QSARINS-Chem

Standalone Version

Quick Start Guide

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

Authors

Nicola Chirico (Research Fellow)
Alessandro Sangion (Research Fellow)
Paola Gramatica (Senior Professor)
Ester Papa (Associate Professor)

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

Information about the Standalone Version

Models

This standalone version of QSARINS-Chem contains QSAR models for the prediction of:

Category	**Model**
1. Physico-Chemical properties	1. Soil Organic Carbon-Water partition Coefficient (K_{OC})
2. Global Indexes	1. Global Half-Life Index 2. Insubria PBT Index
3. Aquatic Toxicity	1. Fish Acute Toxicity (*P.promelas*)
4. Aquatic Toxicity of Personal Care Products (PCPs)	1. PCP Freshwater Algae Growth Inhibition 2. PCP *Daphnia* sp. Acute Toxicity 3. PCP Fish Acute Toxicity Model 1 (logP based) 4. PCP Fish Acute Toxicity Model 2 5. PCP Aquatic Toxicity Index (ATI)
5. Aquatic Toxicity of Pharmaceuticals	1. Pharmaceutical Freshwater Algae Growth inhibition 2. Pharmaceuticals *Daphnia* sp. acute Toxicity 3. Pharmaceuticals fish Acute Toxicity (*O.mykiss*) 4. Pharmaceuticals fish Acute Toxicity (*P.promelas*) 5. Pharmaceuticals Aquatic Toxicity Index (ATI)
6. Metabolic transformation in fish	1. Fish Biotransformation Model 1 2. Fish Biotransformation Model 2 3. Fish Biotransformation Model 3
7. Metabolic transformation in human	1. Human biotransformation Model 1 2. Human biotransformation Model 2 3. Human biotransformation Model 3 4. Human biotransformation Model 4 5. Human Total elimination

Database

This standalone version of QSARINS-Chem contains the database published by Gramatica et al. 2014[1] which includes the following datasets:

Chemical Class	**Endpoint-Type**	**Dataset Name**
1. General	1. Physico-Chemical Properties	1. Soil Organic Carbon-Water partition Coefficient (K_{OC})
	2. Environmental Persistence	1. Sediment Half-Lives 2. Soil Half-Lives 3. Water Half-Lives 4. Air Half-Lives 5. NO_3 reactivity 6. O_3 reactivity 7. OH reactivity 8. Global Half-Life Index
	3. Bioconcentration Factor	1. BCF-Fernandez 2. BCF-Lu
	4. Metabolic Transformation	1. Fish Biotransformation 2. Human Biotransformation Model 1 3. Human Biotransformation Model 2 4. Human Biotransformation Model 3 5. Human Biotransformation Model 4 6. Human Total Elimination
	5. Aquatic Toxicity	1. Fish acute toxicity (*P.promelas*)
	6. Endocrine Disruption	1. Estrogen Receptor Binding
2. Aromatic Amines	1. Mutagenicity	1. Aromatic Amines mutagenicity TA98 2. Aromatic Amines mutagenicity TA100
3. (Benzo)Triazoles	1. Physico-Chemical Properties	1. (B)TAZ Kow 2. (B)TAZ Solubility in Water 3. (B)TAZ Vapor Pressure 4. (B)TAZ Melting Point
	2. Aquatic Toxicity	1. (B)TAZ Algae acute toxicity (*P.subcapitata*) 2. (B)TAZ *Daphnia sp* acute toxicity 3. (B)TAZ Fish acute toxicity (*O.mykiss*)
4. Brominated Flame Retardants	1. Physico-Chemical Properties	1. BFR Kow 2. BFR Koa 3. BFR Vapor Pressure 4. BFR Solubility in Water 5. BFR Henry Law Constant 6. BFR Melting Point
	2. Endocrine Disruption	1. BFR-DR-Ag 2. BFR-ER-Ag 3. BFR-ERODind

Chemical Class	**Endpoint-Type**	**Dataset Name**
		4. BFR-PR-ant 5. BFR-SULT-REP 6. BFR-T4-REP 7. BFR Receptor Binding Affinity
5. Dioxin Analogues	1. Biological Activity	1. Dioxin Analogues pAHH 2. Dioxin Analogues pRB
6. Esters	1. Physico-Chemical Properties	1. Esters Flash Point
	2. Aquatic Toxicity	1. Esters Algae acute toxicity 2. Esters *Daphnia sp* acute toxicity 3. Esters Fish acute toxicity (*P.promelas*) 4. Esters Aquatic Toxicity Index (EATIN)
7. Fragrances	1. Terrestrial Toxicity	1. Fragrances Oral toxicity (Rat)
	2. Biochemical activity	1. Fragrances Inhibition NADHox 2. Fragrances Mitochondrial memb pot
8. Nitrated polycyclic aromatic hydrocarbons	1. Mutagenicity	1. NitroPAH mutagenicity TA100
9. Perfluorinated Compounds	1. Physico-Chemical Properties	1. PFC Critical Micelle Concentration 2. PFC Solubility in Water 3. PFC Vapor Pressure
	2. Terrestrial Toxicity	1. PFC Oral toxicity (Rat) 2. PFC Oral toxicity (Mouse) 3. PFC Inhalation toxicity (Rat) 4. PFC Inhalation toxicity (Mouse)
10. Personal Care Products	1. Aquatic Toxicity	1. PCP Algae acute toxicity (*P.subcapitata*) 2. PCP *Daphnia sp* acute toxicity 3. PCP Fish acute toxicity (*P.promelas*)
11. Pharmaceuticals	1. Aquatic Toxicity	1. Pharm. Algae acute toxicity (*P.subcapitata*) 2. Pharm. *Daphnia sp* acute toxicity 3. Pharm. Fish acute toxicity (*O.mykiss*) 4. Pharm. Fish acute toxicity (*P.promelas*)

NOTE ABOUT DATA AND STRUCTURES: All the chemical structures have been designed and "energetically minimized" (AM1 method in HyperChem v.7.03) by the QSAR Research Unit in Environmental Chemistry and Ecotoxicology at the University of Insubria[1]. SMILES (not in the canonical form) reported in the database have been generated from the 3D structures using OpenBabel v 2.3.2. All the experimental data have been collected from literature. Information about original data, data collection and curation can be found in the publications cited in the database. No experimental data have been generated by QSAR Research Unit in Environmental Chemistry and

---

[1] "QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS", Gramatica P., Cassani S., Chirico N. Journal of Computational Chemistry 2014, 35, 1036-1044. DOI: 10.1002/jcc.23576

Ecotoxicology at the University of Insubria (exceptions are the global indexes generated by Principal Component Analysis).

