Canadian A.I. Removal Rate Estimator (CAIRRE) User Manual

Version 0.12.0 (beta)

March 2nd, 2023

Environment and Climate Change Canada

Table of Contents

Section 1: User Interface Page

Section 2: Running the Models

Section 3: Formatting Graphs

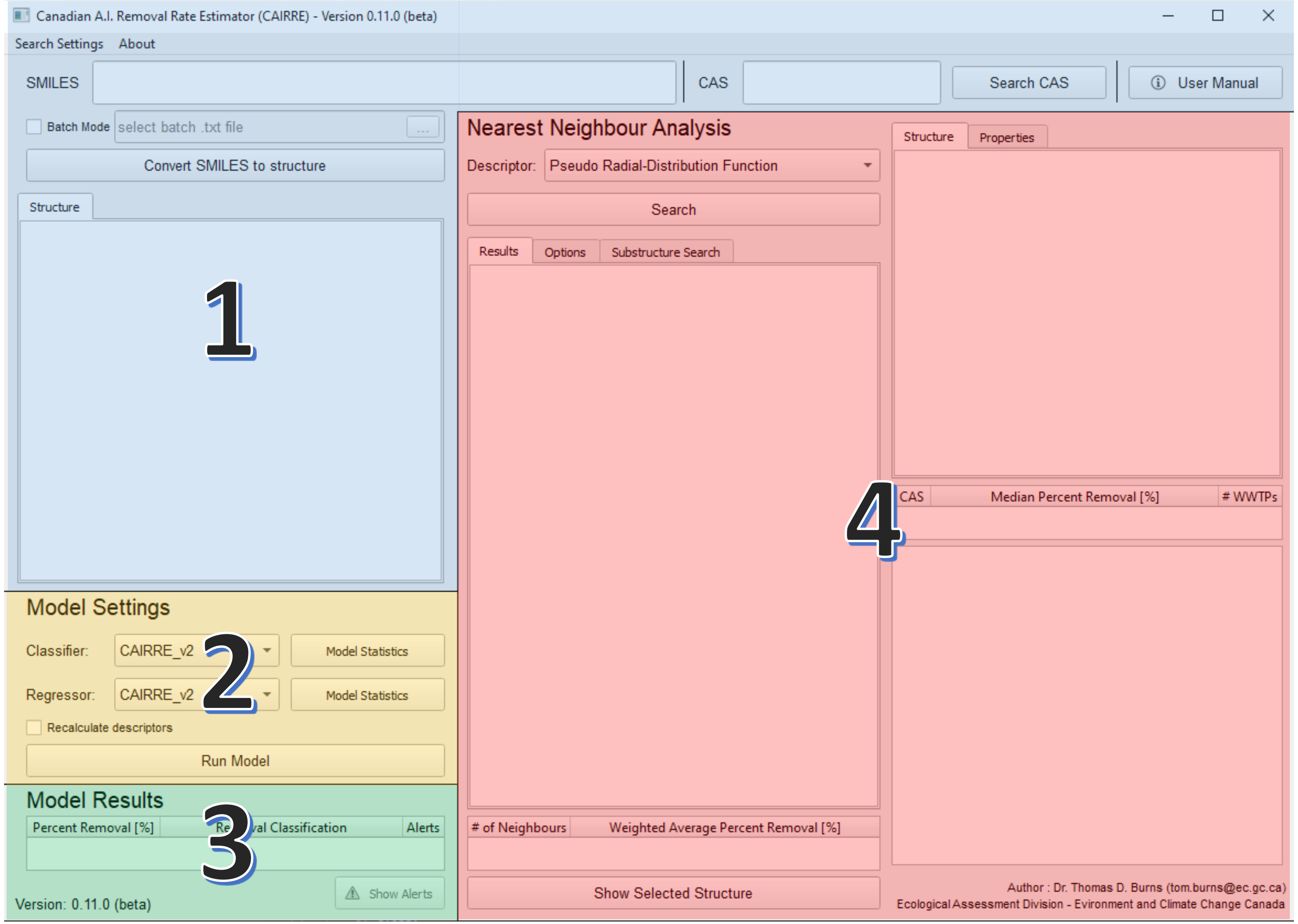
User Interface – Main Window Overview

1 *- User Input*

2 *- Model Settings*

3 *- Model Results*

4 *- Nearest Neighbours*



Main Window: User Input

1 SMILES input line:

User enters a SMILES for the chemical they wish to run through the model

2 CAS input line:

User enters a CAS which searches the selected database for a match. Additional option to perform online search for CAS available.

