in the calculation while “Consensus weighted AD” means that only the prediction within the structural and endpoint domain are used. Concerning single predictions, uncertainties¹⁴ and applicability domain warnings (if any) are reported using “*” if the prediction is outside the experimental endpoint domain and “#” if the chemical is outside the structural domain¹⁵.

3.7.2 MLR models for the prediction of biotransformation *in vitro*

To collect chemicals according to the most likely reactions mediated by cytochrome P-450, the SMARTCyp module embedded in Toxtree software¹⁶ was applied before the development of the QSARs, shipped with QSAR-ME Profiler, to predict the hepatic *in vitro* intrinsic clearance in human, rat and mouse. This procedure allowed to develop mechanistic QSARs based on metabolic reactions. Likelihood of reactions are ranked¹⁷ in descending order by Toxtree, so the report is organized accordingly by the detected rank and reaction, as shown in the example of Figure 25.

ID	Name	SMILES	Endpoint	Consensus weighted ALL	Consensus weighted AD	08.39_InVi...	08.42_InVi...	08.43_InVi...
1	MOL_01	OC(=O)COC1=...	Log_CL in vitro...	4.0e-02±0.19	-0.12±0.24	-0.44±0.51 *	2.3±0.85 #	-0.62±0.82 #
2	MOL_02	CC1=CC(OCCC...	Log_CL in vitro...	0.45±0.19	0.39±0.31	0.18±0.53 #	1.0±0.72	0.39±0.82 #
3	MOL_07	CCOC1=CC=C...	Log_CL in vitro...	0.39±0.20	8.8e-02±0.27	0.41±0.65 #	0.63±0.73	-0.35±0.89 #
4	MOL_08	CCC1=CC=CC(...	Log_CL in vitro...	0.37±0.19	0.54±0.31	0.40±0.61 #	1.3±0.73	0.39±0.73 #

Figure 25. Example of a report of predictions for multiple linear regression (MLR) QSAR developed according to cytochrome mediated reactions. Predictions are collected according to the Toxtree rank (1) and detected reactions (2)

3.7.3 LDA QSAR Reports

LDA QSAR reports are organized similarly to MLR QSARs, as shown in Figure 26. Consensus endpoint can be either the most represented class or a tie. The consensus is calculated using all models, while consensus

¹³ Weighted averages and uncertainties of combined predictions are calculated according to the Italian edition of “An Introduction to Error Analysis, The Study of Uncertainties in Physical Measurements”, Taylor J.R., University Science Books, 1982.

¹⁴ Uncertainties for new chemical unseen in model development were calculated according to Julian J. Faraway, Practical regression and anova using R, 2002, cran.r-project book contribution.

¹⁵ A chemical is considered outside the structural domain if the HAT value is above $3p'/n$ where p' is the number of descriptors + 1 used by the model and n is the number of used chemicals.

¹⁶ Toxtree v. 3.1.0; Patlewicz G. et al. SAR QSAR Environ Res. 2008; Rydberg, P. et al. ACS Med. Chem. Lett. 2010; Rydberg, P. et al. Bioinformatics 2010.

¹⁷ For manually selected reactions, as explained in section “Predicting the endpoint of new chemicals”, QSAR-ME Profiler assigns rank 1 as default.

endpoint filtered by applicability domain is currently under study and is planned to be implemented in a future release of QSAR-ME Profiler.

ID	Name	SMILES	Endpoint	Consensus	BCF Fish t...	BCF Fish t...
1	MOL_01	OC(=O)COC1=...	Bioconcentrati...	tie	B	notB
2	MOL_02	CC1=CC(OCCC...	Bioconcentrati...	tie	B	notB
3	MOL_03	OC1=CC=C(C...	Bioconcentrati...	tie	B	notB
4	MOL_04	O=C(C1=CC=...	Bioconcentrati...	tie	B	notB
5	MOL_05	CNC(C)C(O)C1...	Bioconcentrati...	tie	B	notB
6	MOL_06	OC1N=C(C2=C...	Bioconcentrati...	tie	B	notB
7	MOL_07	CCOC1=CC=C...	Bioconcentrati...	notB	notB	notB
8	MOL_08	CCC1=CC=CC...	Bioconcentrati...	tie	B	notB
9	MOL_09	OC(=O)C(Br)Br	Bioconcentrati...	notB	notB	notB

Figure 26. Example of a prediction report for multiple linear discriminant analysis (LDA) QSAR

4 How to customize QSAR-ME Profiler available QSARs

To customize available QSARs you must edit¹⁸ XML (Extensible Markup Language) and CSV (Comma-Separated Values) files with a text editor of your choice¹⁹. Concerning XML files it is recommended not using special characters, like &, < etc. which may conflict with the xml reader of QSAR-ME Profiler, while it is advised using simple alphanumeric characters and _ for separating words (if needed). Concerning the text editor, it is advised to disable text wrapping, to avoid confusion when editing lines. As a general comment, **it is here strongly advised to modify only the parts suggested in this guide, leaving the others untouched**. In addition, QSAR-ME Profiler checks configuration and QSARs files for consistency, warning the user in case of problems, but this can be accomplished only to a certain extend: **It is thus also advised to carefully check files before running QSAR-ME profiler**.