Acknowledgments

The authors would like to thank Stefano Cassani for the revision of the chemical structures (.hin files) of the database (till 2014). In addition helpful comments for this standalone version were provided by Jon Arnot (Arnot Research and Consulting, Inc.) and James Armitage (Armitage Environmental Sciences, Inc).

How to cite:

Please cite QSARINS-Chem standalone version in your publications as:

Chirico Nicola, Sangion Alessandro, Gramatica Paola, Bertato Linda, Casartelli Ilaria, Papa Ester, QSARINS-Chem standalone version: a free platform-independent software for QSAR-based predictions of properties and activities of organic chemicals, submitted to J.Comput. Chem.

We would also appreciate citations to the website: http://dunant.dista.uninsubria.it/qsar/

Limitations of liability and disclaimer of warranty

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

Overview

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

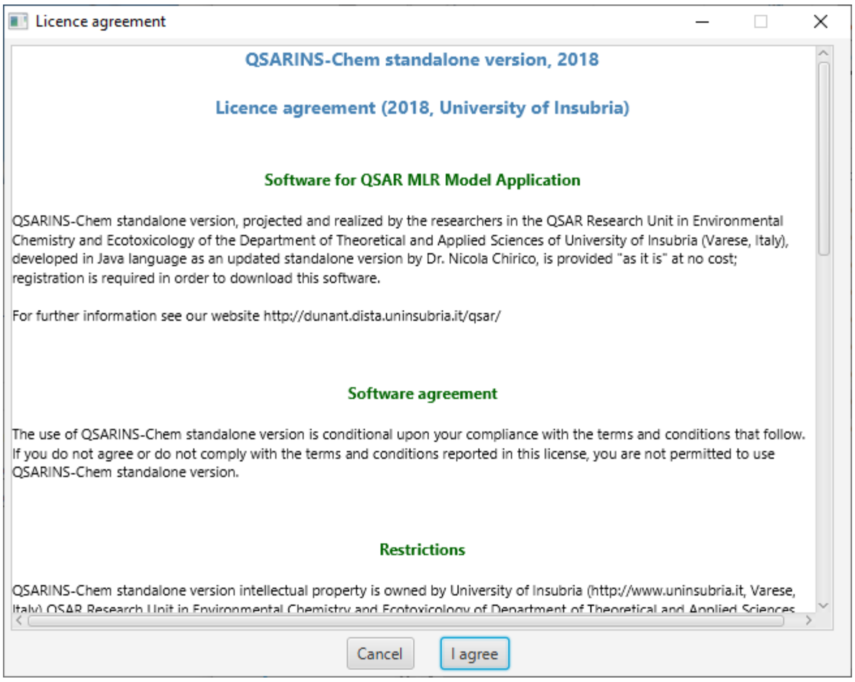
Minimal information required to obtain prediction are:

-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-Chem.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, you need to read the Licence agreement and if you agree, by selecting "I agree", you can use QSARINS-Chem.



QSARINS-Chem standalone version, 2018

Licence agreement (2018, University of Insubria)

Software for QSAR MLR Model Application

QSARINS-Chem standalone version, projected and realized by the researchers in the QSAR Research Unit in Environmental Chemistry and Ecotoxicology of the Department of Theoretical and Applied Sciences of University of Insubria (Varese, Italy), developed in Java language as an updated standalone version by Dr. Nicola Chirico, is provided "as it is" at no cost; registration is required in order to download this software.

For further information see our website http://dunant.dista.uninsubria.it/qsar/

Software agreement

The use of QSARINS-Chem standalone version is conditional upon your compliance with the terms and conditions that follow. If you do not agree or do not comply with the terms and conditions reported in this license, you are not permitted to use QSARINS-Chem standalone version.

Restrictions

QSARINS-Chem standalone version intellectual property is owned by University of Insubria (http://www.uninsubria.it, Varese, Italy) QSAR Research Unit in Environmental Chemistry and Ecotoxicology of Department of Theoretical and Applied Sciences

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 "Fish biotransformation model 1", 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).