3 Opens the user manual

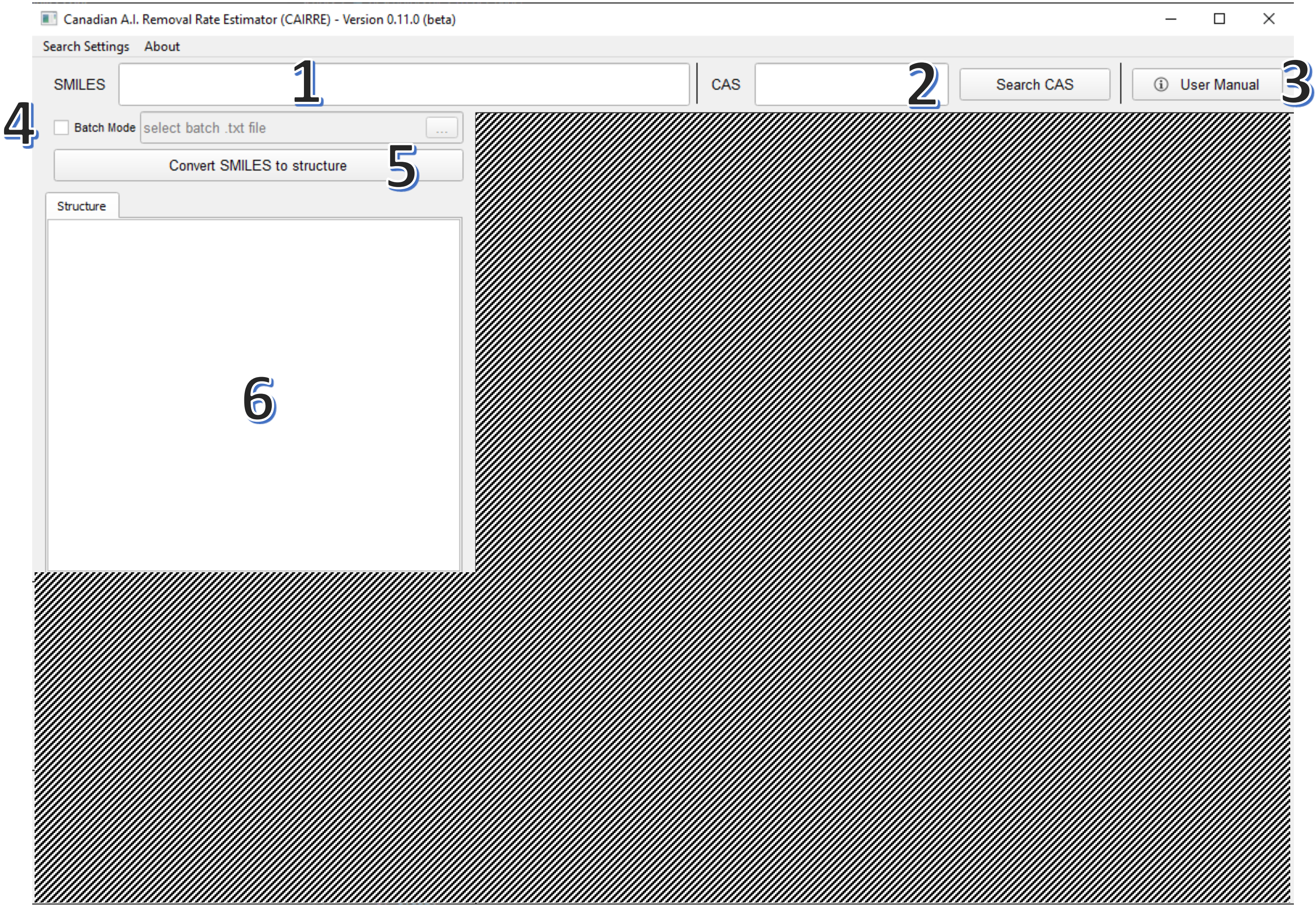
4 Batch Mode Options

Check box toggles Batch mode, once activated the user must select a text file containing the list of SMILES to run through the models.

Note: This disabled Nearest Neighbours

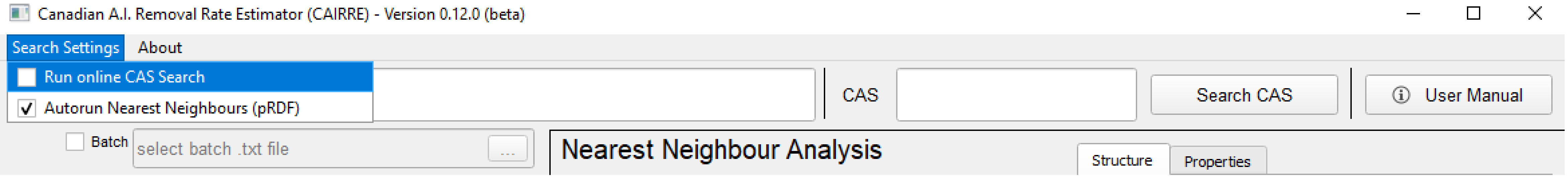
5 Converts the SMILES in 1 to structure

6 Shows the structure of the SMILES in 1



Main Window: User Input (continued)

Two additional search options available on the toolbar:



Canadian A.I. Removal Rate Estimator (CAIRRE) - Version 0.12.0 (beta)

□ Run online CAS Search

☑ Autorun Nearest Neighbours (pRDF)

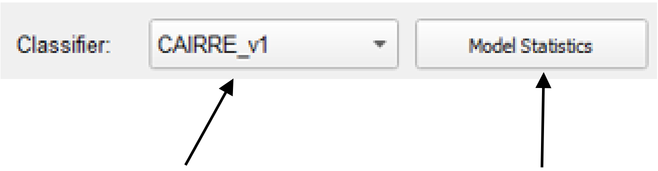
□ Batch select batch .txt file

1. **Run online CAS Search**: allow CAIRRE to search the internet for a SMILES corresponding to a provided CAS. If selected this will automatically be performed when the “Search CAS” button is pressed.

2. **Autorun Nearest Neighbours (pRDF):** automatically run the Nearest Neighbour analysis when the “Search CAS” or “Convert SMILES to Structure” buttons are pressed. This is performed using the pRDF descriptor.