4.1 QSAR categories

A QSAR category in QSAR-ME Profiler is a collection of QSARs, sharing the same endpoint, which represents coherent groups like, for example, bioconcentration factor in fish and carbon-water partition coefficient. The **qsar_layout.xml** file, located in the config folder within the QSAR-ME Profiler one, allows for the manipulation of the categories, and looks like the following example:

```
<?xml version="1.0" encoding="UTF-8"?>
<model_gui>
  <tab name="Cat 1" folder="cat_mlr_1" hint="MLR Category 1 hint" type="mlr" toxtree="yes"/>
  <tab name="Cat 2" folder="cat_mlr_2" hint="MLR Category 2 hint" type="mlr" toxtree="no" />
  <tab name="Cat 1" folder="cat_lda_1" hint="LDA Category 1 hint" type="lda" toxtree="no" />
</model_gui>
```

A QSAR category begins with `<tab` and ends with `/>`, everything in between specifies the QSAR category. **name** is the name of the QSAR category, specified by the format: `name="Name of the category"`. The name between quotes will be displayed in the QSAR-ME Profiler GUI.

folder is the name of the folder containing the QSARs, specified by the format: `folder="name_of_the_folder"` (underscores are used instead of spaces for convenience). This folder must be in the qsar folder within the main QSAR-ME Profiler folder.

¹⁸ Before editing, remember to close QSAR-ME Profiler if running.

¹⁹ It is suggested to use an editor for text only.

hint specifies the hint which appear on the GUI while hovering over the QSAR category with the mouse's cursor.

type specifies the type of QSAR model, which can be either **type="mlr"** (multiple linear regression) or **type="lda"** (linear discriminant analysis).

toxtree specifies whether the QSARs requires the use of Toxtree (**toxtree="yes"**) or not (**toxtree="no"**).

4.1.1 How to add or delete a QSAR category

To delete a QSAR category suffices to delete the corresponding line from the configuration file. For example, let us assume that "MLR Category 2" should not be handled by QSAR-ME Profiler²⁰: just remove the line containing the unwanted category by editing **qsar_layout.xml**. In this case the file content would be like:

```
<?xml version="1.0" encoding="UTF-8"?>
<model_gui>
  <tab name="Cat 1" folder="cat_mlr_1" hint="MLR Category 1 hint" type="mlr" toxtree="yes"/>
  <tab name="Cat 1" folder="cat_lda_1" hint="LDA Category 1 hint" type="lda" toxtree="no" />
</model_gui>
```

To see the effect of the reduction of available categories, save **qsar_layout.xml** and run QSAR-ME Profiler.

To add an existing QSAR category, edit **qsar_layout.xml** adding a line containing the category. The easiest scenario is just reversing the steps of deleting the pre-existing category, as in the above example, while adding a new category from scratch needs writing the .xml and .csv files of the new QSARs (the .pdf QMRF files are optional), as explained in the following sections.

4.1.2 How to create a QSAR category

To create a QSAR category you need to:

- 1) Locate the **qsar** folder within the main folder of QSAR-ME Profiler
- 2) Create a folder which will contain the new category's QSARs files
- 3) Create the .xml and .csv QSARs files
- 4) Optionally create the .pdf QMRFs files

Steps 1 and 2 are straightforward so let us assume that a **new_cat** folder is created. Let us also assume that this is an MLR QSAR category, called "New cat" whose QSARs, hinted in QSAR-ME Profiler GUI as "New Category hint", should use Toxtree. The corresponding **qsar_layout.xml** file would look like the following example, where the new script line is evidenced by an arrow:



```
<?xml version="1.0" encoding="UTF-8"?>
<model_gui>
  <tab name="Cat 1" folder="cat_mlr_1" hint="MLR Category 1 hint" type="mlr" toxtree="yes"/>
  <tab name="Cat 2" folder="cat_mlr_2" hint="MLR Category 2 hint" type="mlr" toxtree="no" />
  <tab name="Cat 1" folder="cat_lda_1" hint="LDA Category 1 hint" type="lda" toxtree="no" />
  <tab name="New cat" folder="new_cat" hint="New Category hint" type="mlr" toxtree="yes" />
</model_gui>
```

As a final note, the order of the categories (i.e., the lines beginning from **<tab** and ending with **/>**), is irrelevant.

²⁰ The category folder and the QSARs within are not deleted from the disk, so the operation is reversible.

4.2 How to create QSAR files

QSAR-ME Profiler loads QSARs as XML files (which, basically, describe the model) and corresponding CSV files (which contain the training set chemicals data). Before starting, remember that for each QSAR **the .xml and .csv file name must be the same**, like new_qsar.xml and new_qsar.csv. To simplify the editing of the .xml files, which are somewhat complex, the following templates are provided in the help folder located in the main folder of QSAR-ME Profile.

mlr_template.xml - template for multiple linear regression QSARs

mlr_toxtree_template.xml - template for multiple linear regression QSARs using Toxtree

lda_template.xml - template for linear discriminant analysis QSARs

4.2.1 How to create MLR QSAR XML and CSV files

We here focus on how to compile a Multiple Linear Analysis (MLR) QSAR to be loaded by QSAR-ME Profiler. Let us assume we would like to add a QSAR called “New QSAR” to a QSAR category called “New category” (see How to create a QSAR category for further details).

Before starting, it is here recalled that in even though the order of the items (the lines starting from `<item` and ending to `/>`) within the XML sections (e.g., `<equation type="normal"> ... </equation>`) is irrelevant, it is advised to keep them in the same order for the sake of readability. It is here also suggested to look to the .xml and .csv files of QSARs shipped with QSAR-ME Profiler (which are in the qsar folder, within the QSAR-ME Profiler one. Examples reported in this text are simplified fictional QSARs, to avoid cluttering the tutorials).

