



QSARINS-Chem ECO.44

Beta version

Quick Start Guide



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/>

Data collection and curation was performed by ARC Arnot Research & Consulting
<https://arnotresearch.com>

Funding information: European Chemical Industry Council (CEFIC), grant number: CEFIC-LRI ECO.44 (2018-2020) and extension ECO.44.2 (2021).

We also acknowledge the PhD Course in Chemical and Environmental Sciences (DISCA-University of Insubria) for a PhD fellowship granted to Linda Bertato.

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 QSARINS-Chem ECO.44 beta version in your publications as:

Chirico N., Bertato L., Casartelli I., Papa E., QSARINS-Chem ECO.44 beta version, 2021, freely downloadable at: <https://dunant.dista.uninsubria.it/qsar/>

Limitations of liability and disclaimer of warranty

QSARINS-Chem 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 QSARINS-Chem 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 QSARINS-Chem ECO.44 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 QSARINS-Chem ECO.44 beta version or the use or other dealings in QSARINS-Chem ECO.44 beta version, even if the authors have been advised of the possibility of such damages.

The QSARINS-Chem ECO.44 beta version software, the accompanying materials and manuals are protected by copyright: 2018, University of Insubria, <http://www.uninsubria.it> - Varese, Italy.

Overview

QSARINS-Chem ECO.44 beta version allows the user to input his molecular structures and get estimations and Applicability Domain for a desired QSAR model.

Minimal information required to obtain prediction are:

-Chemical structures files (e.g. .smi, .mol) (see below to see how to set a .smi file)

Tutorial

Step 1: Select the model:

1) **Run** QSARINS-ChemECO44.jar (usually by double clicking on its icon. If it is not working, please ask your IT staff because it depends on whether, or how, the Java environment is configured on your machine). The first time you execute QSARINS-Chem ECO.44 beta version, you need to read the licence agreement and if you agree, by selecting “I agree”, you can use QSARINS-Chem ECO.44 beta version.



Figure 1. Licence agreement

2) The “Model Selection” page will open; here **you can select** the desired **QSAR model** from the blue drop-down menu. Once selected, the main page summarizes the principal information of the selected model (model’s description, equation and statistics).

As a practical tutorial, select “quick_start_example”, as shown in Figure 2. The endpoint of this model is “logHLn (days)”, that is “the base-10 logarithm of the whole-body biotransformation half-life of chemicals in fish in days normalized for a reference 10g fish at a water temperature of 288K”. This model will be used in the following steps as a tutorial (which is highlighted in this color).