Main Window: Model Settings

1

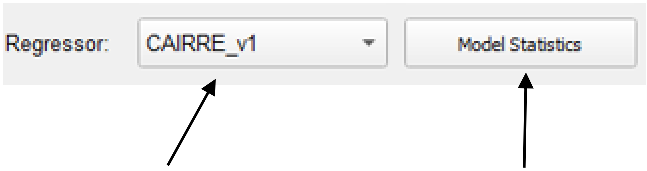


Classifier: CAIRRE_v1

Classifier model selection

Show model training statistics

2



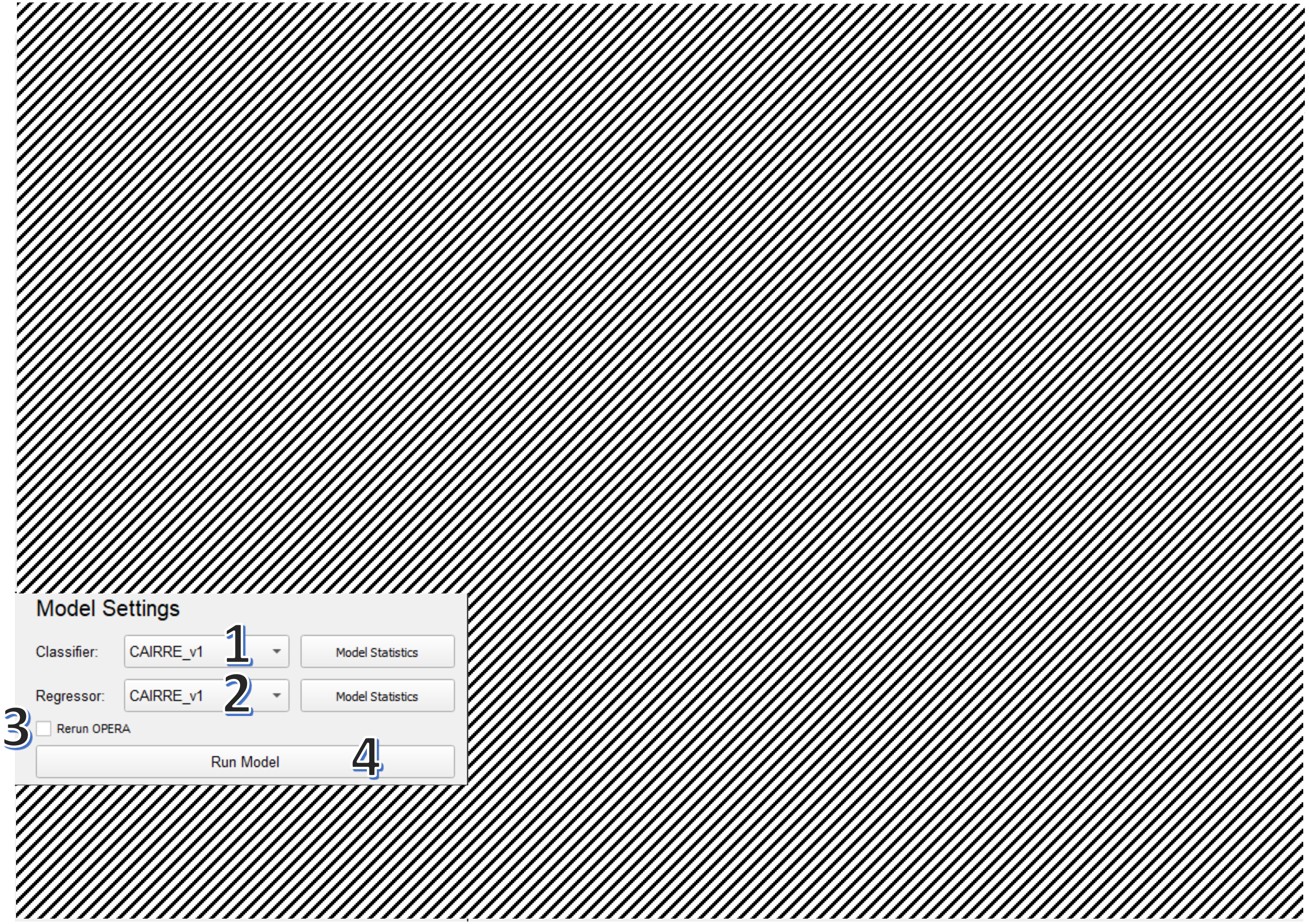
Regressor: CAIRRE_v1

Regressor model selection

Show model training statistics

3 ***Rerun OPERA:*** This option forces CAIRRE to rerun the OPERA calculations from scratch. Use this option when results for CAIRRE read as *Failed*.

4 Run the models



Model Settings

Classifier: CAIRRE_v1

Regressor: CAIRRE_v1

□ Rerun OPERA

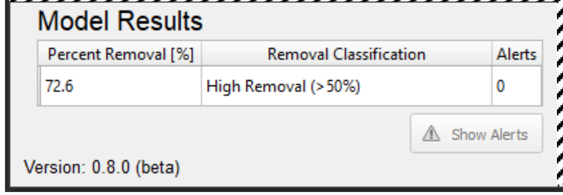
Main Window: Model Results

This section of the main window provides the user with:

- Predicted percent removal
- Predicted removal classification
- Alerts:
  - While running the substance through the A.I. models, CAIRRE will perform a series of checks to ensure that:
    a) The substance is within the effective domain of the models
    b) The Classifier and regressor results do not contradict
- Button to see details on the identified Alerts

Note 1: This section will not be used when batch mode is selected.