Here it follows a workflow for the creation and the editing of the XML and CSV files, where each step is marked as a dot •

- copy the `mlr_template.xml` file and rename it as `new_qsar.xml`, then open it in a text editor (e.g., Windows Notepad). Once opened, locate the `header` section, then locate the item containing `type="name"` and write the name of the QSAR between the quotes (see arrow in the following example) of `value`.



```
<header>
  <item type="name" value="New QSAR"/>
</header>
```



- Locate the `description` section and write the description of the QSAR between `CDATA[` and `]`, without quotes²¹, as indicated by the top arrow in the example below, and then do the same for the description of the endpoint (`endpoint` section), as indicated by the arrows in the example below.

```
<description>
  <![CDATA[Description of the new QSAR]]>
</description>

<endpoint>
  <![CDATA[pEC50 (mol/L)]]>
</endpoint>
```

²¹ The content between `CDATA[` and `]` is read literally, so you can use characters without restriction, excluding the `]]>` sequence.

- Let now assume you developed (or took from the literature) an MLR QSAR with the following equation (the example is fictional):

$$\text{pEC}_{50} (\text{mol/L}) = 0.5473 - 0.1606 \times \text{C1SP2} + 0.1434 \times \text{minHBint4}$$

Locate the `equation` section `type="normal"`, then locate the `type="intercept"` item and write the intercept value²² between the `value` quotes (upper arrow, in the example below). Concerning the descriptors, locate the `type="coefficient"` item, then write the descriptors' names between the `name` quotes (lower left arrow in the example below) and the value of the coefficient between the `value` quotes (lower right arrow in the example below).



```
<equation type="normal">
  <item type="intercept" value="0.5473"/>
  <item type="coefficient" name="C1SP2" value="-0.1606"/>
  <item type="coefficient" name="minHBint4" value="0.1434"/>
</equation>
```

Then locate the `equation` section `type="standardized"` and write the values of the standardized coefficients the same way as the normal coefficients, below follows an example.

```
<equation type="standardized">
  <item type="coefficient" name="C1SP2" value="-0.7053"/>
  <item type="coefficient" name="minHBint4" value="0.6089"/>
</equation>
```

- Statistics concerning the intercept and the coefficients are in the `statistic` sections. Three types of statistics must be provided: `confidence`, `significance` (p-value) and `standard_error`, which must be compiled the same way as for the equation sections. Below follows an example.

```
<statistic type="confidence">
  <item type="intercept" value="0.2280"/>
  <item type="coefficient" name="C1SP2" value="0.06261"/>
  <item type="coefficient" name="minHBint4" value="0.06472"/>
</statistic>

<statistic type="significance">
  <item type="intercept" value="0.00009098"/>
  <item type="coefficient" name="C1SP2" value="0.00004420"/>
  <item type="coefficient" name="minHBint4" value="0.0002081"/>
</statistic>

<statistic type="standard_error">
  <item type="intercept" value="0.1075"/>
  <item type="coefficient" name="C1SP2" value="0.02953"/>
  <item type="coefficient" name="minHBint4" value="0.03053"/>
</statistic>
```

- Fitting performances are located in the `performance` section, `type="fitting"`. Two items are obligatory because involved in calculations: `<item type="value" name="s">` and `<item type="value" name="h*">`, where "s" is the standard error of the estimate (i.e., the square root of the squared residuals mean) and "h*" is the leverage cut-off value (calculated as $3p' / n$, where p' is the number of the descriptors + 1 and n the number of compounds). Other items are optional, like for example `<item type="value" name="R2">`, but it is strongly

²² Numerical values in the QSAR XML files must conform to 64 bits double precision format IEEE 754. This means that a string with up to 15 significant digits and an approximate range from 4.9×10^{-324} to 1.80×10^{308} can be accepted.

advised to add them whatever can help better understand the QSAR's performances. Names and values of optional items can be any since they are not used for calculations, but are only displayed as model's information, while "s" and "h*" cannot be changed. Below follows an example.

```
<performance type="fitting">
  <item type="value" name="R2" value="0.7372"/>
  <item type="value" name="R2adj" value="0.7044"/>
  <item type="value" name="RMSE" value="0.1858"/>
  <item type="value" name="MAE" value="0.1470"/>
  <item type="value" name="CCC" value="0.8487"/>
  <item type="value" name="F" value="22.44"/>
  <item type="value" name="s" value="0.2025"/>
  <item type="value" name="h*" value="0.4737"/>
</performance>
```

- Cross validation performances are optional (even though is strongly advised to provide some for the user's convenience) and are located in the `performance` section, `type="cross_validation"`. Names and values of cross validation items can be any, since are not involved in calculations but are only displayed as information. Below follows an example.

```
<performance type="cross_validation">
  <item type="value" name="Q2L00" value="0.658"/>
  <item type="value" name="RMSE" value="0.1736"/>
  <item type="value" name="CCC" value="0.7973"/>
</performance>
```

- The `dataset` section tells QSAR-ME Profiler where to look for the training set values (while reading this part it is suggested to look at Figure 27 for reference. The upper part of the example concerns the XML file while the lower part the training dataset as it would be shown in a spreadsheet software like Excel or OpenOffice. Numberer circles helps in finding the correspondences between the dataset items and the spreadsheet columns.)

The `value` of `<item type="file_name">` is the name of the CSV file containing the dataset, in this example is `new_qsar.csv` (for coherence, the name excluding the extension corresponds to XML file `new_qsar.xml`). The `value` of `<item type="object">` corresponds to the column name containing the names of the chemicals while the `value` of `<item type="smiles">` corresponds to the column name containing the chemical's SMILES. The values of `<item type="exp_endpoint">`, `<item type="pred_endpoint">`, `<item type="residual">` and `<item type="std_residual">` corresponds respectively to the column's names of the experimental and predicted endpoints followed by the normal and standardized endpoint's residuals. The `value` of `<item type="hat">` is the name of the column containing the chemical's hat (leverage) values. The remaining `value` of the `<item type="descriptor">` items are the names of the columns containing the QSAR's descriptors.