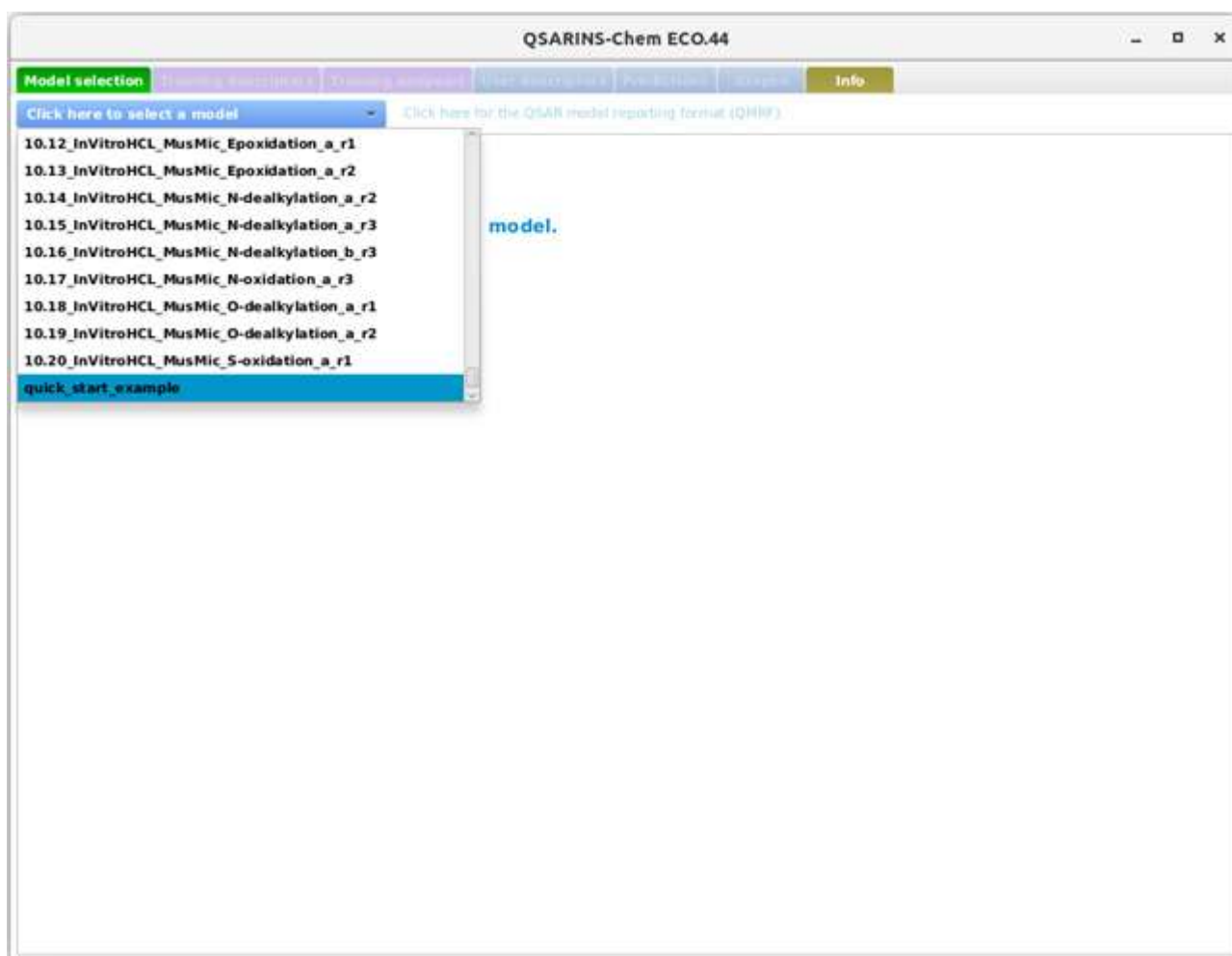


Figure 2. Model selection

3) Once the model is selected, two tabs (“Training descriptors”, Figure 3, and “Training endpoint”, Figure 4) are activated in order to provide information about the Training set used for the model development:

The “Training descriptors” tab shows the values of the training set descriptors of the selected QSAR model. These values can be used, as a reference, to check the validity of the descriptors values calculated for the new molecules, i.e. those provided by the user (see Step 2).

The “Training endpoint” tab shows the experimental and estimated endpoint for the training set objects as well as the residuals and the leverage values (i.e. the diagonal elements of the Hat matrix). Outliers for the response and influential objects are highlighted in red. These values can be used as a reference for checking the validity of the predictions (see Step 3 for further information).

Model selection		Training descriptors	Training endpoint	User descriptors		Predictions	Graphs	Info
No.	SMILES	gmax	VAdjMat	nHBAcc	nX	SaaaC	PubchemFP...	
1	S(=O)(=O)(O)...	11	5.5	3	0	0	0	
2	OCCOCCOCCO...	8.6	6.2	9	0	0	0	
3	S(=O)(=O)(O)...	11	5.5	3	0	0	0	
4	FC(F)(Oc1ccc(...	13	6	0	8	0	0	
5	FC(F)(Oc1ccc(...	13	6.1	0	8	0	0	
6	FC(F)(Oc1ccc(...	13	6.1	0	8	0	0	
7	FC(F)(Oc1ccc(...	13	5.4	0	4	0	0	
8	FC(F)(Oc1ccc(...	13	5.4	0	4	0	0	
9	FC(F)(Oc1ccc(...	13	5.3	0	4	0	0	
10	S=P(OC)(OC)O...	9	5	3	1	0	1	
11	c1c(c(ccc1Oc1...	6	5.1	0	4	0	0	
12	c1c(Br)c(Br)cc...	6	5.1	0	4	0	0	
13	Brc1cc(Br)c(Br...	6.4	5.2	0	6	0	0	
14	ClC(Cl)(C(c1cc...	2.4	5.2	0	5	0	1	
15	c1cc2c3ccc4c...	2.3	5.3	0	0	11	0	
16	ClC(Cl)(C@@H...	2.4	5.2	0	4	0	1	
17	c12ccccc1c1c(...	2.3	5.5	0	0	11	0	
18	CSclc(C)cc(cc...	6	5	2	0	0	0	
19	ClC(Cl)(Cl)Cl	1.4	3.3	0	4	0	0	