Note 2: If *Percent Removal* or *Removal Classification* reads as *Failed*, click ***Rerun OPERA*** option and rerun the model *(see page 6).*



Model Results

Percent Removal [%]	Removal Classification	Alerts
72.6	High Removal (>50%)	0

Version: 0.8.0 (beta)

Main Window: Nearest Neighbour Analysis

1 Select the descriptor to use to check for similarity

2 Run the search
Note: *search is run automatically when models are run*

3 Tanimoto similarity for each substance, a value from 0.0 to 1.0:
1 = Identical, 0 = completely dissimilar

4 SMILES of the substance in the training set

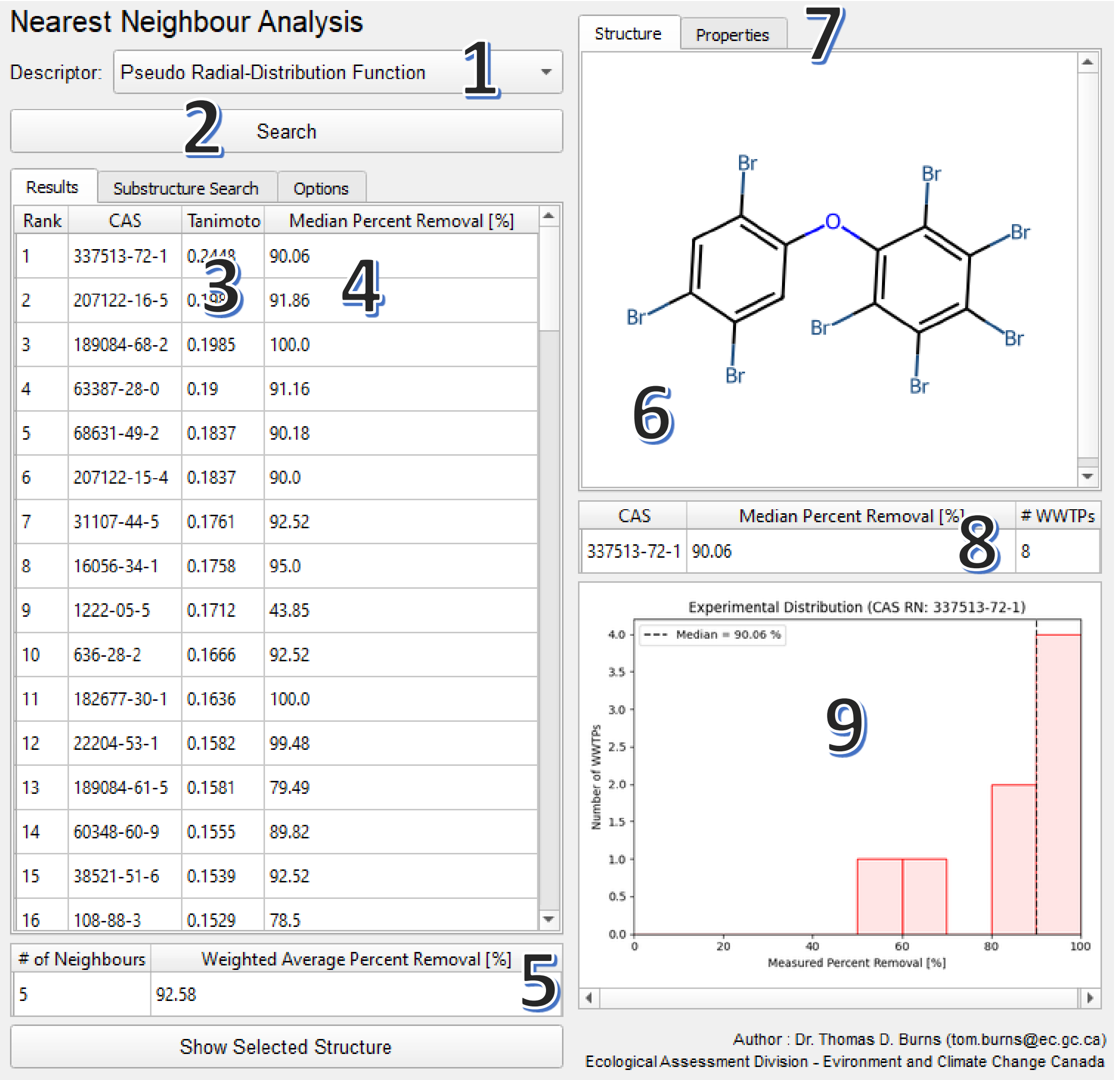
5 A weighted average of N nearest neighbours (can be edited under the *Options* tab)

6 Structure of selected SMILES (default SMILES with highest Tanimoto)

7 Phys-chem properties for the substance calculated using OPERA version 2.7.

8 Information on identity and median experimental removal for the selected SMILES

9 Histogram of the experimental removals for the selected SMILES



Descriptor: Pseudo Radial-Distribution Function

Rank	CAS	Tanimoto	Median Percent Removal [%]
1	337513-72-1	0.2448	90.06
2	207122-16-5	0.1[illegible]	91.86
3	189084-68-2	0.1985	100.0
4	63387-28-0	0.19	91.16
5	68631-49-2	0.1837	90.18
6	207122-15-4	0.1837	90.0
7	31107-44-5	0.1761	92.52
8	16056-34-1	0.1758	95.0
9	1222-05-5	0.1712	43.85
10	636-28-2	0.1666	92.52
11	182677-30-1	0.1636	100.0
12	22204-53-1	0.1582	99.48
13	189084-61-5	0.1581	79.49
14	60348-60-9	0.1555	89.82
15	38521-51-6	0.1539	92.52
16	108-88-3	0.1529	78.5

# of Neighbours	Weighted Average Percent Removal [%]
5	92.58

CAS	Median Percent Removal [%]	# WWTPs
337513-72-1	90.06	8

Author : Dr. Thomas D. Burns (tom.burns@ec.gc.ca)
Ecological Assessment Division - Evironment and Climate Change Canada

Main Window: Substructure Search

CAIRRE also has the option to search the removal database using a specific substructure.

This is a binary search and only returns chemicals containing the full substructure (excluding H atoms).