Additional columns in the CSV files containing data not referenced in the `dataset` section will be ignored, thus can be left in place if you prefer to do so.

```
<dataset>
  <item type="file_name" value="new_qsar.csv"/>
  <item type="object" value="Name"/> ①
  <item type="smiles" value="SMILES"/> ②
  <item type="exp_endpoint" value="Exp. endpoint"/> ③
  <item type="pred_endpoint" value="Pred. endpoint"/> ④
  <item type="residual" value="Residual"/> ⑤
  <item type="std_residual" value="Std. Residual"/> ⑥
  <item type="hat" value="Hat"/> ⑦
  <item type="descriptor" value="C1SP2"/> ⑧
  <item type="descriptor" value="minHBint4"/> ⑨
</dataset>
```

1

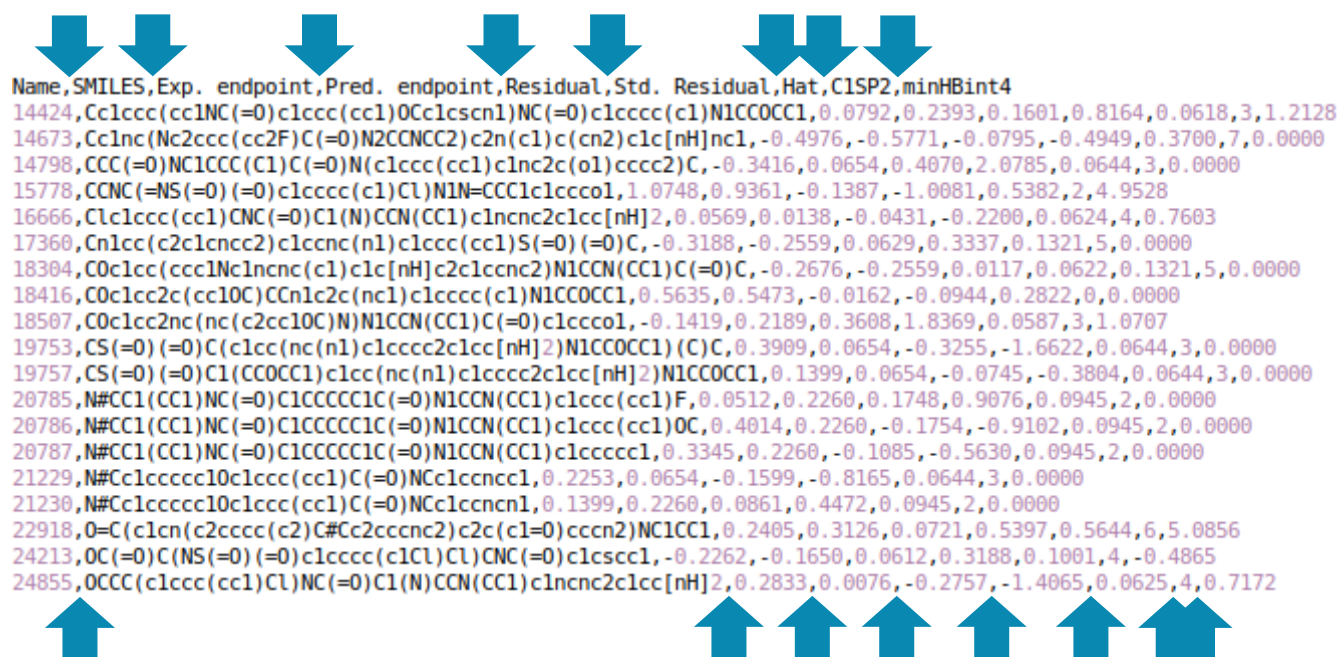
2

①	②	③	④	⑤	⑥	⑦	⑧	⑨
Name	SMILES	Exp. endpoint	Pred. endpoint	Residual	Std. Residual	Hat	C1SP2	minHBint4
14424	Cc1ccc(cc1)N	0.0792	0.2393	0.1601	0.8164	0.0618	3	1.2128
14673	Cc1nc(Nc2cc	-0.4976	-0.5771	-0.0795	-0.4949	0.3700	7	0.0000
14798	CCC(=O)NC>	-0.3416	0.0654	0.4070	2.0785	0.0644	3	0.0000
15778	CCNC(=NS(→	1.0748	0.9361	-0.1387	-1.0081	0.5382	2	4.9528
16666	Clc1ccc(cc1)>	0.0569	0.0138	-0.0431	-0.2200	0.0624	4	0.7603
17360	Cn1cc(c2c1c>	-0.3188	-0.2559	0.0629	0.3337	0.1321	5	0.0000
18304	COc1cc(ccc1>	-0.2676	-0.2559	0.0117	0.0622	0.1321	5	0.0000
18416	COc1cc2c(cc>	0.5635	0.5473	-0.0162	-0.0944	0.2822	0	0.0000
18507	COc1cc2nc(r>	-0.1419	0.2189	0.3608	1.8369	0.0587	3	1.0707
19753	CS(=O)(=O)O>	0.3909	0.0654	-0.3255	-1.6622	0.0644	3	0.0000
19757	CS(=O)(=O)O>	0.1399	0.0654	-0.0745	-0.3804	0.0644	3	0.0000
20785	N#CC1(CC1>	0.0512	0.2260	0.1748	0.9076	0.0945	2	0.0000
20786	N#CC1(CC1>	0.4014	0.2260	-0.1754	-0.9102	0.0945	2	0.0000
20787	N#CC1(CC1>	0.3345	0.2260	-0.1085	-0.5630	0.0945	2	0.0000
21229	N#Cc1ccccc>	0.2253	0.0654	-0.1599	-0.8165	0.0644	3	0.0000
21230	N#Cc1ccccc>	0.1399	0.2260	0.0861	0.4472	0.0945	2	0.0000
22918	O=C(c1cn(c2>	0.2405	0.3126	0.0721	0.5397	0.5644	6	5.0856
24213	OC(=O)C(NS>	-0.2262	-0.1650	0.0612	0.3188	0.1001	4	-0.4865
24855	OCCC(c1ccc>	0.2833	0.0076	-0.2757	-1.4065	0.0625	4	0.7172