Figure 3. Training descriptors used for model development

Model selection		Training descriptors	Training endpoint	User descriptors	Predictions	Graphs	Info
No.	SMILES	Experimental endpoint	Estimated endpoint	HAT V/I (h* = 4.7e-02)			
1	S(=O)(=O)(O)...	0.34	-0.45	1.3e-02			-0.79
2	OCCOCCOCCO...	-0.61	-0.42	7.8e-02			0.19
3	S(=O)(=O)(O)...	8.1e-02	-0.45	1.3e-02			-0.53
4	FC(F)(Oc1ccc(...	2.6	1.7	3.2e-02			-0.93
5	FC(F)(Oc1ccc(...	1.7	1.8	3.3e-02			0.10
6	FC(F)(Oc1ccc(...	2.4	1.8	3.3e-02			-0.68
7	FC(F)(Oc1ccc(...	0.36	0.60	1.8e-02			0.24
8	FC(F)(Oc1ccc(...	0.4	0.60	1.8e-02			0.20
9	FC(F)(Oc1ccc(...	0.29	0.54	1.7e-02			0.25
10	S=P(OC)(OC)O...	0.65	0.18	1.3e-02			-0.47
11	c1c(c(ccc1Oc1...	2.3	1.1	6.1e-03			-1.1
12	c1c(Br)c(Br)cc...	1.4	1.1	6.1e-03			-0.29
13	Brc1cc(Br)c(Br...	2	1.5	1.1e-02			-0.51
14	ClC(Cl)(C(c1cc...	1.9	2.3	8.5e-03			0.36
15	c1cc2c3ccc4c...	4.7e-02	0.15	5.7e-02			0.11
16	ClC(Cl)(C@@H...	1.7	2.1	8.3e-03			0.44
17	c12ccccc1c1c(...	0.21	0.33	5.4e-02			0.12
18	CSclc(C)cc(cc...	-0.36	0.25	5.5e-03			0.61
19	ClC(Cl)(Cl)Cl	-1.2	-0.19	2.1e-02			1.0

Figure 4. Endpoint values used for model development and related additional information

Step 2: Calculate and edit the molecular descriptors – apply them to predict the endpoint for your molecules

In this section you will learn how to calculate the descriptors for your molecules. This step is obligatory in order to apply the model to the user-entered molecules, for the endpoint prediction.

1) **Select** “User descriptors” tab, as shown in Figure 5.

