**User Manual for “MS2Purifier.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

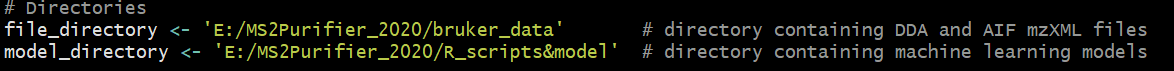
E-mail: thuan@chem.ubc.ca

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

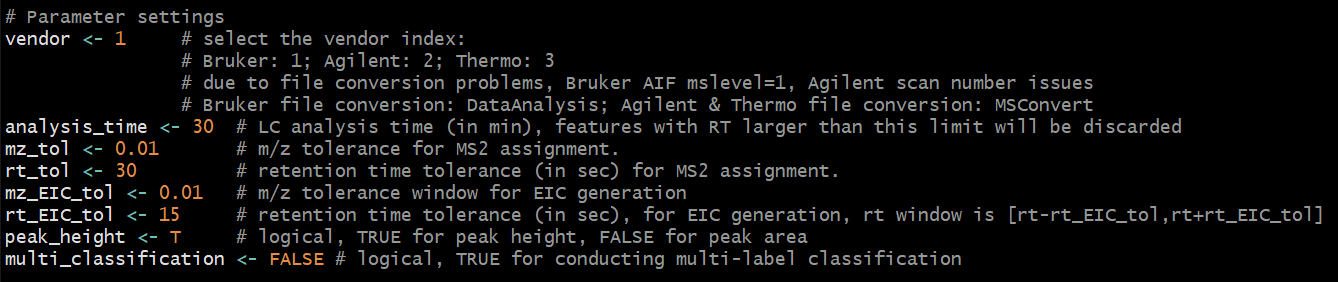
* This R script performs MS2 spectra purification collected in data-dependent acquisition (DDA) mode with help of all-ion fragmentation (AIF) data in data-independent acquisition (DIA) mode.
* The program is written in R and is now publicly available on GitHub (https://github.com/HuanLab/MS2Purifier).
* 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.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. The model directory should contain the random forest models “RF\_model\_binary.rds” and “RF\_model\_multi.rds”. 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.



**Table.** Parameter settings.

|  |  |
| --- | --- |
| **Parameter name** | **Function** |
| 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. |
| multi\_classification | Logical, TRUE for performing the multi-label classification. |

1. Run the R script by clicking “Source” on the top right of the Rstudio panel.
2. Three csv files will be output with the names of “Feature\_table\_DDA\_MS2assigned.csv”, “Fragment\_prediction\_result.csv” and “Feature\_table\_DDA\_MS2purified.csv”. They refer to the DDA feature table with MS2 spectra assigned, the binary classification prediction result of each fragment ion and the feature table with purified MS2 spectra. If multi-label classification is enabled, another csv file named “Fragment\_prediction\_multi\_result.csv” will be output.