Figure 27. The script (1) is an example of the dataset part of the QSAR XML file, which tells QSAR-ME Profile the name of the CSV dataset (2) file containing the training data and how to read it (circled numbers in red show the correspondences between the XML items and the CSV table data). See main text for further details

The above table in the example concerns data display in a spreadsheet software. It is here recalled that a CSV file is a text table whose fields are separated by a comma. Figure 28 shows, for clarity, the same table above as a text file (some commas are evidenced by arrows). In case fields contain commas, like for example the chemical name 2,4-Diaminotoluene, the field must be embraced by quotes²³ (e.g., "2,4-Diaminotoluene"), otherwise QSAR-ME Profiler would confuse these commas as field separators.

²³ Using quotes for fields containing commas is a common procedure usually followed by spreadsheet software when saving tables as CSV files.



Name	SMILES	Exp. endpoint	Pred. endpoint	Residual	Std. Residual	Hat	CISP2	minHBint4
14424	Cc1ccc(cc1NC(=O)c1ccc(cc1)OCc1cscn1)NC(=O)c1cccc(c1)N1CCOCC1	0.0792	0.2393	0.1601	0.8164	0.0618	3	1.2128
14673	Cc1nc(Nc2ccc(cc2F)C(=O)N2CCNCC2)c2n(c1)c(c2)c1c[nH]nc1	-0.4976	-0.5771	-0.0795	-0.4949	0.3700	7	0.0000
14798	CCC(=O)NC1CCC(C1)C(=O)N(c1ccc(cc1)c1nc2c(o1)cccc2)C	-0.3416	0.0654	0.4070	2.0785	0.0644	3	0.0000
15778	CCNC(=NS(=O)(=O)c1cccc(c1)Cl)N1N=CCc1c1cccc1	1.0748	0.9361	-0.1387	-1.0081	0.5382	2	4.9528
16666	C1c1ccc(cc1)CNC(=O)C1(N)CCN(CC1)c1ncnc2c1cc[nH]2	0.0569	0.0138	-0.0431	-0.2200	0.0624	4	0.7603
17360	Cn1cc(c2c1cnc2)c1ccnc(n1)c1ccc(cc1)S(=O)(=O)C	-0.3188	-0.2559	0.0629	0.3337	0.1321	5	0.0000
18304	C0c1cc(ccc1Nc1ncnc(c1)c1c[nH]c2c1cnc2)N1CCN(CC1)C(=O)C	-0.2676	-0.2559	0.0117	0.0622	0.1321	5	0.0000
18416	C0c1cc2c(cc1OC)CCn1c2c(nc1)c1cccc(c1)N1CCOCC1	0.5635	0.5473	-0.0162	-0.0944	0.2822	0	0.0000
18507	C0c1cc2nc(nc(c2cc1OC)N)N1CCN(CC1)C(=O)c1cccc1	-0.1419	0.2189	0.3608	1.8369	0.0587	3	1.0707
19753	CS(=O)(=O)C(c1cc(nc(n1)c1cccc2c1cc[nH]2)N1CCOCC1)C(C)C	0.3909	0.0654	-0.3255	-1.6622	0.0644	3	0.0000
19757	CS(=O)(=O)C1(CC0CC1)c1cc(nc(n1)c1cccc2c1cc[nH]2)N1CCOCC1	0.1399	0.0654	-0.0745	-0.3804	0.0644	3	0.0000
20785	N#CC1(CC1)NC(=O)C1CCCC1C(=O)N1CCN(CC1)c1ccc(cc1)F	0.0512	0.2260	0.1748	0.9076	0.0945	2	0.0000
20786	N#CC1(CC1)NC(=O)C1CCCC1C(=O)N1CCN(CC1)c1ccc(cc1)OC	0.4014	0.2260	-0.1754	-0.9102	0.0945	2	0.0000
20787	N#CC1(CC1)NC(=O)C1CCCC1C(=O)N1CCN(CC1)c1cccc1	0.3345	0.2260	-0.1085	-0.5630	0.0945	2	0.0000
21229	N#Cc1cccc10c1ccc(cc1)C(=O)NCc1ccnc1	0.2253	0.0654	-0.1599	-0.8165	0.0644	3	0.0000
21230	N#Cc1cccc10c1ccc(cc1)C(=O)NCc1ccnc1	0.1399	0.2260	0.0861	0.4472	0.0945	2	0.0000
22918	O=C(c1cn(c2cccc(c2)C#Cc2ccnc2)c2c(c1=O)ccc2)NC1CC1	0.2405	0.3126	0.0721	0.5397	0.5644	6	5.0856
24213	OC(=O)C(NS(=O)(=O)c1cccc(c1Cl)Cl)CNC(=O)c1scs1	-0.2262	-0.1650	0.0612	0.3188	0.1001	4	-0.4865
24855	OCCC(c1ccc(cc1)Cl)NC(=O)C1(N)CCN(CC1)c1ncnc2c1cc[nH]2	0.2833	0.0076	-0.2757	-1.4065	0.0625	4	0.7172

Figure 28. Example of a CSV file. Items are separated by a comma, as indicated by the arrows

4.2.2 How to create Toxtree derived MLR QSAR XML and CSV files

Toxtree (using the SMARTCyp module Cytochrome P450-Mediated Drug Metabolism and metabolites prediction) can be used to select training set chemicals according to their potential reactivity based on putative cytochrome P450 (CYP) mediated reactions. Before starting, remember to set `toxtree="yes"` as explained in section “How to create a QSAR category in case you are creating a new QSAR category”.

