QSRR Automater Instruction Manual

**Introduction:**

QSRR Automater is a program designed to automate generation of QSRR models to predict retention times of compounds on chromatographic columns. The software is available in two forms, a .exe for those who wish the most ease of use and the source code for those who wish to adjust and modify the code in any way. Instructions on installation, use, and troubleshooting are below.

**Installation:**

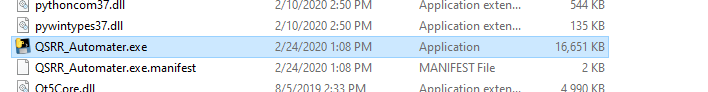
*.exe version:*

The following are required

1. Windows operating system
2. .exe file in its folder

The windows operating system is the only operating system the software has been tested on. While Apple and Linux systems may work, neither has been attempted so errors are likely.

Download the .exe from https://github.com/UofUMetabolomicsCore/QSRR\_Automator. It is in the form of a zipped folder. Simply download this folder and put it wherever you like in your computer. It will need to be unzipped to work. If you right-click on the folder and hover over 7-zip. This will expand a window to the side. Select either the “Extract files…” or “Extract Here” option. If your version of windows is older and does not provide a 7-zip option, zipping/unzipping software can be found relatively easily.



When opened the folder will have many files including QSRR\_Automater.exe shown above. The .exe should stay in this folder. If you wish a way to access the .exe without going into the folder, right click on the .exe and select “Create shortcut”. This will create a small file called “QSRR\_Automater.exe – Shortcut”. This shortcut can be moved wherever you wish and renamed if desired.

*Python code version:*

Software was coded using python 3.7. The following are required packages:

* mordred
* rdkit
* pyqt5
* scikit-learn

The following are also required but are likely to be installed as dependencies of the above:

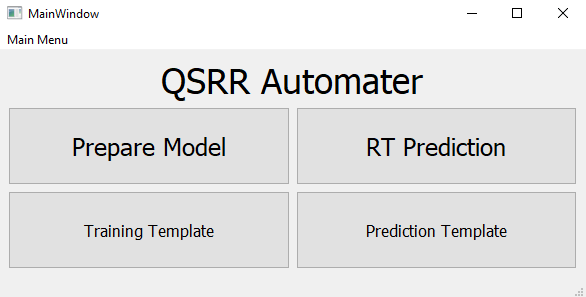
* pandas
* scipy
* numpy
* matplotlib

The only dependency that might cause issues is rdkit. Rdkit is difficult to install by pip though if you wish to do so the instructions are at <https://www.rdkit.org/docs/Install.html>. The easiest way to install rdkit is by using anaconda so the standard “conda install rdkit” works fine. For those looking at the installation link above who are not familiar with anaconda, it recommends creating a new environment for rdkit. That is not necessary, a standard install command works fine so long as you use “conda” instead of “pip”. If you do choose to create the environment, simply remember to activate it before attempt to run QSRR Automater.

The source code and all required files are at https://github.com/UofUMetabolomicsCore/QSRR\_Automator

**Basic Functions**

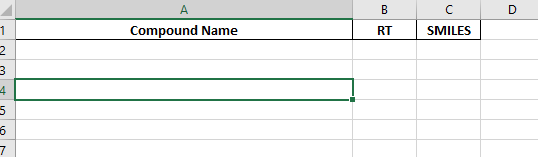
To open QSRR Automater double click on the .exe or run \_\_main\_\_.py from your command line. This will open a window that looks like this:



**Training a Model**

*Put Data into Proper Form*

First the data must be put into the proper form. The Training Template button will provide a way to save a .xlsx or .csv file with whatever name is desired. This template file looks like this:



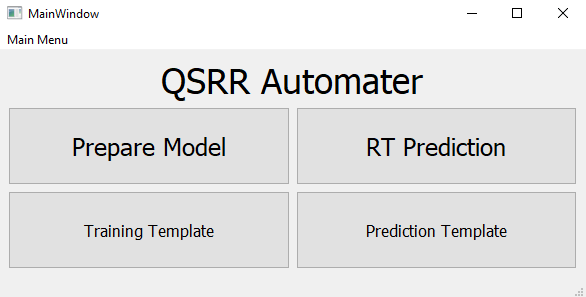
It must be filled out to be used by QSRR Automater. “Compound Name” can be any identifier for your compounds. The software just uses this as a tag so as long as each compound name is unique, they can be whatever you want. “RT” stands for Retention Time, the time that the compound was observed. This is unit agnostic, so long as the units are all the same they can be in minutes, hours or seconds and QSRR Automater will not care. Finally, the “SMILES” column. This is the structure for the compound in the SMILES format. If you do not know the SMILES notation for your compounds, several publicly available resources such as the pubchem website or the human metabolome database have SMILES for most known compounds. Do note that RT and SMILES can have duplicate values, only the Compound Names must be unique.