1 Enter the SMILES representing the desired substructure

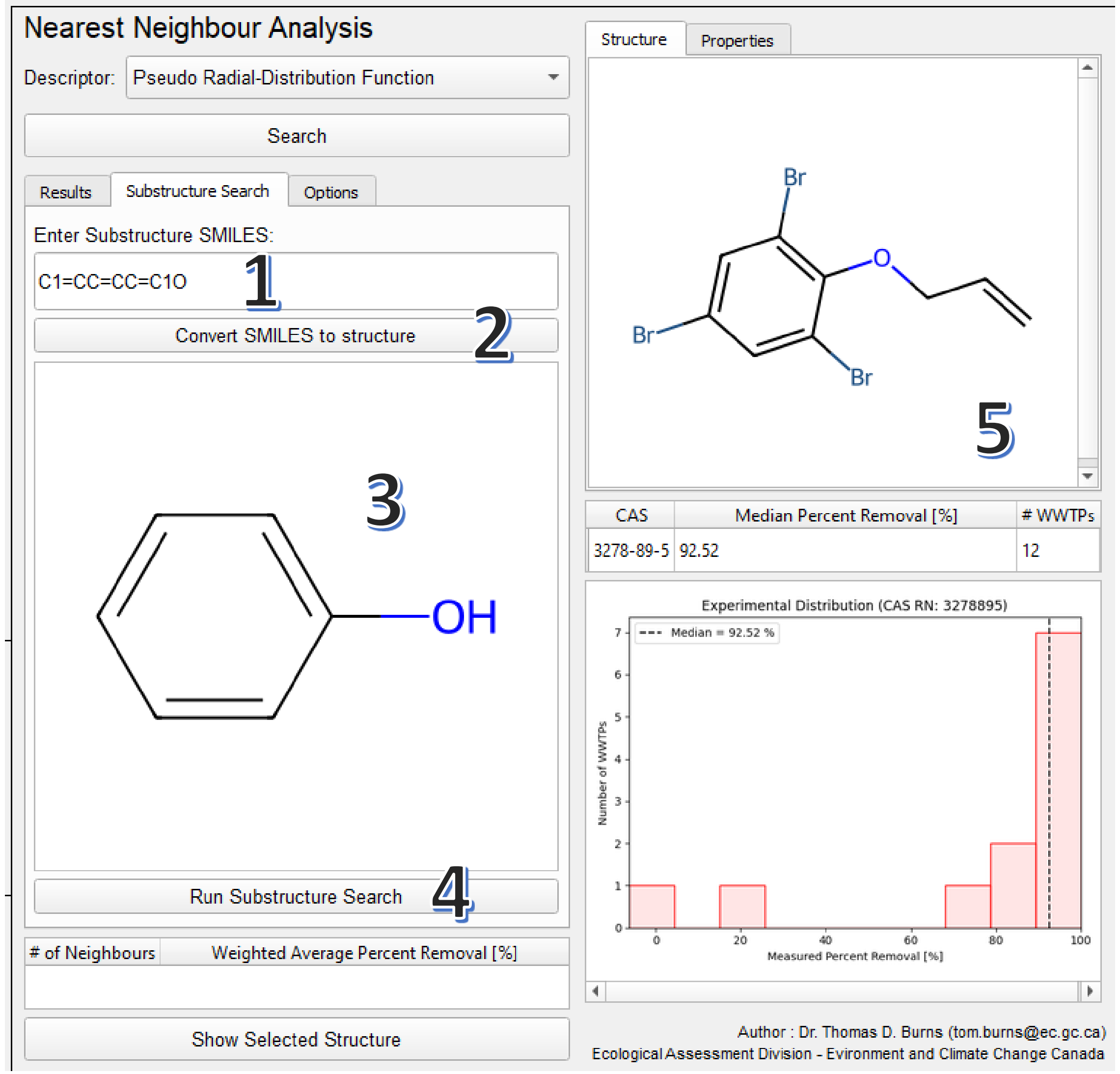
2 Display the chemical substructure entered in 1 (optional)

3 A graphical representation of the SMILES entered in 1

4 Click this button to run the substructure search

5 Information on chemical identified to contain substructure

6 Results window shows all chemicals with substructure in 1, in no particular order



Nearest Neighbour Analysis

Descriptor: Pseudo Radial-Distribution Function

Enter Substructure SMILES:

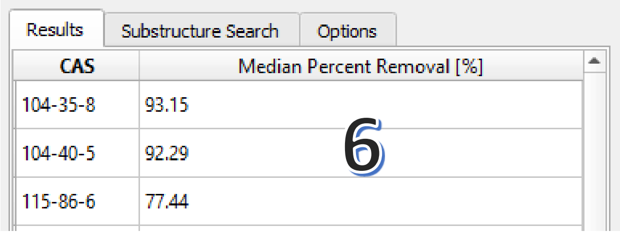
C1=CC=CC=C1O

CAS	Median Percent Removal [%]	# WWTPs
3278-89-5	92.52	12

# of Neighbours	Weighted Average Percent Removal [%]

Author : Dr. Thomas D. Burns (tom.burns@ec.gc.ca)
Ecological Assessment Division - Evironment and Climate Change Canada

***Note**: Substructure search ignores hydrogen atoms, and includes hybridization.*



CAS	Median Percent Removal [%]
104-35-8	93.15
104-40-5	92.29
115-86-6	77.44

Options : Nearest Neighbour Analysis

1 Select the number of nearest neighbours to include in the weighted avearge

2 Set the minimum number of observations of removal efficiency are needed for a chemical to be included in the database query

3 Set the minimum number of WWTPs with experimental removal efficiency observations for a chemical to be included in the database query

