**User Manual for “MS2Purifier\_model\_generation.r”**

(Version 1.0, Nov 20, 2020)

Shipei Xing, Tao Huan\*

Department of Chemistry, Faculty of Science, University of British Columbia, Vancouver Campus, 2036 Main Mall, Vancouver, V6T 1Z1, BC, Canada

\* Author to whom correspondence should be addressed:

Dr. Tao Huan

Tel: (+1)-604-822-4891

E-mail: thuan@chem.ubc.ca

Internet: <https://huan.chem.ubc.ca/>

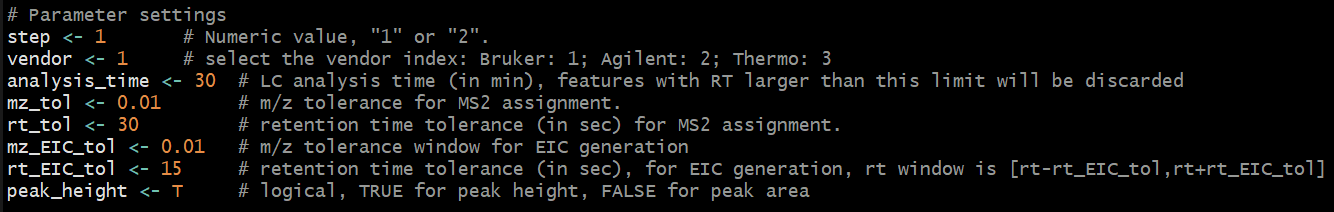
* This R script allows users to generate their own random forest models that can be embedded in MS2Purifier.
* The program is written in R and is now publicly available on GitHub (https://github.com/HuanLab/MS2Purifier).
* Please see below for the code instructions.

1. Install the R language, Rstudio and following packages: xcms, randomForest.
2. Download and open the R script “MS2Purifier\_model\_generation.r” in Rstudio.
3. Change the file directory and model directory in the R script. The file directory should contain mzXML files collected in DDA and DIA (AIF) mode. Use “/” instead of “\” in the directory.

Note that the mzXML files of the same sample in the file directory should be named as “XXX\_DDA.mzXML” and “XXX\_AIF.mzXML”.



1. Set the parameters in the following table. Set the “step” as 1.



**Table.** Parameter settings.

|  |  |
| --- | --- |
| **Parameter name** | **Function** |
| step | Numeric value: 1 or 2. |
| vendor | MS vendor index. Bruker: 1; Agilent: 2; Thermo: 3. |
| analysis\_time | LC analysis time (in min), features with RT larger than this limit will be discarded. |
| mz\_tol | The *m/z* tolerance for MS2 assignment. |
| rt\_tol | The retention time tolerance (in sec) for MS2 assignment. |
| mz\_EIC\_tol | The *m/z* tolerance window for EIC generation in AIF data. |
| rt\_EIC\_tol | The retention time tolerance (in sec) for EIC generation in AIF data. RT window is [rt - rt\_EIC\_tol, rt + rt\_EIC\_tol]. |
| peak\_height | Logical, TRUE for peak height, FALSE for peak area. |

1. Run the R script by clicking “Source” on the top right of the Rstudio panel.
2. A new csv file “Fragment\_for\_manual\_check.csv” and a subfolder “EIC\_plots” containing the EIC plots of precursor-fragment pairs will be created. Manual inspections will then be needed to fill in the column of “manual\_check” in the csv file. Please use “T” for true fragments, “F” for contamination ions and “NA” for unchecked fragments.
3. After manual checking, save the edited csv file and change the “step” into 2 in the R script.



1. Click “Source”, and a new random forest model “RF\_new\_model.rds” will be generated in the file directory.