If you have extra columns present they will be assumed to be descriptors. When you are creating the model (detailed below) you will be given the option to keep these descriptors or to drop and re-calculate the usual QSRR Automater descriptors. The descriptors must be numerical or they will be uninterpretable, so if you have descriptors that are text (true false values, names of compound classes, etc.) represent the unique text as a numerical value. If they are true false, 1 and 0 work well. If they are groups, if there is a clear relationship (number of aromatic rings or similar) providing a number is good. If the values are non-connected classes it is often best to represent each separate possibility as a true false column. So if it could be A, B, C or D, make all four into columns and then in the A column put a 1 for every compound that is an A and 0 for all the others. Repeat for all options.

You must have at least 50 compounds for a model to be created, though 100-200 will be more accurate if possible.

*Create the Model*

Push the “Prepare Model” button on the main window:

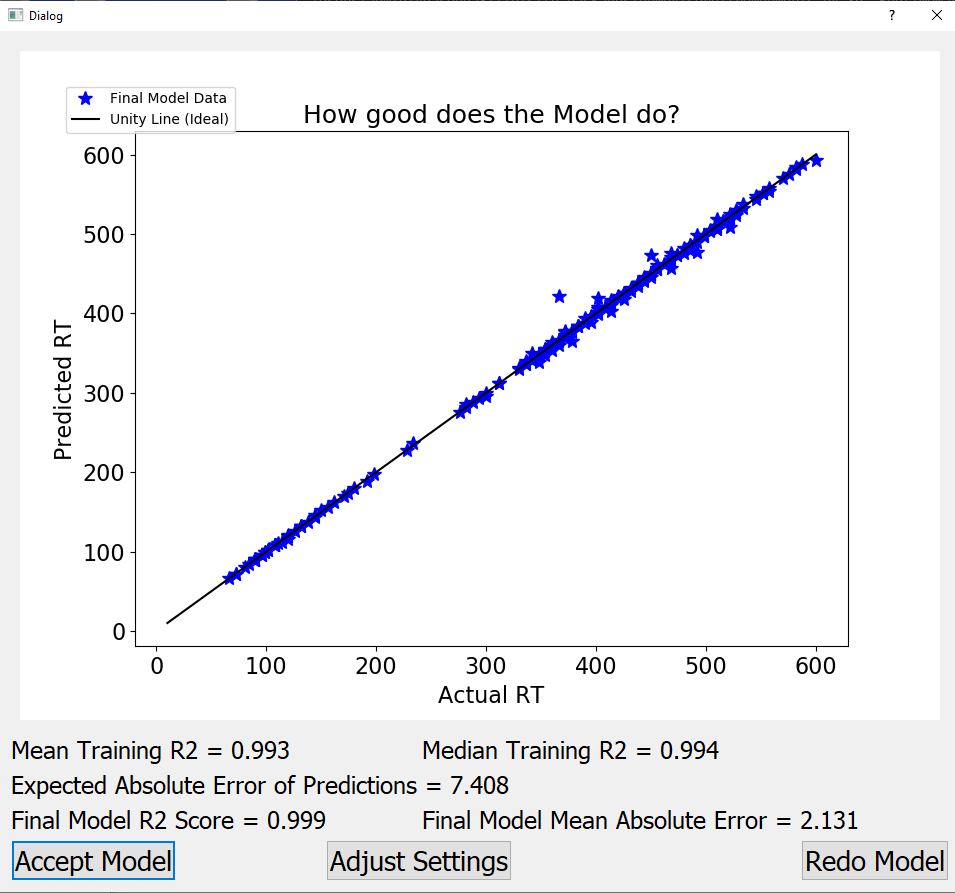


An option will appear asking if you wish to create a new model or load an old one. We will first go over creating a new model.