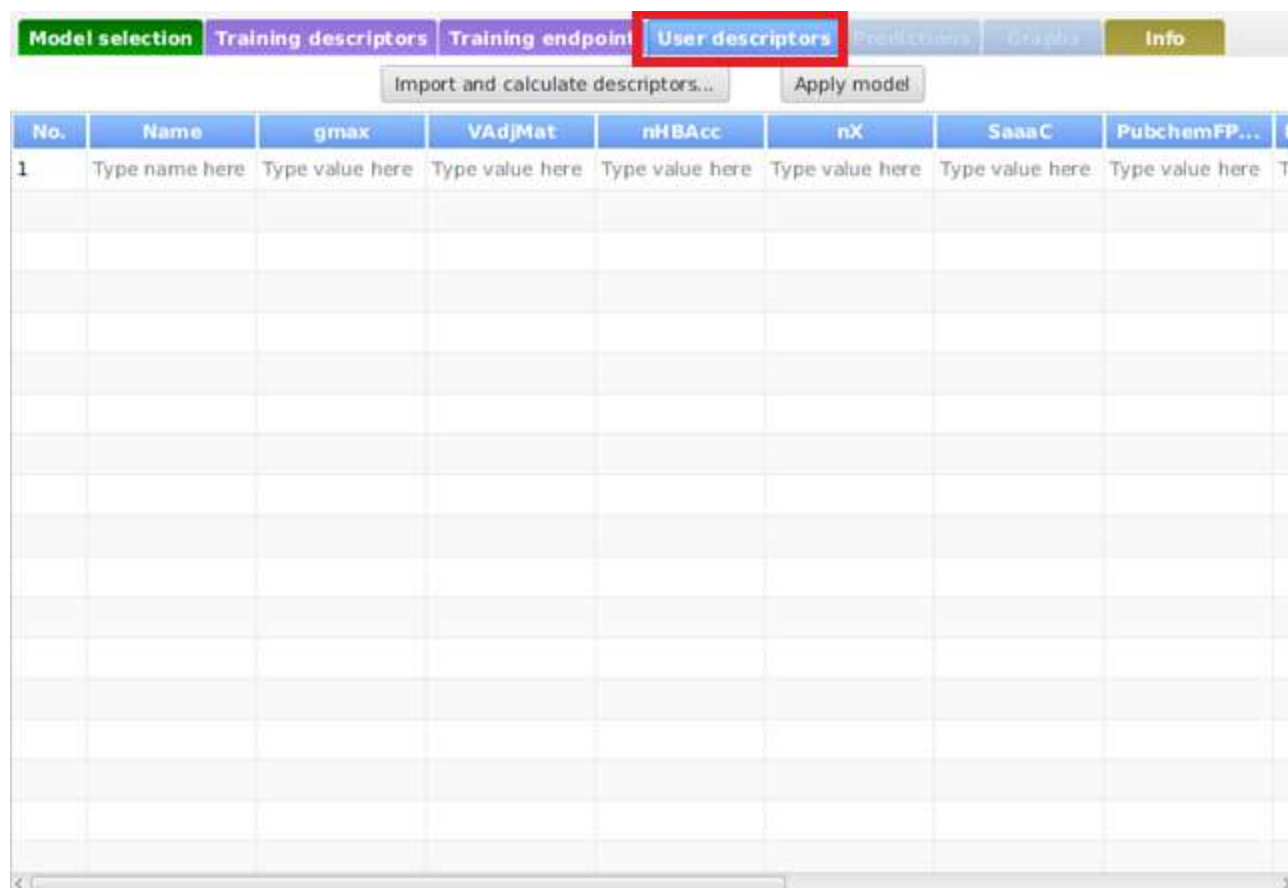


Figure 5. Descriptors calculation

2) **Press** “Import and calculate descriptors”: you will be asked to select a folder containing the molecular structures (acceptable file formats must be checked in the PaDEL-Descriptor software documentation. This software, for descriptors calculation, is freely available and included in the QSARINS-Chem ECO.44 folder, see licence in the “Info” tab for further information). **As an example for this tutorial, go to the QSARINS-Chem main folder (it should be “QSARINS-Chem-ECO44”), then enter the “help” folder and subsequently the “quickstart_example” folder. Finally enter the “smiles” folder and confirm the folder from your dialog (the dialog layout depends on your O.S.). Wait until PaDEL-Descriptor completes the calculations.**

You can also enter the descriptor values manually, usually by means of copy/paste, in case you prefer using a different software for their calculation.

***Optional:** if available, **you can enter** the experimental value of your response (in the same units of the selected QSAR models). **Type in** your value/values in the user response column (**in the tutorial example is “logHLn (days)”**) and **press** “Enter”.

****Warning:** The column “Status” will report the presence of issues in your data by means of a red warning (i.e. descriptors and/or user response out of range of the training set or missing descriptors, see Figure 6 as

Model selection Training descriptors Training endpoint User descriptors Predictions Graphs Info

[illegible]

Figure 6. Descriptors values calculated for user-entered molecules

In the example of this tutorial (see Figure 6) four experimental values of the endpoint (“logHLn (days)”) have been manually entered. These values are optional, but when provided they help to evaluate the reliability of the predictions.