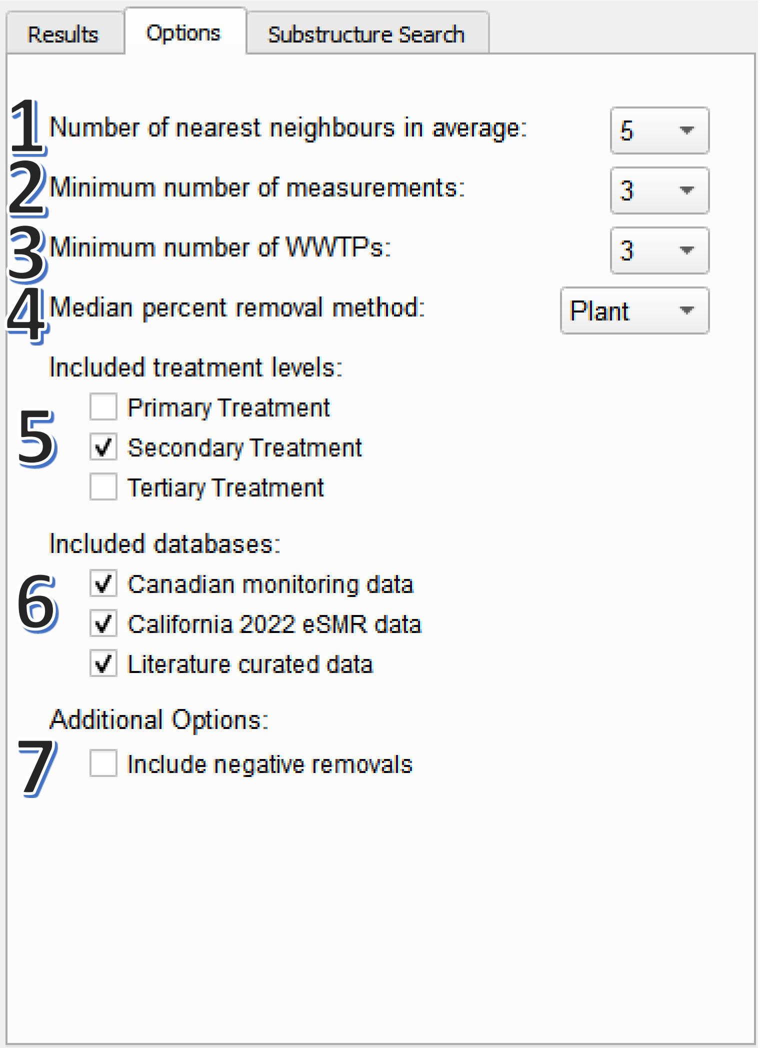
4 How should the median removal efficiency be calculated:

- **Plant:** Median value of the median removal values calculated for each individual WWTP (recommended)
- **Global:** Median of all individual removal observations with no consideration for the WWTP where it was measured

5 WWTP Treatment levels included in the database query

6 Select which database to query

7 Include negative removal efficiencies in the WWTP database queries



1 Number of nearest neighbours in average: 5

2 Minimum number of measurements: 3

3 Minimum number of WWTPs: 3

4 Median percent removal method: Plant

Included treatment levels:

5
- □ Primary Treatment
- ☑ Secondary Treatment
- □ Tertiary Treatment

Included databases:

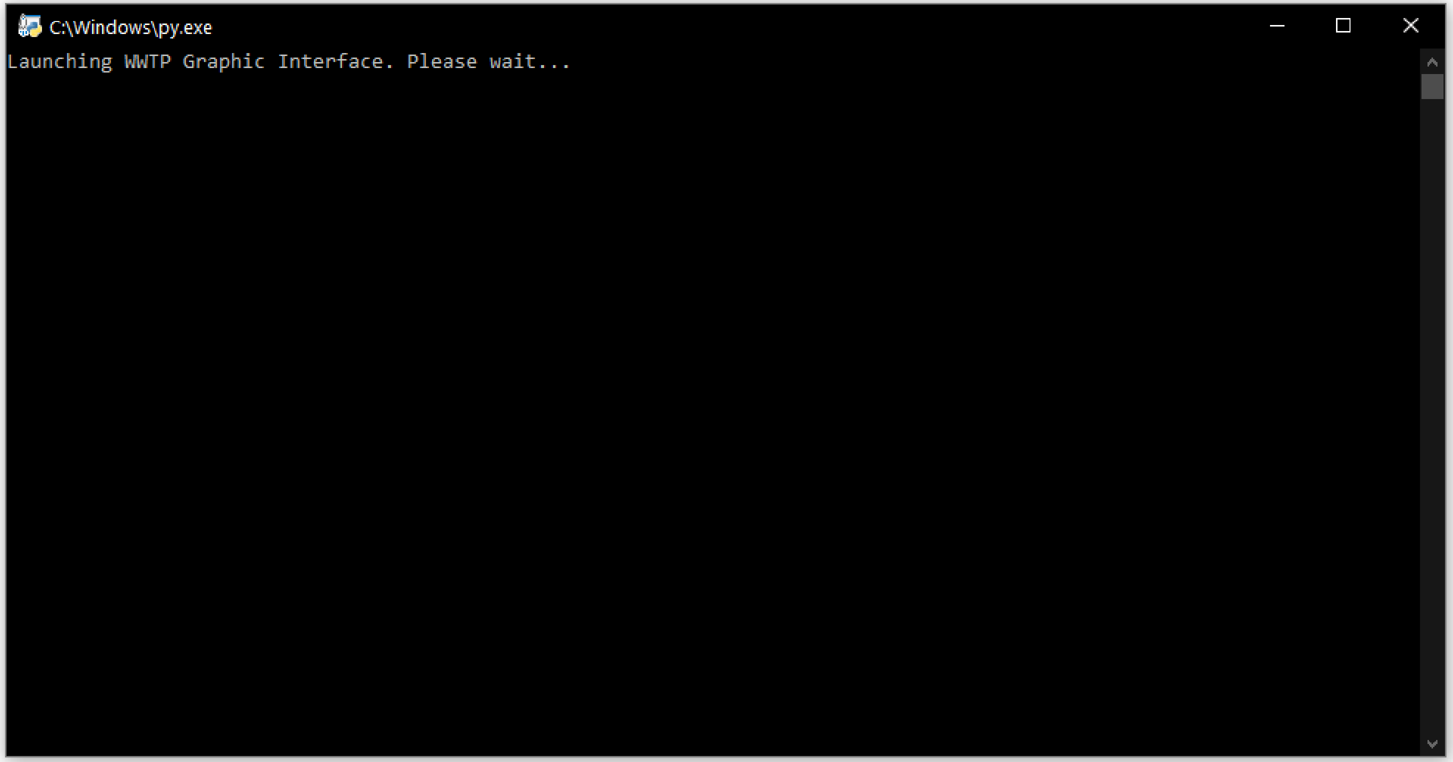
6
- ☑ Canadian monitoring data
- ☑ California 2022 eSMR data
- ☑ Literature curated data