*Creating a New Model*

You will be asked to select a file to train the model on. Choose a model in the style of the Training Template as described above. If you have extra columns in your input template, they will be assumed to be descriptors for the prediction and you will be asked if you wish to use your provided descriptors or drop them and have QSRR Automater calculate new descriptors. Choose whichever option you feel appropriate to your data set.

QSRR Automater will begin analyzing your data. This will take some time during which the program will not respond. There may be some progress shown on the command prompt that opens along with the software, but this may take a while as well. Eventually a window will pop up:



This is your completed model. The graph shows how your compounds Actual RT provided by you, match the final Predicted RT. The statistics on the graph are the lowest values provided “Final Model R2 Score” and “Final Model Mean Absolute Error”, which are the r2 of the line, and the mean absolute error the average error values of all points (in the above case 2.131 seconds).

While the Final model usually looks good, it is often deceptive since the final model uses all the data for model construction. The performance on unobserved data (and thus how it will likely perform on new data) is represented by the other three numbers. Multiple models (default 5 but this can be changed in the settings menu) were created while leaving out a test set to try the new model on. The results are the average and median of the r2 of their fits (Mean Training R2 and Median Training R2) and the average absolute error of predictions (Expected absolute Error of Predictions). The most useful is likely the Expected Absolute error of predictions, since in testing 2/3-3/4 of truly new data in the test set will be accurate to within this value. If it is too high you may wish to adjust the settings or improve the training set.

You must give a response to proceed. The Accept Model button accepts this model, which QSRR Automater will save for later use. Adjust Settings allows access to a limited settings menu, so settings can be altered if desired. Redo Model will just redo the analysis, useful if you adjusted the settings.

After the model is saved you will be informed with a message.

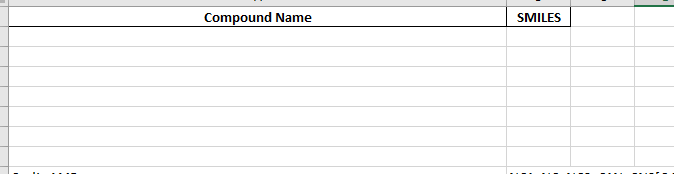
*Loading a Model*

If you have already created a model you can press the Prepare Model and choose the “Load model” button when given the option. You will be prompted to select the folder you put the model into last time. This will be your default folder if you have not adjusted the folder, or the folder you used using the “Choose Output Folder” in the pull down menu. You will be told if there was a problem loading the model or if this was successful.

**Using the Model**

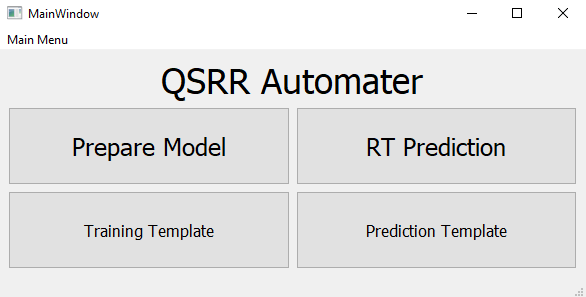
*Putting Compounds in the Proper Format*

Like the Training of a model there is a “Prediction Template” button which will prompt you to save the template as an excel or .csv file. The file will look like this:



As above, the Compound Name can be any identifier so long as it is unique and SMILES are the structure using the SMILES system. Also as above, you may include extra columns which will be assumed to be descriptors. Do note that these descriptors must at least contain the descriptors used in your final model or problems will result (so use whatever descriptors you used in the model training set). Also none can be named “RT” as doing so will make QSRR Automater assume that you are attempting to put in a training template.

*Predicting Retention Times*



Push the RT Prediction button. This will use the currently loaded model, the most recent successful use of the Prepare Model button whether this was creating a new model or loading a previously created model.

If there is not a model loaded, you will be told and will not be allowed to proceed until a model is created or loaded. As before if you provided extra columns in your input you will be asked to keep or drop them. You will be told if there is a problem and when the predictions are complete. This is much faster than preparing a new model.

**Outputs**

Before we discuss outputs be aware that there are several settings we have not discussed yet which change which output files are produced or what form they take. These will be noted in the appropriate section.

All files will be the folder pointed to in the “Choose Output Folder” option.

*Training Outputs*

Training Processing Steps.xlsx – created if the excel output option is selected in the setting menu. This is a collection of the initial input and the rest of the outputs on separate tabs for easy storage.