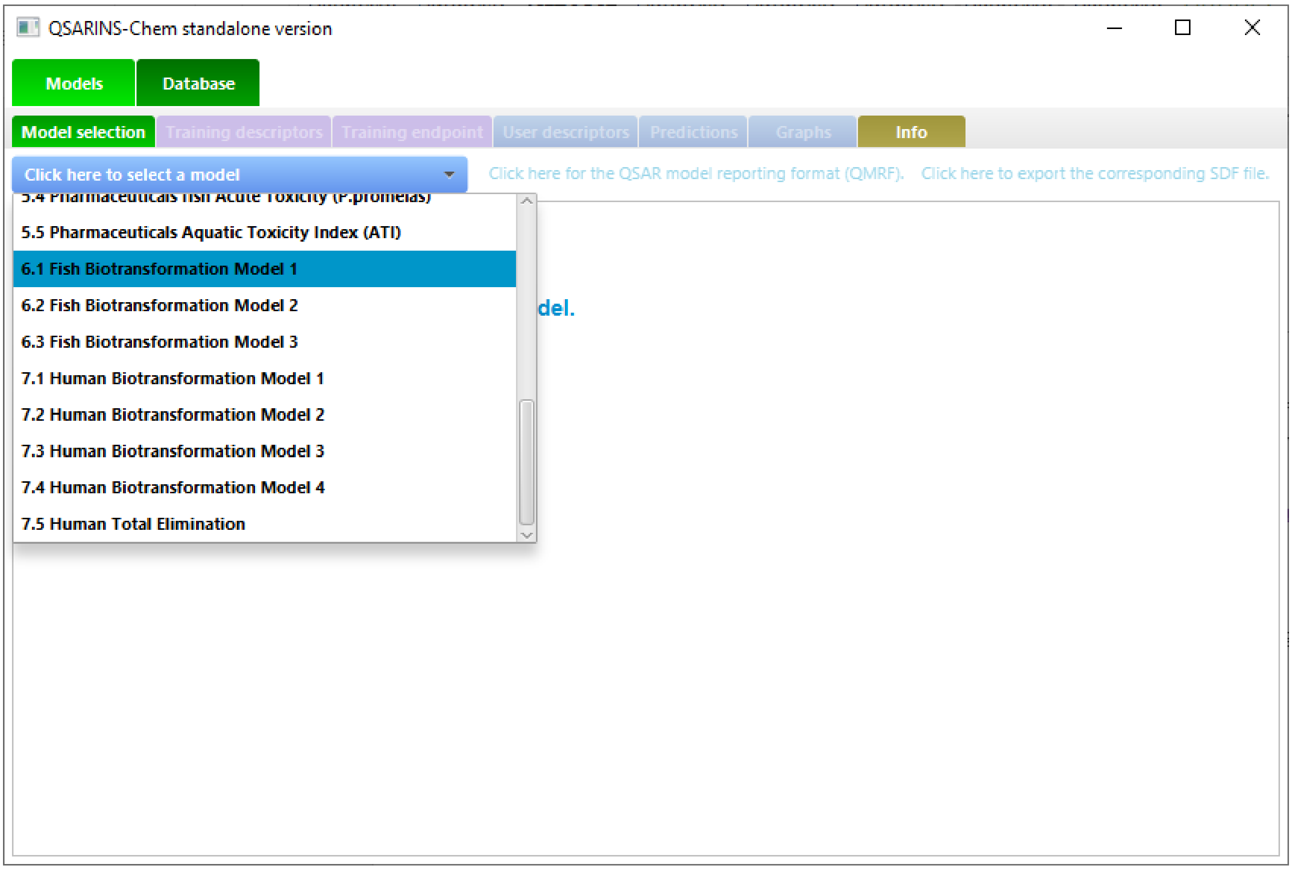
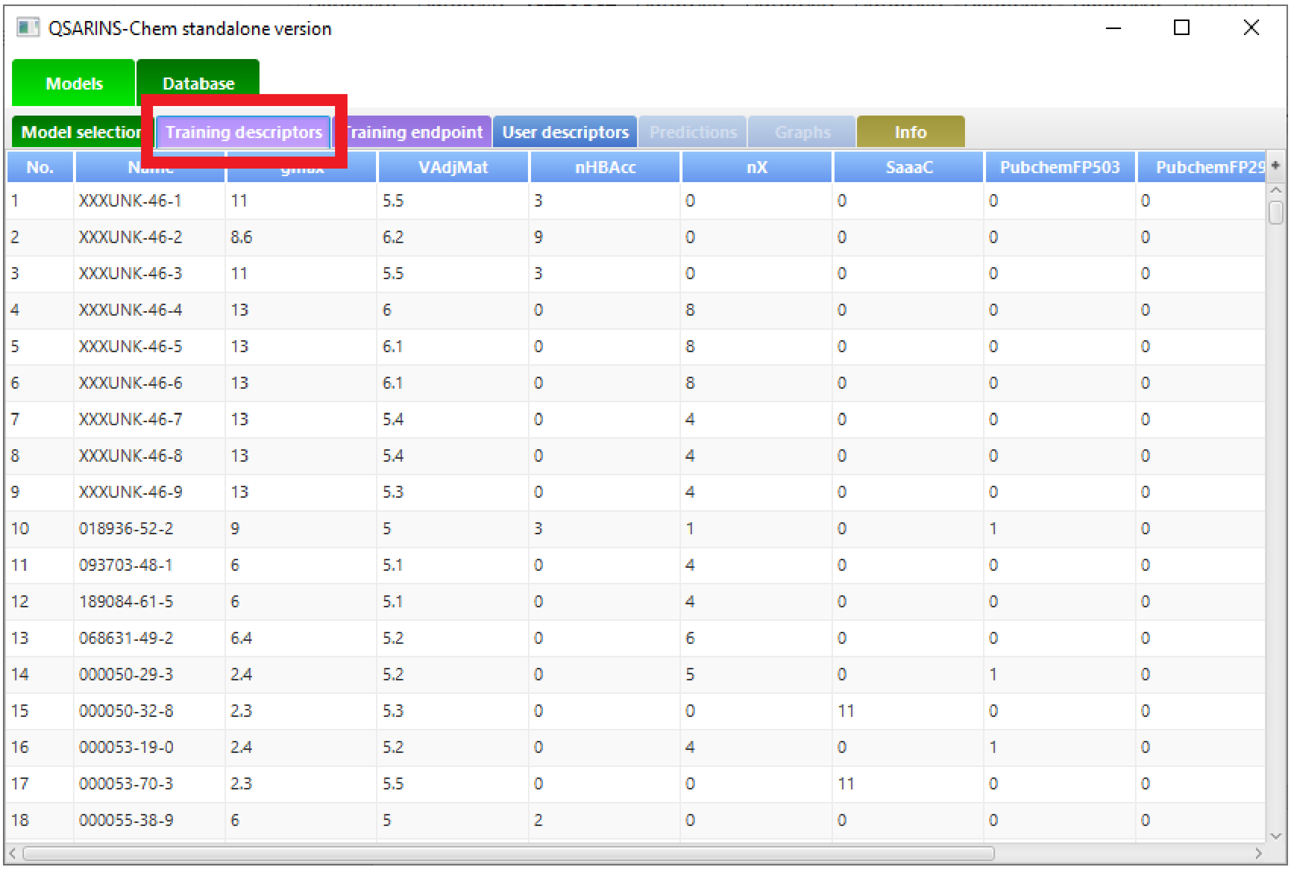

QSARINS-Chem standalone version
Models
Database
Model selection
Training descriptors
Training endpoint
User descriptors
Predictions
Graphs
Info
Click here to select a model
Click here for the QSAR model reporting format (QMRF).
Click here to export the corresponding SDF file.
5.4 Pharmaceuticals fish Acute Toxicity (P.promelas)
5.5 Pharmaceuticals Aquatic Toxicity Index (ATI)
6.1 Fish Biotransformation Model 1
6.2 Fish Biotransformation Model 2
6.3 Fish Biotransformation Model 3
7.1 Human Biotransformation Model 1
7.2 Human Biotransformation Model 2
7.3 Human Biotransformation Model 3
7.4 Human Biotransformation Model 4
7.5 Human Total Elimination
del.