3) **Press** “Apply Model” to run the model and generate predictions.

Step 3: get estimated values and evaluate the applicability domain

Once the model is applied, the “Predictions” tab will be activated (see Figure 7). This tab contains, in addition to the ID number and the molecule names, the following information:

- 1) **Experimental endpoint.** This information is optional, see Step 2 for further information.
- 2) **Estimated endpoint.** This is the **endpoint predicted by the model** of the user-entered molecules. The prediction of the endpoint is **the aim of using QSARINS-Chem ECO.44 beta version**.
- 3) **HAT values (leverages).** These values represent the “distance of the molecular structures” of the molecules entered by the user, respect to the ones used for the model development. When the HAT value of the user-entered molecules is above the calculated threshold (h^*) that means they could be structurally different from the ones used for the development of the model. These molecules need further checking. *An example of this check is shown in the following Step 4: graphical inspection.*

- 4) **Residual and standardized residuals.** These values are a measure of the distance between the predicted and the experimental values if the user provides the latter. The smaller the residual, the lower the error in prediction.
- 5) **Status.** If problems are detected, the status is displayed as “Warning” in red. Moving the mouse pointer on the red value/s (e.g. 26 or 9.2e-02 of the tenth molecule in the example of Figure 7) on the same row a tooltip will appear indicating the reason of the problem.

Model selection	Training descriptors	Training endpoint	User descriptors	Predictions	Graphs	Info	
No.	Name	Experimental endpoint	Estimated endpoint	HAT VI ($h^+ = 4.7e-02$)	Residual	Standardized residual	Status
1	Pyrene	0.32	0.13	3.2e-02	-0.19	-0.33	OK
2	BaP	4.7e-02	0.15	5.7e-02	0.11	0.19	Warning
3	Methoxychlor	Not provided	1.5	7.5e-03	Need exp...	Need experimental end...	OK
4	Deltamethrin	0.51	0.47	1.6e-02	-3.8e-02	-6.6e-02	OK
5	4-nonyl-phenol	-0.23	-0.20	1.8e-02	2.8e-02	4.8e-02	OK
6	Cyclohexyl-salicylate	Not provided	-0.87	1.1e-02	Need exp...	Need experimental end...	OK
7	Tetrahydropyrene	Not provided	1.1	8.8e-03	Need exp...	Need experimental end...	OK
8	1-Chloromethylpyrene	Not provided	0.38	3.4e-02	Need exp...	Need experimental end...	OK
9	1-Aminopyrene	Not provided	-0.69	3.5e-02	Need exp...	Need experimental end...	OK
10	Decafluoro-pyrene	Not provided	2.6	9.2e-02	Need exp...	Need experimental end...	Warning

Figure 7. Endpoint predictions for user-entered molecules

You can **select** the chemicals of your interest and **right click to copy** the predictions. **Paste** in “Excel” or in other format to save the estimations.

Step 4: graphical inspection

To visualize the graphs for the estimation of the predicted values performances, **select the “Graphs” tab** and then **press the “Calculate graphs” button**. The following graphs (Figures 8-11) will be displayed.

1) HAT vs. estimated endpoint Graph: plot of HAT values (leverages) (see also point 3 in Step 3) vs. estimated values of the model. The red points (Training) are the predicted values of the molecules endpoint used for the model development, the light blue dots (User Set N) the user-entered molecules without the experimental value (optional) while the blue dots (User Set E) are the ones with the experimental value.