Additional Options:

7
- □ Include negative removals

User Interface – Output Window

After Launching the WWTP program, you may notice a second output window also launched that resembles a command terminal. This window will display the raw output from your calculations, including a progress bar used during the generation of the model descriptors. It will also display error information if any errors arise during the calculation.



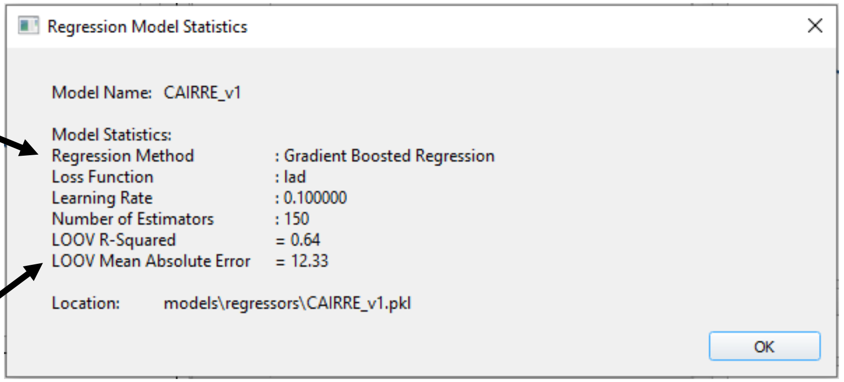
```
Launching WWTP Graphic Interface. Please wait...
```

User Interface – Model Statistics

Each model selected will have separate pop-up windows, one for the regressor and another for the classifier. These windows contain information on the model's training performance, based on Leave-One-Out-Validation (LOOV) results.

Regressor Model Statistics

Model parameters, including model type and the model specific hyperparameters (tuning parameters)



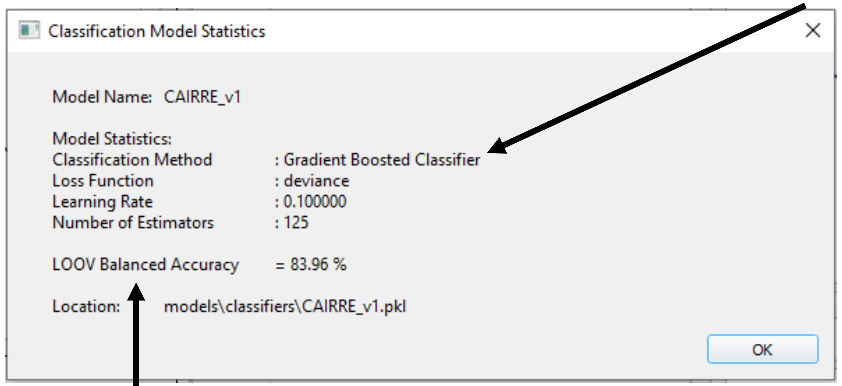
```
Regression Model Statistics

Model Name:  CAIRRE_v1

Model Statistics:
Regression Method              : Gradient Boosted Regression
Loss Function                  : lad
Learning Rate                  : 0.100000
Number of Estimators           : 150
LOOV R-Squared                 = 0.64
LOOV Mean Absolute Error       = 12.33

Location:       models\regressors\CAIRRE_v1.pkl
```

The results of the Leave One Out Validation (LOOV) for the selected model, including the Pearson R^2 and the Mean Absolute Error for the predictions.

Note: Clicking the Regressor's Model Statistics button also provides a scatter plot of the LOOV-validation, showing the predicted vs actual percent removals.

Classifier Model Statistics

Model parameters, including model type and the model specific hyperparameters (tuning parameters)



```
Classification Model Statistics

Model Name:  CAIRRE_v1

Model Statistics:
Classification Method          : Gradient Boosted Classifier
Loss Function                  : deviance
Learning Rate                  : 0.100000
Number of Estimators           : 125

LOOV Balanced Accuracy         = 83.96 %

Location:       models\classifiers\CAIRRE_v1.pkl
```

The Balanced Accuracy value obtained by the model from the LOOV analysis.