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:

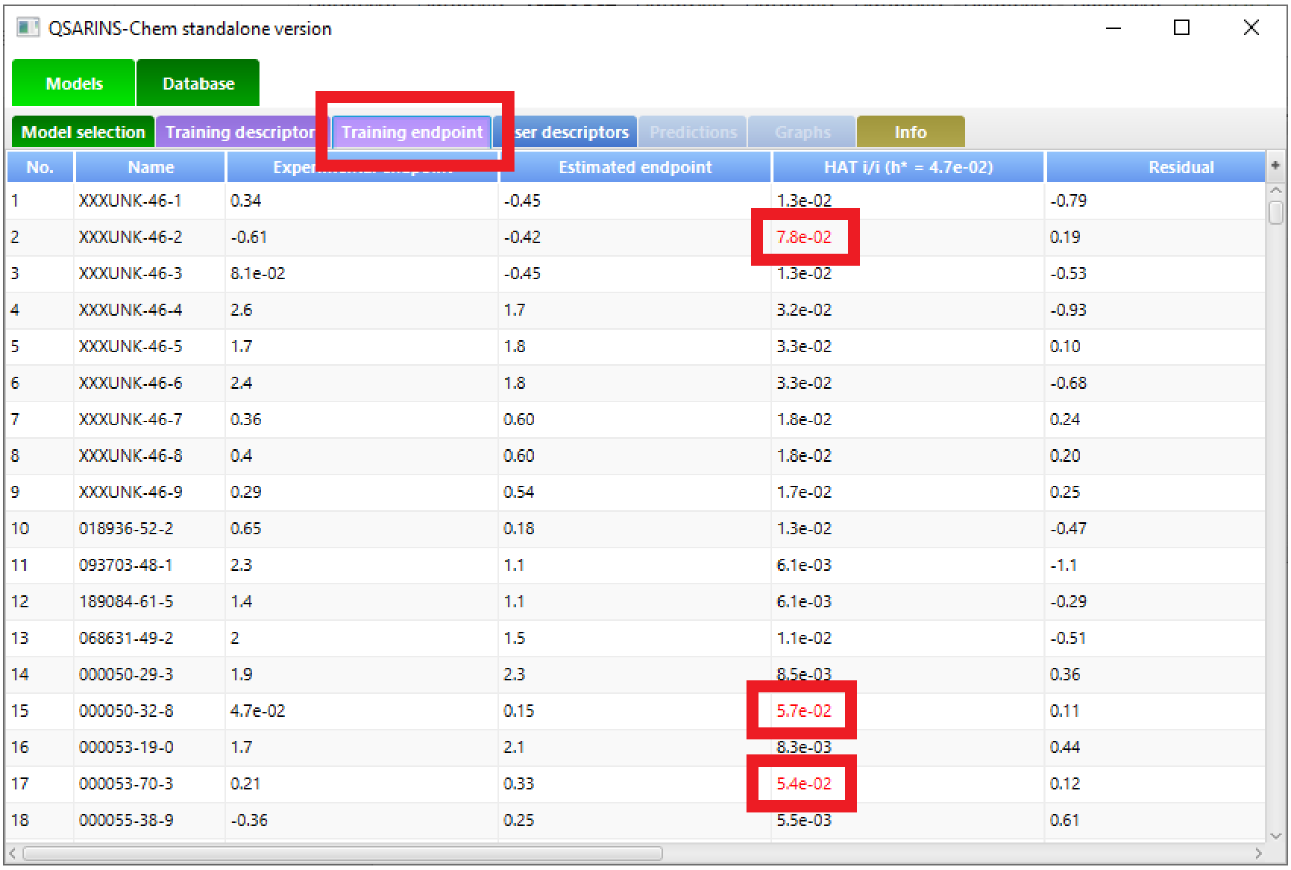
The "Training descriptors" tab shows the values of the training set descriptors of the selected QSAR model. These values can be used, as a reference, to check the validity of the descriptors values calculated for the new molecules, i.e. those provided by the user (see Step 2).

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).



No.	Name	gmax	VAdjMat	nHBAcc	nX	SaaaC	PubchemFP503	PubchemFP29
1	XXXUNK-46-1	11	5.5	3	0	0	0	0
2	XXXUNK-46-2	8.6	6.2	9	0	0	0	0
3	XXXUNK-46-3	11	5.5	3	0	0	0	0
4	XXXUNK-46-4	13	6	0	8	0	0	0
5	XXXUNK-46-5	13	6.1	0	8	0	0	0
6	XXXUNK-46-6	13	6.1	0	8	0	0	0
7	XXXUNK-46-7	13	5.4	0	4	0	0	0
8	XXXUNK-46-8	13	5.4	0	4	0	0	0
9	XXXUNK-46-9	13	5.3	0	4	0	0	0
10	018936-52-2	9	5	3	1	0	1	0
11	093703-48-1	6	5.1	0	4	0	0	0
12	189084-61-5	6	5.1	0	4	0	0	0
13	068631-49-2	6.4	5.2	0	6	0	0	0
14	000050-29-3	2.4	5.2	0	5	0	1	0
15	000050-32-8	2.3	5.3	0	0	11	0	0
16	000053-19-0	2.4	5.2	0	4	0	1	0
17	000053-70-3	2.3	5.5	0	0	11	0	0
18	000055-38-9	6	5	2	0	0	0	0