Successful\_Graph.pdf – the final graph of the successfully trained model from the “How good did the model do?” window so it can be referenced later.

saved\_model.joblib – this is the actual saved model in a form QSRR Automater can easily load.

saved\_model\_settings.csv – contains some basic information on the saved model, such as the machine learning model type used, how descriptors were selected, and the file path of the training model used to create the model. Is present in Training Processing Steps.xlsx as the “Model Settings” tab.

chosen\_features.csv – if feature selection of any sort is used, this file is created. It is a list of the descriptors (features) actually chosen. Is present in Training Processing Steps.xlsx as the “Selected Features” tab.

cross\_validation\_and\_final\_sores.csv – has values from the grid searches, a process of splitting the data into a training and test set, training on the training set and analyzing performance on the test set. This is done repeatedly. The results on the test sets and the final results of the models are save here. Is present in Training Processing Steps.xlsx as the “Model Score Values” tab.

Descriptors\_for\_training – this will be a .csv or a .xlsx file depending on the settings menu. This file contains your input file with values for all potentially usable descriptors calculated by QSRR Automater. Is present in Training Processing Steps.xlsx as the “Final Descriptors Output” tab. This can be useful for adding new descriptors to the QSRR Automater provided descriptors and making a new model. Is also used to check that a loaded model has not been altered in any way

Model building settings.csv – this contains all the settings set in the settings menu for easy reproducibility or information if a redo is needed. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in Training Processing Steps.xlsx as the “User Settings” Tab.

Model building programmer settings.csv – this contains some settings that should never need to be changed by the program user, but are recorded in case a change is needed. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in Training Processing Steps.xlsx as the “Programmer Settings” Tab.

Model building numerical coercion.csv – when creating descriptors for the model, or especially when loading descriptors from the user, all values must be numerical. Any values that cannot be converted, usually text strings, will be turned into “nan” (not a number) values. This holds the output of that coercion. Linked to SMILES not compound name. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in Training Processing Steps.xlsx as the “Added Descriptors” Tab.

Model building replicates removed.csv – the next step in feature selection is the removal of descriptors that are too similar. This is the results after that removal. Linked to SMILES not compound name. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in Training Processing Steps.xlsx as the “Replicates Removed” Tab.

Model building bad samples removed.csv – this is the feature selection after removed samples with too many missing values in the descriptors will be removed. Linked to SMILES not compound name. This is the result. If excel is selected it is present in Training Processing Steps.xlsx as the “Bad Samples Removed” Tab.

Model building Final Descriptors.csv – the final descriptors used. Similar to Descriptors\_for\_training except linked to SMILES not compound names. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in Training Processing Steps.xlsx as the “Final Descriptors” Tab.

*Testing Ouput*

RT Prediction Steps.xlsx - created if the excel output option is selected in the setting menu. This is a collection of the initial input and the rest of the outputs on separate tabs for easy storage.

Descriptors for prediction - this will be a .csv or a .xlsx file depending on the settings menu. It contains the calculated descriptors for the compounds in your input file for prediction. If excel is selected it is present in RT Prediction Steps.xlsx as the “Final Descriptor Output” Tab.

Final RT Prediction - this will be a .csv or a .xlsx file depending on the settings menu. It contains your prediction input with the predicted retention times added as the “RT” column. If excel is selected it is present in RT Prediction Steps.xlsx as the “RT prediction” Tab.

Full Error File - this will be a .csv or a .xlsx file depending on the settings menu. This is only created if there is an error in assigning descriptors. It is the descriptors with True and False values instead of numerical values. TRUE indicates a problem. This does not mean the model will not predict, but can indicate that the compound is different from the compounds used to make the model and so may have more error in prediction

Relevant Error File - this will be a .csv or a .xlsx file depending on the settings menu. This is similar to the “Full Error File” but indicates a true problem. This has only the descriptors that are needed for the training model, and one of your compounds has failed. The “TRUE” value which compound and which feature have the problem.