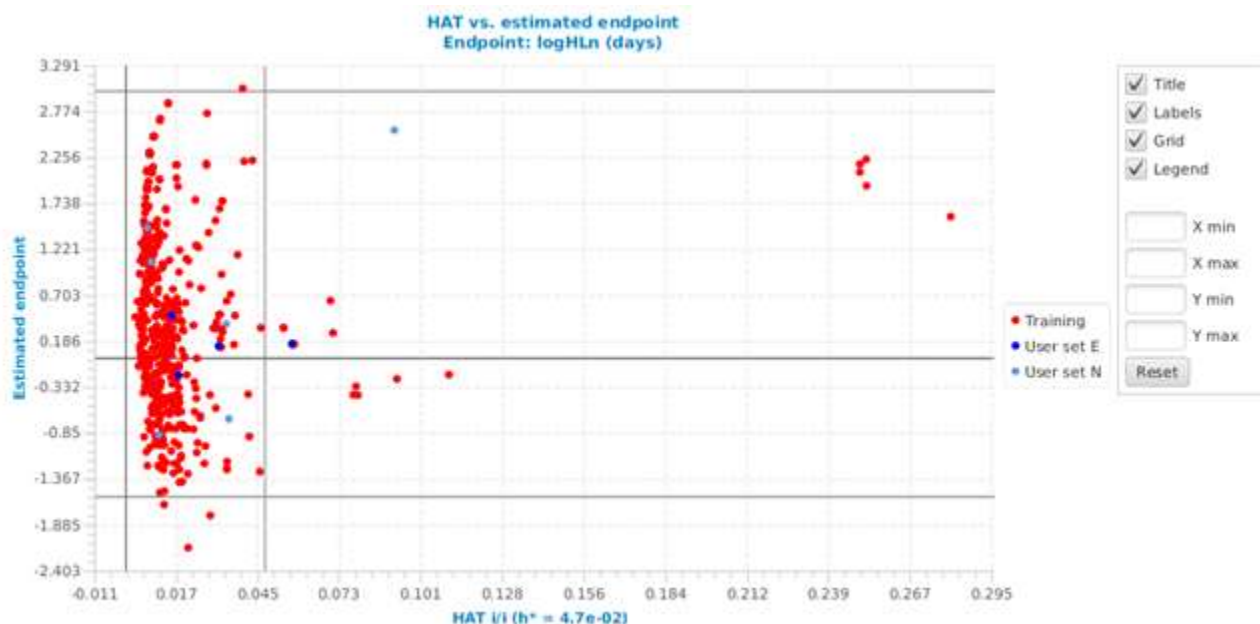


Figure 8. HAT vs. estimated endpoint

2) Experimental vs. Estimated values. This plot provides visual information of the fitting of the model (Training, red points). If experimental values are provided by the user-entered molecules, they will appear in this graph (User set, blue points). *In the example of this tutorial, see Figure 9, all blue dots are within the scatter plot of the training set (red dots). That means that the user-entered molecules predictions are within the experimental and predictivity domain of the model. In case they are outside, further checking of user-entered molecules should be performed.*