Figure 3. Training descriptors used for model development



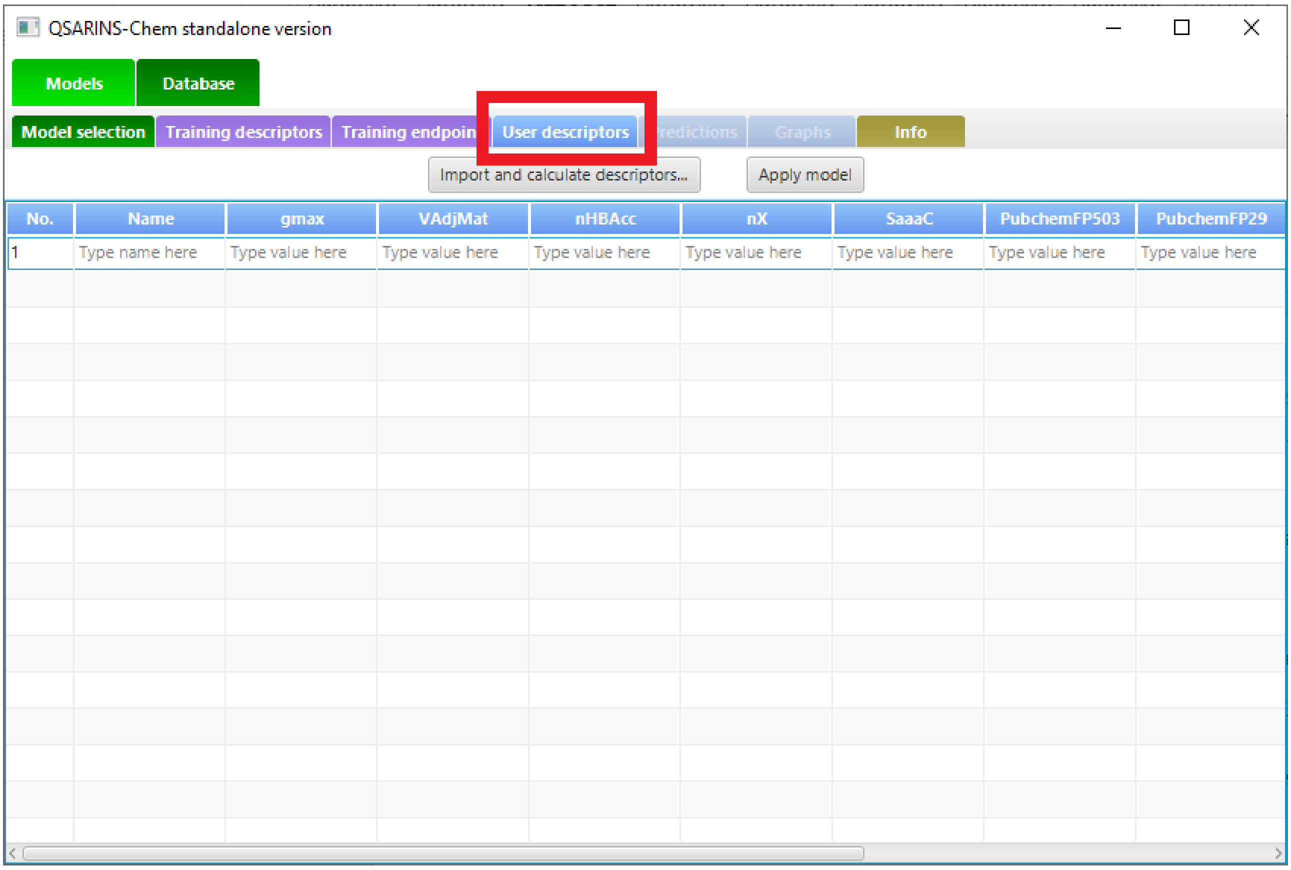
No.	Name	Experimental endpoint	Estimated endpoint	HAT i/i (h* = 4.7e-02)	Residual
1	XXXUNK-46-1	0.34	-0.45	1.3e-02	-0.79
2	XXXUNK-46-2	-0.61	-0.42	7.8e-02	0.19
3	XXXUNK-46-3	8.1e-02	-0.45	1.3e-02	-0.53
4	XXXUNK-46-4	2.6	1.7	3.2e-02	-0.93
5	XXXUNK-46-5	1.7	1.8	3.3e-02	0.10
6	XXXUNK-46-6	2.4	1.8	3.3e-02	-0.68
7	XXXUNK-46-7	0.36	0.60	1.8e-02	0.24
8	XXXUNK-46-8	0.4	0.60	1.8e-02	0.20
9	XXXUNK-46-9	0.29	0.54	1.7e-02	0.25
10	018936-52-2	0.65	0.18	1.3e-02	-0.47
11	093703-48-1	2.3	1.1	6.1e-03	-1.1
12	189084-61-5	1.4	1.1	6.1e-03	-0.29
13	068631-49-2	2	1.5	1.1e-02	-0.51
14	000050-29-3	1.9	2.3	8.5e-03	0.36
15	000050-32-8	4.7e-02	0.15	5.7e-02	0.11
16	000053-19-0	1.7	2.1	8.3e-03	0.44
17	000053-70-3	0.21	0.33	5.4e-02	0.12
18	000055-38-9	-0.36	0.25	5.5e-03	0.61

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.



No.	Name	gmax	VAdjMat	nHBAcc	nX	SaaaC	PubchemFP503	PubchemFP29
1	Type name here	Type value here	Type value here	Type value here	Type value here	Type value here	Type value here	Type value here