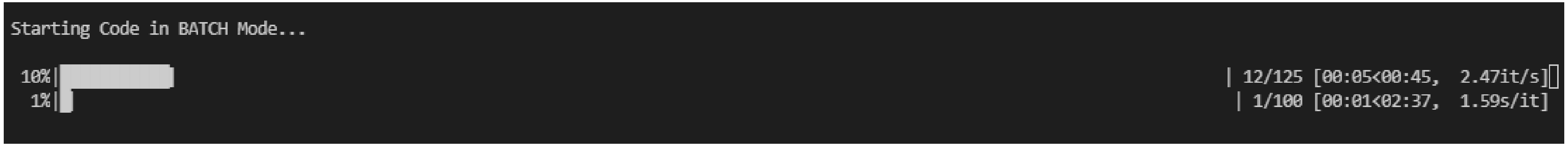
$$Balanced\ Accuracy = 100\%\left[\frac{1}{2}\left(\frac{True\ High}{All\ High} + \frac{True\ Low}{All\ Low}\right)\right]$$

Where:
High >= 50%
Low < 50%

Running the Code – Batch Mode

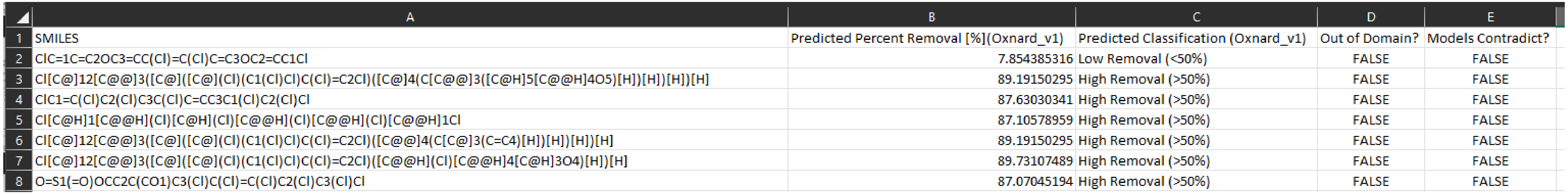
If a large number of substances need to be processed, the WWTP prediction code can also be run in ***Batch Mode***. To run in ***Batch Mode***, toggle the option in the main window, and then press the '…' button (...) to navigate to your text file. This text file should contain the list of SMILES you wish to run, with one SMILES per line. Once the substance is loaded, simply press "***Run Model***".

This calculation will likely take several minutes, the progress bar can be found in the ***Output Window***.



```
Starting Code in BATCH Mode...

 10%|█████                                                          | 12/125 [00:05<00:45,  2.47it/s]
  1%|▌                                                               | 1/100 [00:01<02:37,  1.59s/it]
```

Once completed the code will prompt the user to save the results to a csv file.



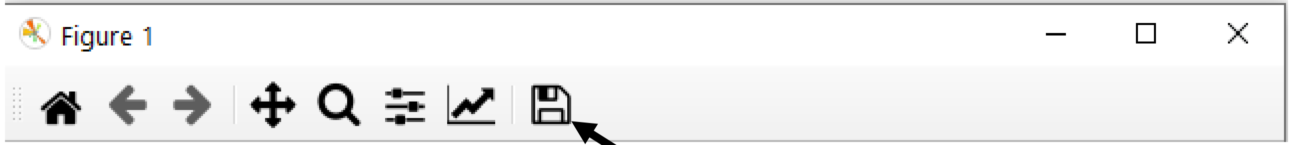
	A	B	C	D	E
1	SMILES	Predicted Percent Removal [%](Oxnard_v1)	Predicted Classification (Oxnard_v1)	Out of Domain?	Models Contradict?
2	ClC=1C=C2OC3=CC(Cl)=C(Cl)C=C3OC2=CC1Cl	7.854385316	Low Removal (<50%)	FALSE	FALSE
3	Cl[C@]12[C@@]3([C@]([C@](Cl)(C1(Cl)Cl)C(Cl)=C2Cl)([C@]4(C[C@@]3([C@H]5[C@@H]4O5)[H])[H])[H])[H]	89.19150295	High Removal (>50%)	FALSE	FALSE
4	ClC1=C(Cl)C2(Cl)C3C(Cl)C=CC3C1(Cl)C2(Cl)Cl	87.63030341	High Removal (>50%)	FALSE	FALSE
5	Cl[C@H]1[C@@H](Cl)[C@H](Cl)[C@@H](Cl)[C@@H](Cl)[C@@H]1Cl	87.10578959	High Removal (>50%)	FALSE	FALSE
6	Cl[C@]12[C@@]3([C@]([C@](Cl)(C1(Cl)Cl)C(Cl)=C2Cl)([C@@]4(C[C@]3(C=C4)[H])[H])[H])[H]	89.19150295	High Removal (>50%)	FALSE	FALSE
7	Cl[C@]12[C@@]3([C@]([C@](Cl)(C1(Cl)Cl)C(Cl)=C2Cl)([C@@H](Cl)[C@@H]4[C@H]3O4)[H])[H]	89.73107489	High Removal (>50%)	FALSE	FALSE
8	O=S1(=O)OCC2C(CO1)C3(Cl)C(Cl)=C(Cl)C2(Cl)C3(Cl)Cl	87.07045194	High Removal (>50%)	FALSE	FALSE

Note: The model used to generate each prediction is included in brackets in the column header, and Out of domain and model contradiction errors are flagged as True in the relevant columns. If the value in these columns is False, that alert was ***NOT*** present for the substance.

Formatting Graphs – Formatting & Saving

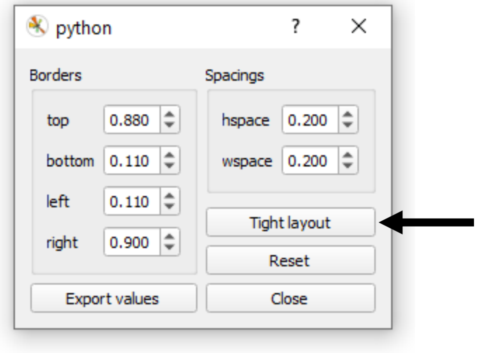
The WWTP prediction code provides the user with several different plots. Depending on the user's screen resolution or aspect ratio, the plots may need some formatting.

Plot Toolbar



Save Plot

Format Settings



Borders

top 0.880

bottom 0.110

left 0.110

right 0.900

Spacings

hspace 0.200

wspace 0.200

Click "Tight Layout"