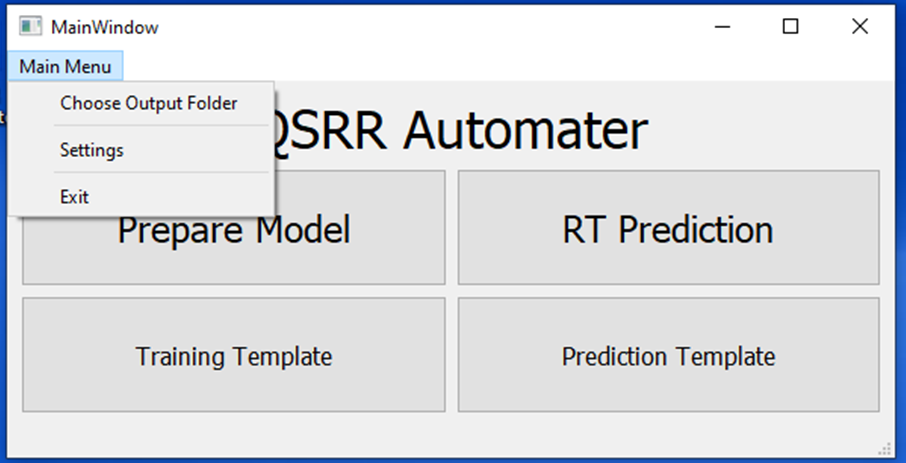
RT prediction settings.csv - this contains all the settings set in the settings menu for easy reproducibility or information if a redo is needed. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in RT Prediction Steps.xlsx as the “User Settings” Tab.

RT prediction programmer settings.csv - this contains some settings that should never need to be changed by the program user, but are recorded in case a change is needed. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in RT Prediction Steps.xlsx as the “Programmer Settings” Tab.

RT prediction numerical coercion.csv - when creating descriptors for the model, or especially when loading descriptors from the user, all values must be numerical. Any values that cannot be converted, usually text strings, will be turned into “nan” (not a number) values. This holds the output of that coercion. Linked to SMILES not compound name. This is only produced if csv is selected for “Excel or csv output setting”. If excel is selected it is present in RT Prediction Steps.xlsx as the “Added Descriptors” Tab. Note the trimming of descriptors from training does not occur so the removal and bad sample removal does not occur (so their files are not created).

**Pull down menu**

In the top left of the interface there is a pull down menu shown below



*Choose Output folder*

Since there are so many outputs all will be placed in a folder. By default, this will be a folder next to your QSRR Automater folder created by QSRR Automater called “QSRR\_Automater\_Output”. So if the QSRR\_Automater.exe file (or .py file) is in “C:\Science Software\QSRR\_Automater” QSRR Automater will attempt to create “C:\Science Software\QSRR\_Automater\_Output” and will default to this. If you select “Choose Output Folder” you will have the option to choose a new folder to put your outputs into.

*Change Settings*

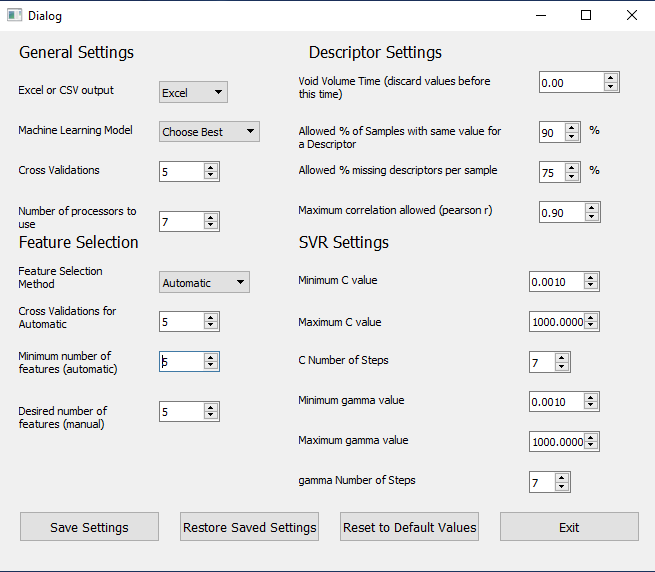
Opens a settings menu to adjust basic settings. This requires a large explanation so is in its own section later

*Exit*

Close the program

**Settings Menu**

*Main Settings Menu*



The settings are spit into various sections. The General Setting section has the following options

Excel or CSV output – this has the options of “Excel” or “CSV”. This changes if the output files are .xlsx files or .csv files. How this affects individual files in the “Outputs” section above.

Machine Learning Method – allows specifying which machine learning model should be used. “Choose Best” will try all and attempt to choose the best. “MLR” is multiple linear regression, “Random Forest” is Random Forest Regression, and “SVR” is using Support Vector Machines for regression.