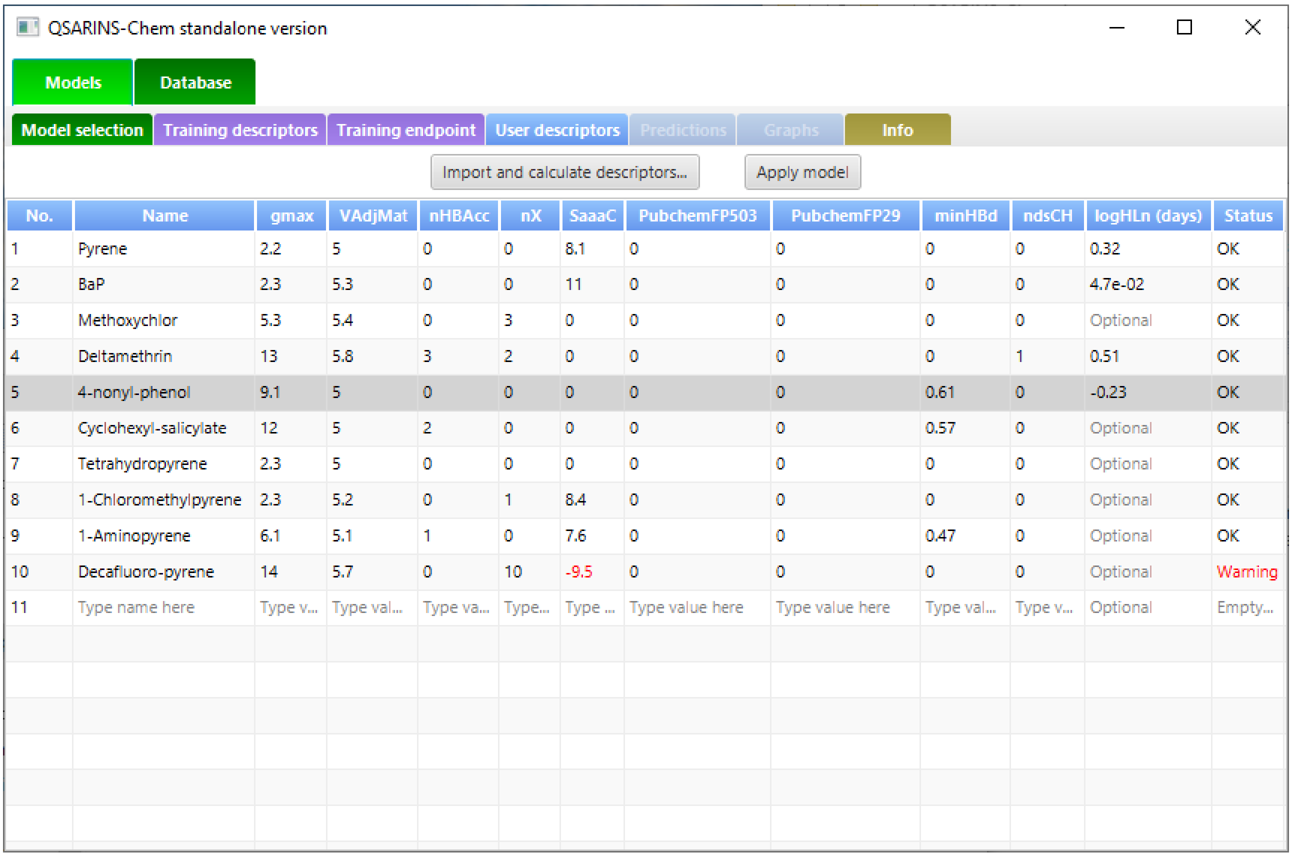
Figure 5. Descriptors calculation

2) **Press** "Import and calculate descriptors": you will be asked to select a folder containing the molecular structures (acceptable file formats must to be checked in the PaDEL-Descriptor software documentation. This software, for descriptors calculation, is freely available and included in the QSARINS-Chem folder, see licence in the "Info" tab for further information). ***As an example for this tutorial, go to the QSARINS-Chem main forlder (it should be "QSARINS_Chem_STANDALONE_v_100"), 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* *an example of calculated descriptors).* ***Note: QSARINS-Chem cannot process molecules with missing descriptors; in this case you have to manually enter the value or delete the chemical (right click and press delete).***



No.	Name	gmax	VAdjMat	nHBAcc	nX	SaaaC	PubchemFP503	PubchemFP29	minHBd	ndsCH	logHLn (days)	Status
1	Pyrene	2.2	5	0	0	8.1	0	0	0	0	0.32	OK
2	BaP	2.3	5.3	0	0	11	0	0	0	0	4.7e-02	OK
3	Methoxychlor	5.3	5.4	0	3	0	0	0	0	0	Optional	OK
4	Deltamethrin	13	5.8	3	2	0	0	0	0	1	0.51	OK
5	4-nonyl-phenol	9.1	5	0	0	0	0	0	0.61	0	-0.23	OK
6	Cyclohexyl-salicylate	12	5	2	0	0	0	0	0.57	0	Optional	OK
7	Tetrahydropyrene	2.3	5	0	0	0	0	0	0	0	Optional	OK
8	1-Chloromethylpyrene	2.3	5.2	0	1	8.4	0	0	0	0	Optional	OK
9	1-Aminopyrene	6.1	5.1	1	0	7.6	0	0	0.47	0	Optional	OK
10	Decafluoro-pyrene	14	5.7	0	10	-9.5	0	0	0	0	Optional	Warning
11	Type name here	Type v...	Type val...	Type va...	Type...	Type ...	Type value here	Type value here	Type val...	Type v...	Optional	Empty...

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.

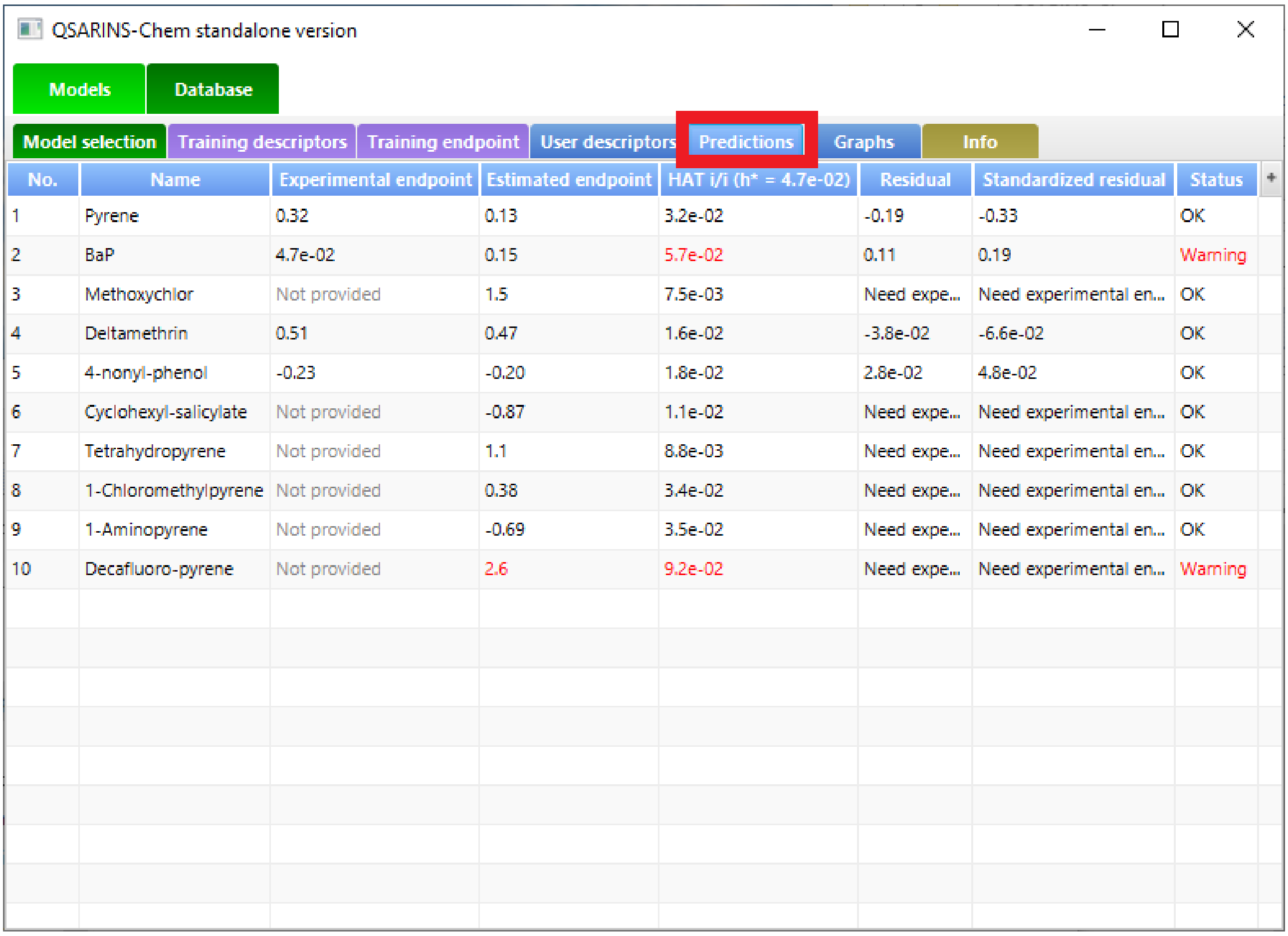
3) **Press** "Apply Model" to run the model and generate predictions.