The CSV file must be compiled as explained in the “How to create MLR QSAR XML and CSV files” section, the same for the XML file, except the `header` section which requires additional information like the `organism`, the `assay`, the `rank` and the `reaction`. Below follows an example.

```
<header>
  <item type="name" value="New QSAR"/>
  <item type="organism" value="rat"/>
  <item type="assay" value="hepatocytes"/>
  <item type="rank" value="2"/>
  <item type="reaction" value="aromatic hydroxylation"/>
</header>
```

Organism and assay must be known *a priori*, while the rank and the reaction can be read from the output of Toxtree. Since QSAR-ME Profiler parser is case sensitive, remember to be consistent with values, for example do not write “Rat” in one XML file and “rat” in another. Either you use “Rat” or “rat” (for simplicity it is here suggested to use lowercases only). Concerning reactions names, they must be the same as the ones reported in the Toxtree output files.

4.2.3 How to create LDA QSAR XML and CSV files

Linear Discriminant Analysis (LDA) QSAR XML and CSV files are compiled similarly as explained in the “How to create MLR QSAR XML and CSV files” section. The `header`, `description` and `endpoint` sections must be compiled the same way as MLR QSAR while the remaining sections must conform for LDA, as explained below.

- Locate the `equation` section, `type="discrim_function"`. This section, containing the discriminant functions, is organized the same way as for MRL models, except that the class names must be specified. In the example below three classes named 1, 2 and 3 are specified by `class="1"`, `class="2"` and `class="3"`.

```
<equation type="discrim_function">
  <item type="intercept" class="1" value="-11.52"/>
  <item type="coefficient" class="1" name="MW" value="15.09"/>
  <item type="coefficient" class="1" name="Mp" value="29.61"/>
  <item type="intercept" class="2" value="-1.860"/>
  <item type="coefficient" class="2" name="MW" value="3.266"/>
  <item type="coefficient" class="2" name="Mp" value="10.74"/>
  <item type="intercept" class="3" value="-2.670"/>
  <item type="coefficient" class="3" name="MW" value="-6.085"/>
  <item type="coefficient" class="3" name="Mp" value="8.024"/>
</equation>
```

- To make LDA equations easier to be handled numerically by QSAR-ME Profiler, descriptors need being normalized (by QSAR-ME Profiler), so their minimum and maximum values must be provided. Locate the `equation` section, `type="descr_range"` and then write between quotes the name of the descriptor (e.g., `name="MW"`), its minimum (e.g., `min="30.05"`) and maximum (e.g., `max="498.6"`) values, as in the example below.

```
<equation type="descr_range">
  <item type="value" name="MW" min="30.05" max="498.6"/>
  <item type="value" name="Mp" min="0.500" max="1.190"/>
</equation>
```

- The fitting performances for classification are not obligatory, since are only shown but are not used for calculations, anyway it is strongly recommended to report some to help the user of the model to better understand performances, as in the following example.

```
<performance type="fitting">
  <item type="value" name="Accuracy" value="0.8056"/>
  <item type="value" name="No Information Rate" value="0.4583"/>
  <item type="value" name="P-Value [Acc greater than NIR]" value="1.553e-09"/>
  <item type="value" name="Sensitivity class 1" value="0.8947"/>
  <item type="value" name="Specificity class 1" value="0.9623"/>
  <item type="value" name="Sensitivity class 2" value="0.7273"/>
  <item type="value" name="Specificity class 2" value="0.8718"/>
  <item type="value" name="Sensitivity class 3" value="0.8500"/>
  <item type="value" name="Specificity class 3" value="0.8654"/>
</performance>
```

- Prior classification probabilities must be reported in the `probability` section, `type="prior"` as in the following example for class 1, 2 and 3. These are the *a priori* classes probabilities used to calculate the final discriminant score (for each class, the final score is calculated by adding the natural logarithm of the *a priori* probability to the result of the discriminant function).

```
<probability type="prior">
  <item type="value" class="1" value="0.2638"/>
  <item type="value" class="2" value="0.4584"/>
  <item type="value" class="3" value="0.2778"/>
</probability>
```


- The [dataset](#) section is similar to the same section explained in “How to create MLR QSAR XML and CSV files”, as shown in the example below.

```
<dataset>
  <item type="file_name" value="lda_qsar.csv"/>
  <item type="object" value="CAS"/>
  <item type="smiles" value="SMILES"/>
  <item type="exp_endpoint" value="Exp. class"/>
  <item type="pred_endpoint" value="Pred. class"/>
  <item type="post_prob" class="1" value="Post. prob. 1"/>
  <item type="post_prob" class="2" value="Post. prob. 2"/>
  <item type="post_prob" class="3" value="Post. prob. 3"/>
  <item type="descriptor" value="Mw"/>
  <item type="descriptor" value="Mp"/>
</dataset>
```

It is here recalled that the value of `<item type="file_name">` is the name of the CSV file containing the dataset, in this example is `lda_qsar.csv`. As for MLR, the value of `<item type="object">` is for the CSV file table column containing the names of the chemicals while the value of `<item type="smiles">` corresponds to the column name containing the chemical's SMILES. The values of `<item type="exp_endpoint">`, `<item type="pred_endpoint">` corresponds respectively to the column's names of the experimental and predicted classes. The values of `<item type="post_prob">` items corresponds to the CSV columns containing the post probabilities associated to the classes (the latter must be specified in `class=` within the items). The remaining value of the `<item type="descriptor">` items are the names of the columns containing the QSAR's descriptors. As for MLR, additional columns in the CSV files containing data not referenced in the [dataset](#) section will be ignored, thus can be left in place if you prefer doing so.