Cross Validations – a cross validation takes your training set and splits it into a training set and a test set to see how the final model can be expected to do on predictions (this allows all of the data to be used in the final model). Fewer can free up processors or speed up calculation while more may take more time but is more accurate. Standard is 5 or 10.

Number of processors to use – how many processors should be used for calculation. One will be assigned to each cross validation. It is recommended to use at least one less processor than your computer has available (the default setting is number of processors on the computer -1) so that the computer can still function at a reasonable speed during calculation. If the processors exceed the number of cross-validations nothing is gained. If the number of processors is less than the number of cross-validations analysis the processors will keep performing a new cross-validation as soon as they are done with their current one until all the cross-validations are complete.

The Feature Selection settings are based on how features (descriptors) are selected

Feature Selection Method – this determines the criteria used to select features (descriptors). “Automatic” allows a random forest algorithm to determine the number of features that is based. “Manual” uses a random forest algorithm to select a specific number of features. “None” uses all features. It is recommended that “None” not be selected unless you are providing the features, since many irrelevant features are calculated by QSRR Automater and this can lead to overfitting unless they are trimmed in some way.

Cross Validations for Automatic – “Automatic” in the Feature Selection Method uses cross validations like the general analysis. For each cross validations in the General Settings “Cross Validations” setting this number of feature selection cross validations will be done.

Minimum number of features (automatic) – the minimum number of features (descriptors) allowed to be selected for the “Automatic” Feature Selection Method

Desired number of features (manual) – the number of features (descriptors) that will be selected in the “Manual” feature selection method option

The Descriptor Settings involve how the descriptors (features) are selected.

Void Volume Time (discard values before this time) – samples eluting too early can cause problems in prediction. If the void volume is a concern set a value and all times less than that in the training set will be removed.

Allowed % Samples with same value for a Descriptor – if this percent of the training samples have the same value for a descriptor it is unlikely to be predictive or sufficiently represented to be accurately modeled and so is removed

Allowed % of missing descriptors per sample – if the if a sample has this percent of features (descriptors) it is considered unpredictable by these descriptors and so the compound is removed.

Maximum correlation allowed (pearson r) – if two features (descriptors) have this much or more correlation between them for these compounds (according to pearson r) they are assumed to be measuring essentially the same thing for these compounds so only one is retained.

SVR settings manage some values needed for SVR with an ‘rbf’ kernel.

C and gamma are variables used for SVR with an ‘rbf’ kernel. Each can have many values but checking too many will slow down calculations. Therefore, the values are limited to a number of steps between a minimum and maximum values. Minimum and maximum are included. So the default of 7 steps with a minimum of .001 and max of 1000 translates to the values .001, .01, .1, 1, 10, 100, and 1000.

Buttons on the bottom of the menu.

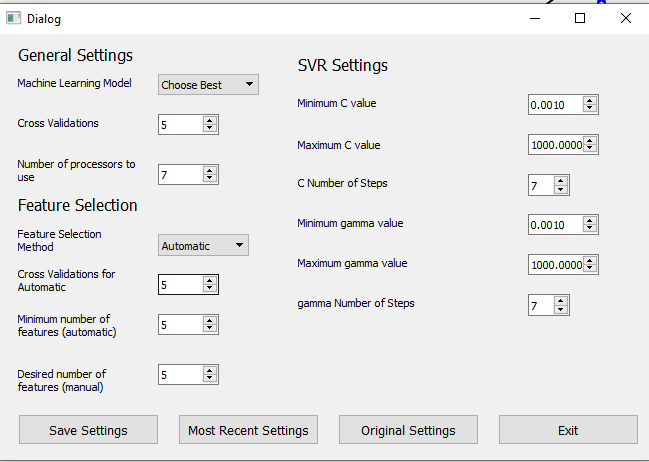
Save Settings – saves the current settings

Restore Saved Settings – reset the settings to the last saved settings

Reset to Default Values – reset to the default values when program was originally opened.

Exit- exits the settings menu. Will prompt to save if settings are not the same as the last saved settings

*Simplified Settings Menu*



This menu is only accessible after a model has been trained and you are choosing to keep or save the model. This adjusts the setting for a redo. Changes from the normal settings menu include:

Excel or CSV output is not present nor are the settings for Descriptors since it is too late to change these things.