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**.
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.



No.	Name	Experimental endpoint	Estimated endpoint	HAT i/i (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 expe...	Need experimental en...	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 expe...	Need experimental en...	OK
7	Tetrahydropyrene	Not provided	1.1	8.8e-03	Need expe...	Need experimental en...	OK
8	1-Chloromethylpyrene	Not provided	0.38	3.4e-02	Need expe...	Need experimental en...	OK
9	1-Aminopyrene	Not provided	-0.69	3.5e-02	Need expe...	Need experimental en...	OK
10	Decafluoro-pyrene	Not provided	2.6	9.2e-02	Need expe...	Need experimental en...	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) Insubria 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.

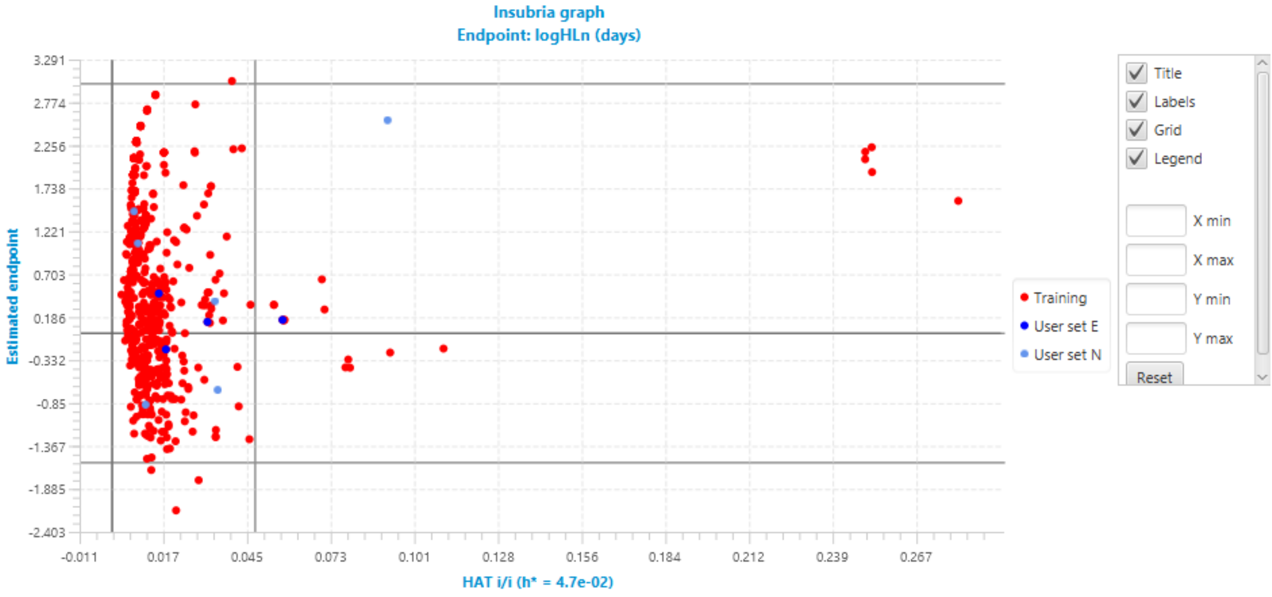


Figure 8. Insubria graph

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.***

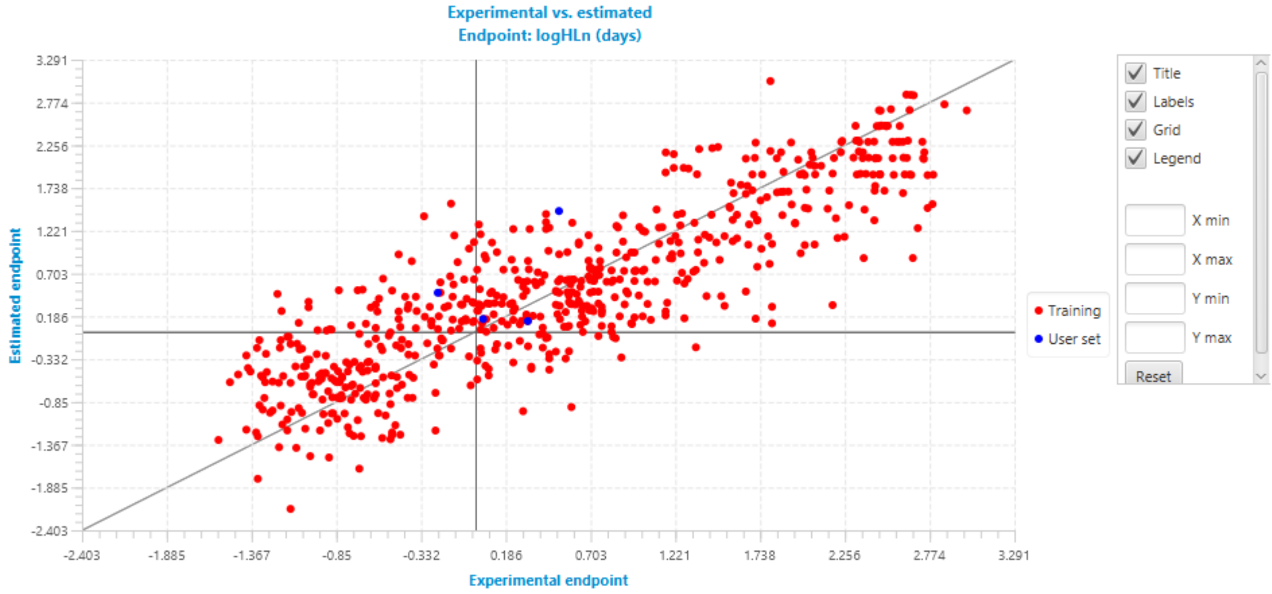


Figure 9. Experimental vs. estimated endpoints

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).

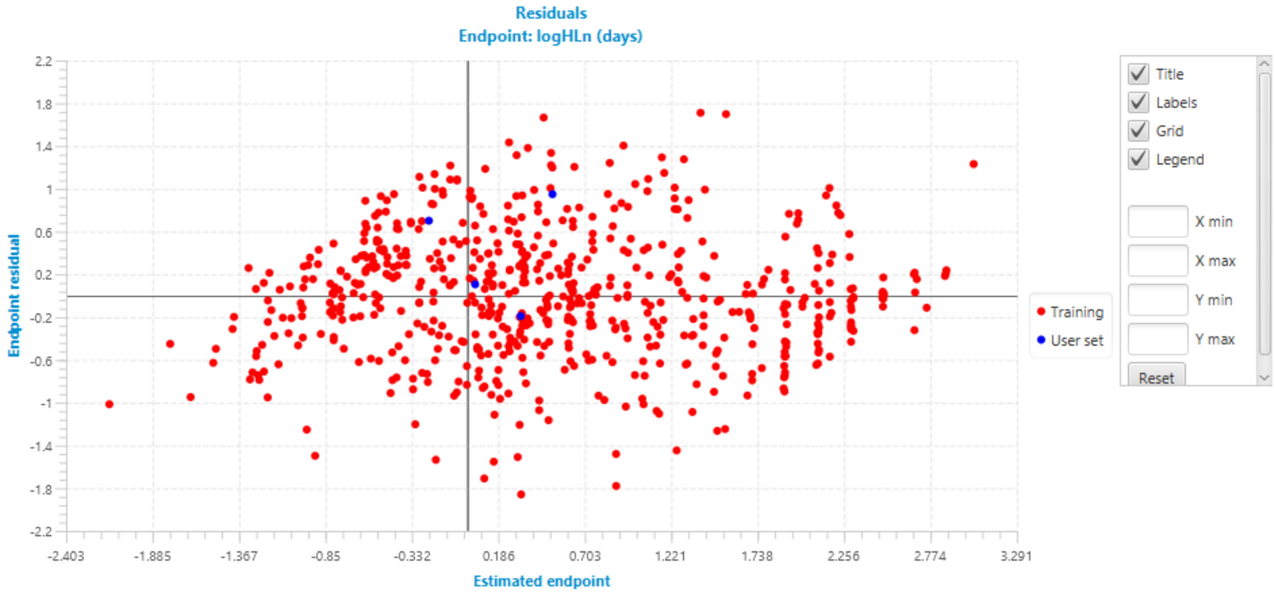


Figure 10. Endpoint residual

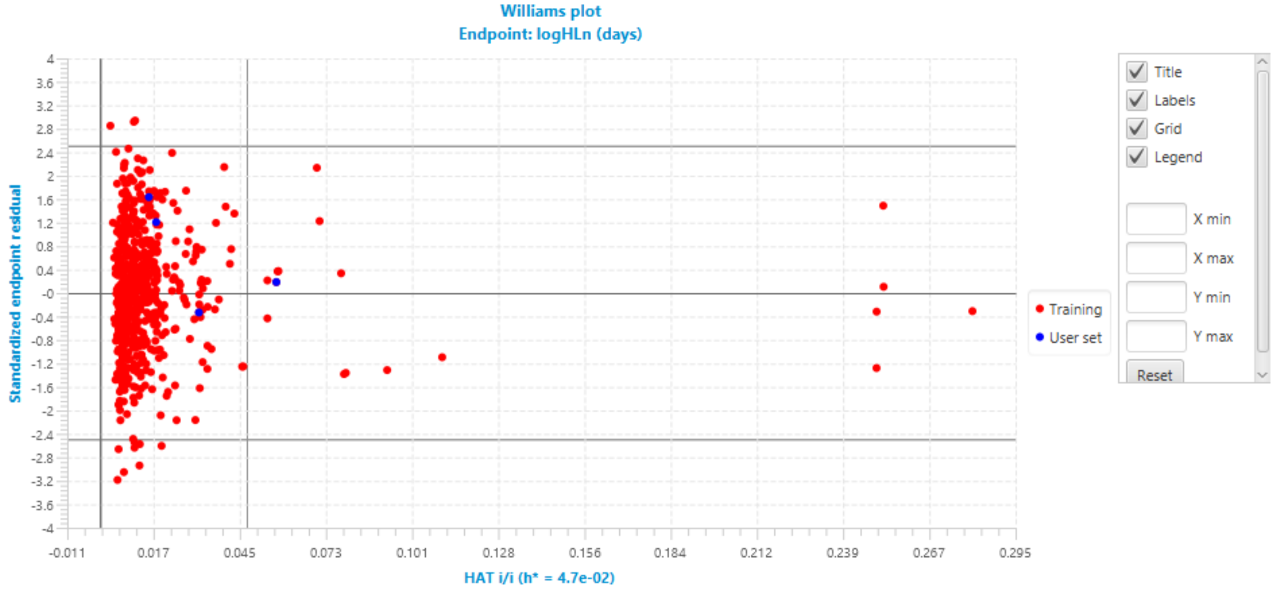


Figure 11. Williams plot.

4) Williams plot. This graph is similar to the Insubria Graph. The difference is in the ordinate axis that reports the standardized residuals of the responses. The Insubria 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 Williams graph 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**