4.3 How to create cache data

Repetitive time-consuming calculations are avoided by QSAR-ME Profiler by performing these calculations once and then saving the results in cache data files, to be loaded later when needed. When new QSARs files are created, as explained in the sections above, QSAR-ME Profiler automatically detects missing cache data, so you will be warned as shown in Figure 29.

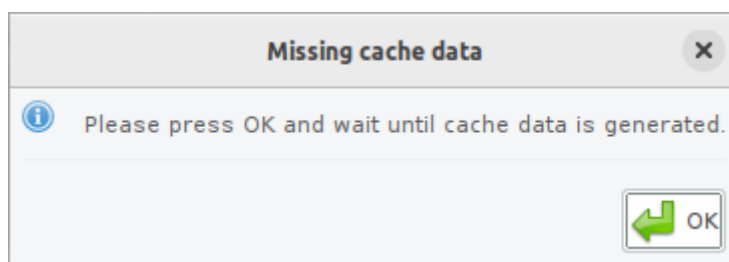


Figure 29. When new QSARs are added, or some cache data has been deleted accidentally, QSAR-ME Profiler will ask for its generation

By pressing OK, QSAR-ME Profiler proceeds by creating the cache data. Since it can take a long while, progress can be followed in the output monitor of QSAR-ME Profiler, as shown in Figure 30.

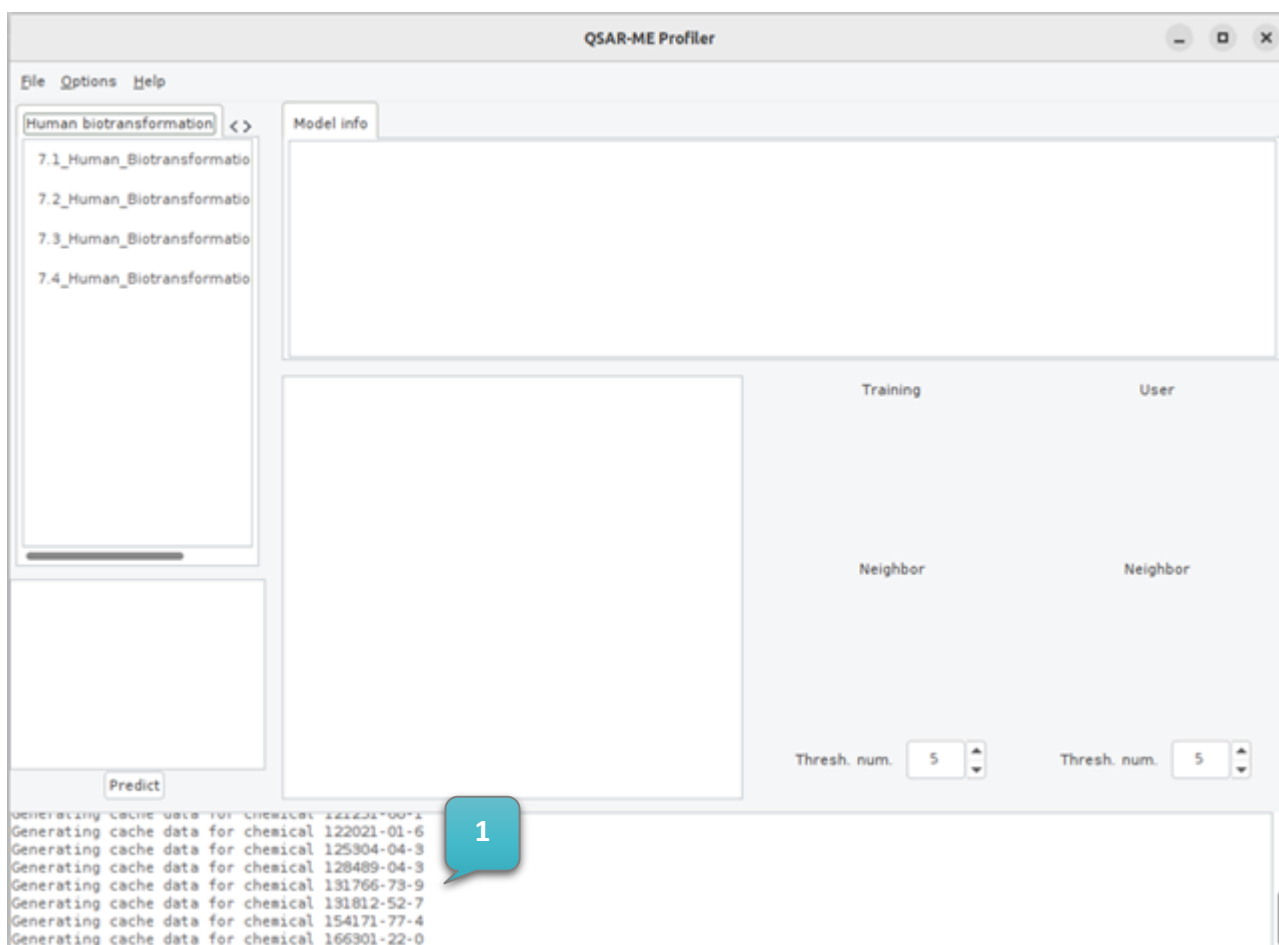


Figure 30. While generating cache data you will be informed on its progress by the output monitor (1)

Sometimes cache generation cannot proceed because of the SMILES format, as shown in Figure 31.