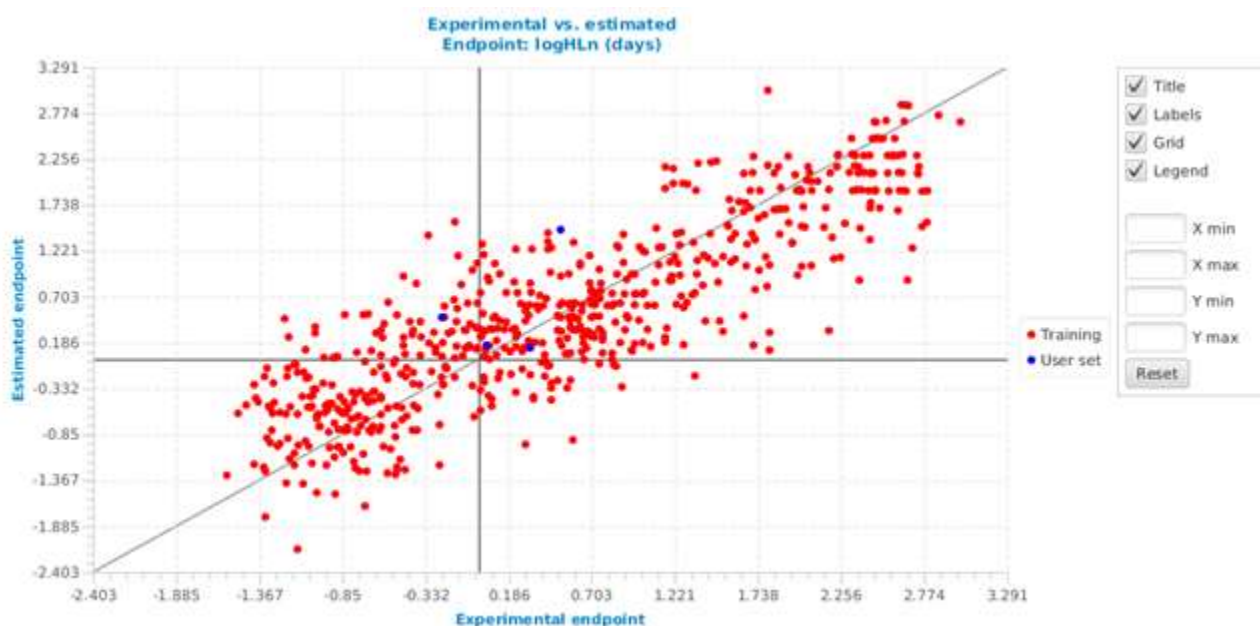


Figure 9. Experimental vs. estimated endpoints

3) Residuals. This graph works similarly to the previous one (“Experimental vs Estimated values”) but shows residuals instead. It is commonly used to evaluate the appropriateness of using a linear model. When used on user-entered data, the blue points (User set) should fall within the range of dispersion of the red points (Training).