Original Settings Button will change the settings to the those used to create the original model

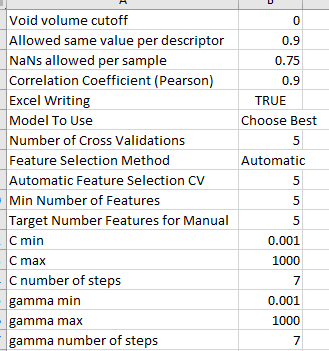
Most Recent Settings will change the settings to those used to create the most recent model (same as Original Settings Button unless you create multiple models)

**Changing Default Settings**

The default settings currently in the software may not be appropriate for all applications. The instructions to perform this are below

*User Settings*

This covers the settings that are altered in the settings menu. In the folder with your .exe or .py file there is a file called “User\_Settings\_Defaults.csv”. Adjust this as desired and save the file. Your defaults will be loaded when you next open QSRR Automater. The file looks like this



Most of these should be fairly self-explanatory but the translations are below:

In csv – in settings menu

Void volume cutoff – Void Volume Time (discard values before this time)

Allowed same value per descriptor - Allowed % of Samples with same value for a Descriptor (percentage in the settings menu a decimal in the .csv)

NaNs allowed per sample – Allowed % missing descriptors per sample (percentage in the settings menu a decimal in the .csv)

Correlation Coefficient (Pearson) = maximum correlation allowed (pearson r)

Excel Writing – Excel or CSV output (True is Excel, False is CSV)

Model To Use – Machine Learning Model

Number of Cross Validations – Cross Validations

Feature Selection Method – Feature Selection Method

Automatic Feature Selection CV – Cross Validations for Automatic

Min Number of Features – Minimum number of features (automatic)

Target Number Features for Manual – Desired number of features (manual)

C min – Minimum C value

C max – Maximum C value

C number of steps – C Number of Steps

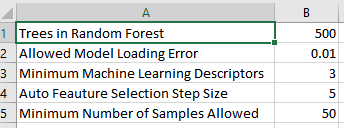
gamma min – Minimum gamma value

gamma max – Maximum gamma value

gamma number of steps – gamma Number of Steps

*Programmer Settings*

These are settings that should not be altered lightly and cannot be altered in the program itself. To adjust open the Programmer\_Settings\_Defaults.csv in the same folder as your .py or .exe file. It looks like this:



Trees in Random Forest – in the feature selection or the Random Forest which creates many decision tree models. This determines the number of trees in both feature selection and attempts to use a random forest as the final model.

Allowed Model Loading Error – when loading a previously trained model, QSRR Automater will compare the original score on the training set with a re-analysis of that training set to ensure no problems in the model. This is how much error is allowed.

Minimum Machine Learning Descriptors – the minimum allowed number of descriptors for any model

Auto Feature Selection Step Size – the auto feature selection step proceeds by taking off a certain number of the worst features (descriptors) at a time. This sets that size. Do note that it can choose a number of features in between the step sizes, this just helps narrow the range.

Minimum Number of Samples Allowed – sets the minimum number of compounds present in the training set. Prediction sets can have any amount.

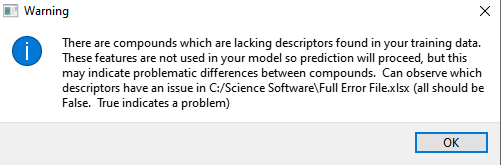
**Troubleshooting**

*Large error in prediction*

This can come from many sources but there are two that are most common:

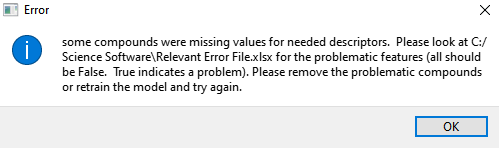
1. extrapolation. Your training set has a minimum retention time and a maximum retention time. Values beyond those are likely to lead to problems. This can also happen if the training set has one value minutes before or after the others. So if the value you think is proper or the one that is predicted is outside the range of your training set, be cautious about trusting it
2. The data in your training set is not representative. This can happen because your training set was too small, the compounds to predict are too different from your training set, or something about your instrument setup has changed (column has aged, new column, different solvents, different gradient) In this case either add more to your training set or re-train for your new instrument state.