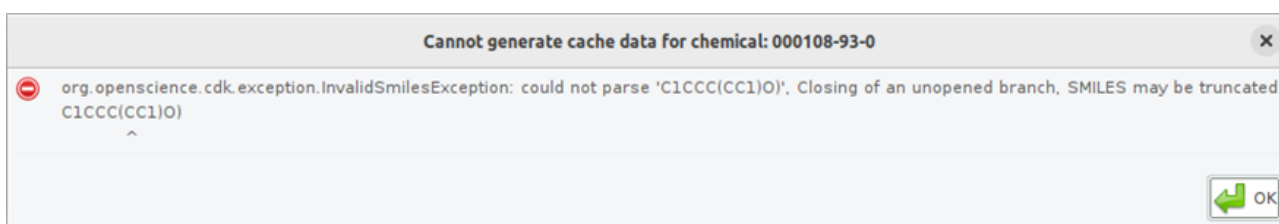


Figure 31. Sometimes cache cannot be generated. See main text for possible solutions.

Such a problem is usually solved by canonicalizing the SMILES (otherwise try different options from your SMILES software generator and/or website). Here it follows some suggested steps to solve the issue.

- 1) Close QSAR-ME Profiler
- 2) Load the CSV file containing the problematic SMILES and substitute the SMILES with one having a different format (as suggested above, try first a canonicalized one) and save the file
- 3) Locate the cache folder in the main QSAR-ME Profiler folder
- 4) Locate the folder corresponding to the QSAR category (see “How to create a QSAR category” for further details) you are working on
- 5) Locate and delete the files with the same name (excluding the extensions) of the problematic model
- 6) Run again QSAR-ME Profiler and press the OK button of the missing cache data dialog. Data cache generation will follow. If the new SMILES is acceptable cache data generation will proceed further.

In case the SMILES is still not acceptable, or the problem is the SMILES of a following chemical, redo steps from 1.

5 Models included in QSAR-ME Profiler

Category	Models
Physico-Chemical properties	1 x Soil organic carbon-water partition coefficient (K_{oc}) ^{24,25}
Global Indexes	1 x Global Half-Life Index (GHLI) ^{24,26} 1 x Insubria PBT index ^{24,27}
Aquatic Toxicity	1 x Fish acute toxicity ^{24,28}
Aquatic Toxicity of personal care products (PCPs)	1 x PCP freshwater algae growth inhibition ²⁹ 1 x PCP <i>Daphnia</i> sp. acute toxicity ²⁹ 2 x PCP fish acute toxicity ²⁹ 1 x PCP Aquatic Toxicity Index (ATI) ²⁹
Aquatic Toxicity of Pharmaceuticals	1 x Pharmaceutical freshwater algae growth inhibition ³⁰ 1 x Pharmaceutical <i>Daphnia</i> sp. acute toxicity ³⁰ 2 x Pharmaceutical fish acute toxicity ³⁰ 1 x Pharmaceutical Aquatic Toxicity Index (ATI) ³⁰
Metabolic transformation	3 x Fish biotransformation ³¹ 4 x Human biotransformation ³² 1 x Human total elimination ^{Errore. Il segnalibro non è definito.} 22 x Rat CYP P-450 microsomes biotransformation ³³ 10 x Rat CYP P-450 hepatocytes biotransformation ³³ 18 x Mouse CYP P-450 microsomes biotransformation ³³ 43 x Human CYP P-450 microsomes biotransformation ³³ 10 x Human CYP P-450 hepatocytes biotransformation ³³
Bioaccumulation	1 x Fish biomagnification factor ³⁴

6 Acknowledgments

We acknowledge the University of Insubria for funding the post doc grant “In silico solutions for the assessment of biotransformation related endpoints of organic chemicals in multiple organisms” (2021-2022), to Dr. Nicola Chirico.

²⁴ Gramatica, P., Cassani, S., Chirico, N. J. Comput. Chem. 2014, 35 (13), 1036–1044.

²⁵ Gramatica, P., Giani, E., Papa, E. J. Mol. Graph. 2007, 25 (6), 755–766

²⁶ Gramatica, P., Papa, E. Environ. Sci. Technol. 2007, 41 (8), 2833–2839.

²⁷ Papa, E., Gramatica, P. Green Chem. 2010, 12 (5), 836–843.

²⁸ Papa, E., Villa, F., Gramatica, P. J. Chem Inf. Model. 2005, 45 (5), 1256–1266.

²⁹ Gramatica, P., Cassani, S., Sangion, A. Green Chem. 2016, 18 (16), 4393–4406.

³⁰ Sangion, A., Gramatica, P. Environ. Int. 2016, 95, 131–143.

³¹ Papa, E., van der Wal, L., Arnot, J. A., Gramatica, P. Sci. Total Environ. 2014, 470, 1040–1046.

³² Papa, E., Sangion, A., Arnot, J. A., Gramatica, P. Food Chem. Toxicol. 2018, 112, 535–543.

³³ IVBP-Suite beta version: in vitro biotransformation prediction suite, 2021, <https://dunant.dista.uninsubria.it/qsar>

³⁴ Bertato, L., Taboureau, O., Chirico, N., Papa, E. SAR QSAR Environ. Res. 2022, 33, 259–271.