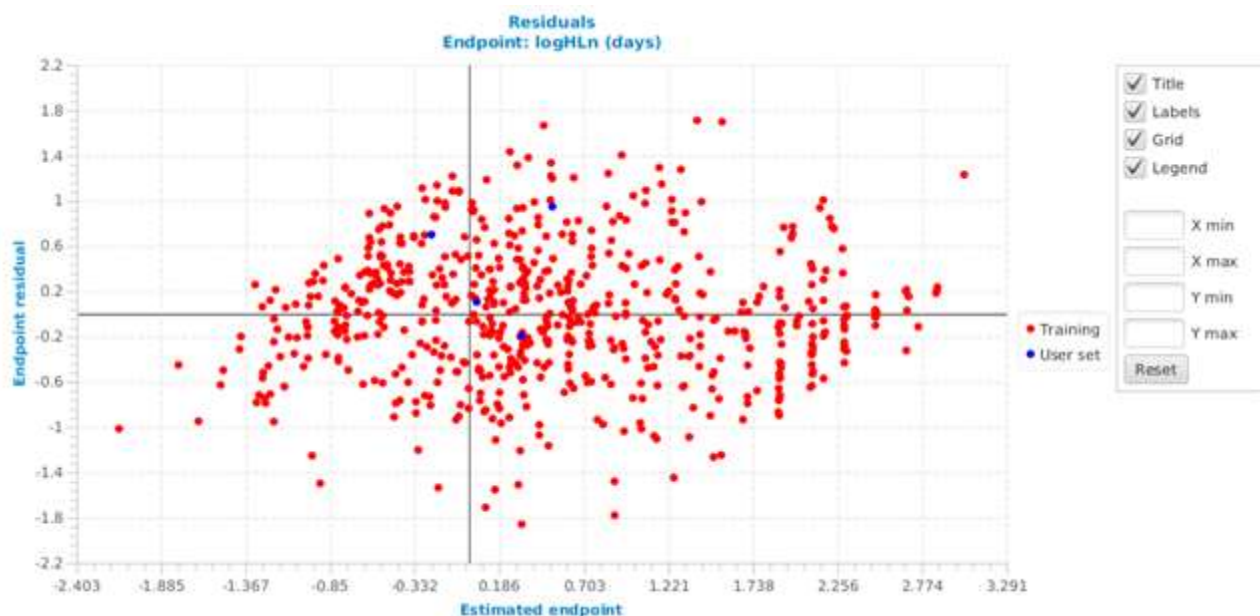


Figure 10. Endpoint residuals

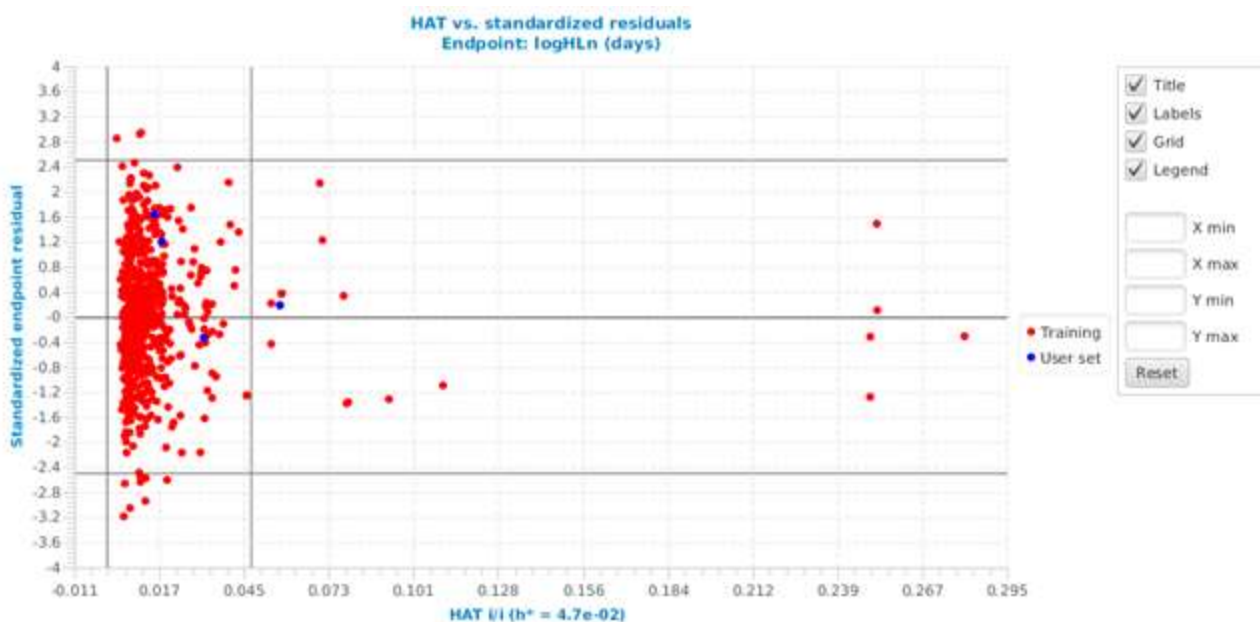


Figure 11. HAT vs. standardized residuals

4) HAT vs. standardized residuals graphs. This graph is similar to the HAT vs. estimated endpoint graph. The difference is in the ordinate axis that reports the standardized residuals of the responses. The HAT vs. estimated endpoint graph allows for the visualization of “User set” chemicals in the applicability domain

defined by leverage distance and experimental range of the related model. Differently, the HAT vs. standardized residuals graphs plots the leverage distance vs. standardized residuals. Therefore, the user-entered molecules (blue dots) must be provided with the experimental values otherwise the residual cannot be calculated. The use of the residuals helps in better evaluating the reliability of the predictions.

Note: you can copy and paste or save any of these graphs.

How to set up a SMILES structural file (.smi)

PaDEL-Descriptor software can calculate descriptors from SMILES placed in a **.smi** structural file. This is a “tab delimited” text file containing the **SMILES structures in the first column** and the Identifiers (optional) in the second column, **no header**. The extension of file must be **.smi**

To generate this file in “Excel” (or similar software):

- 1) **Open** an empty document
- 2) **Paste** your SMILES in the first column
- 3) If present, **paste** the identifier (ID or CAS or NAME) in the second column
- 4) If present, **delete** the header
- 5) **Save** as tab delimited text file [**TEXT (tab delimited)(*.txt)**]
- 6) Manually **change** the extension from **.txt** to **.smi**

(To see file extension on

Windows 7: Open **Windows Explorer** and click the **Organize** button towards the top left. Choose **Folder and search options** from the menu. Click the **View** tab in the window that opens, then scroll down and untick the box next to ‘**Hide file extensions for known file types**’

Windows 8/10: open a **File Explorer** window (the new name for Windows Explorer) and click the **View** tab.

Mac: Click on the **Finder** menu and select **Preferences**. Select **Advanced** button and put a check mark in the checkbox labeled **Show all filename extension**)

- 7) **Place** the file in an empty folder to use in **Step 2 point 3**