*Error in Prediction but will proceed*



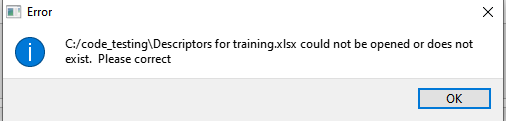
This warning will occur in an RT Prediction step. It indicates that some of the compounds you wish to predict the retention time of has molecular descriptors that worked with your training set. In this case these descriptors were not used in the model. The prediction will proceed, though you may wish to see which compounds had the issue. Open Full Error File.xlsx and any True values are the problem.

*Error in Prediction will not proceed*



This error will occur in an RT Prediction step. It indicates that some of the compounds you wish to predict the retention time of has molecular descriptors that are required by the model. Open the Relevant Error File.xlsx and find the compounds with True values. You can remove the compounds and save a new prediction set, or train a new model to see if it does not require this descriptor.

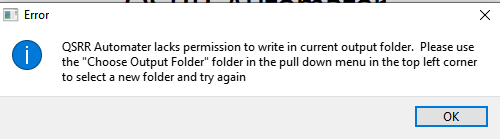
*Load Model errors.*



Sometimes the error above will occur when trying to load a previously created model. What is happening is that a file called saved\_model\_settings.csv is created when the model is. In it is the “Training file” cell with a file path next to it. If the file there does not exist, then the model can’t be loaded. However, the file path in saved\_model\_settings.csv can be adjusted if you need to move the folder.

You will also get error if the saved\_model.joblib (the actual model) or the saved\_model\_settings.csv files can’t be opened or have been altered in some way. Just ensure they are there and undo changes to correct.

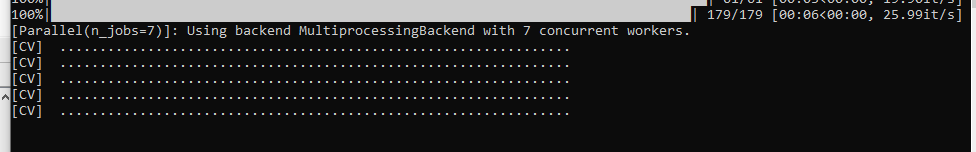
*Admin error*



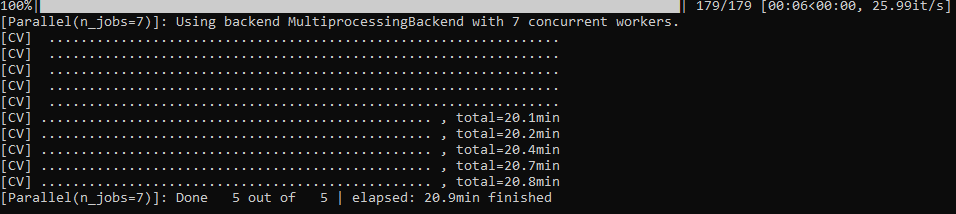
This error will occur if the QSRR\_Automater is placed within a folder that requires administrator permission to alter. Simply us the “Choose Output Folder” option to point to a new folder or close QSRR Automater and move it to a new folder which does not require these permissions and open it again.

*Prediction taking an unusually long time*

A few minutes after the beginning of an analysis the associated command window should look like this:



After a period of time it will look something like this:



It will then take further time to finish the analysis. There are a few reasons this can slow down:

1. feature selection – “None” for feature selection is far faster than “Automatic” or “Manual”
2. number of cross validations – the more cross validations the longer this will take
3. other things going on in the computer, the more complex analyses that are going on the computer the longer all of them will take
4. sometimes the cmd window will pause.

If the problem is 1-3 just wait and maybe turn off other processes. If the issue is 4 click on the window and press the down arrow key on your keyboard. If this is the problem, it should cause an immediate change in the command prompt or cause a window to pop up in QSRR Automater. If none of these work, force QSRR Automater to shut off and open it and try again.

*Program closes automatically at a certain point or Software will not open –*

This requires contacting the programmers. It could be the result of numerous errors.

It is helpful to get the full error message. In python just copy it from the command prompt or editor. If using the .exe open a command prompt (type cmd in the windows start bar) and enter the full path of the .exe into it. This means if the QSRR\_Automater Folder is in a Science Software folder on the C drive type in “C:\Science Software\QSRR\_Automater\QSRR\_Automater.exe”. keep the quotation marks if you have any spaces in the path. Run QSRR Automater until the problem occurs and an error message should appear on the command prompt. Take a screenshot and send it with your request for assistance.