**User Manual for “DNMS2Purifier.r” and “DNMS2Purifier\_model\_generation.r”**

(Version 1.0, October 3rd, 2022)

Tingting Zhao, 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/>

DNMS2Purifier is a machine learning-based (XGBoost) bioinformatic solution to purify the chimeric MS/MS spectra collected in DDA-based LC-MS/MS untargeted metabolomics. The program is written in R (ver 4.2.1). The R script “DNMS2Purifier.r” is the main program for MS/MS purification (Part **I**), we also provide the script “DNMS2Purifier\_model\_generation.r” for customized model retraining (Part **II**). All source codes are publicly available on GitHub (<https://github.com/HuanLab/DNMS2Purifier>).

**Prerequisite**

To run the above R scripts, the user needs to prepare their computer with R software, RStudio software and R packages as below:

* install the R language ([www.r-project.org](http://www.r-project.org))
* install the RStudio ([www.rstudio.com](http://www.rstudio.com))
* install the R packages of “xcms” and “xgboost” using the following R scripts:

if (!require("BiocManager", quietly = TRUE))

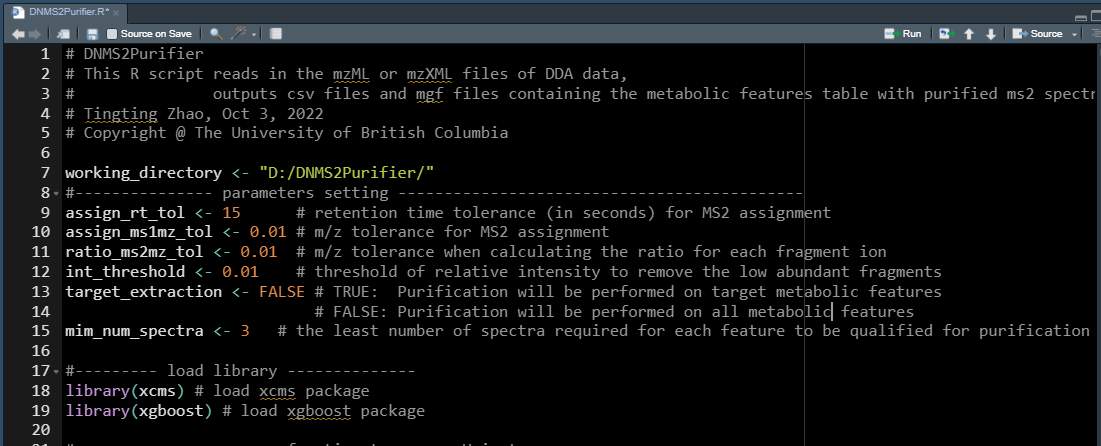
install.packages("BiocManager")

BiocManager::install("xcms")

install.package(“xgboost”)

**Part I: instructions for “DNMS2Purifier.r”**

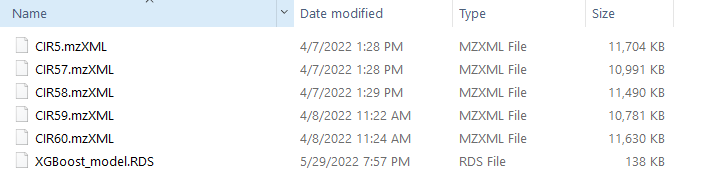
1. Download and open the R script of “DNMS2Purifier.r” in RStudio.



1. Change the working directory in the R script (line 7). Use “/” instead of “\” in the directory as shown below:



The above working directory should contain all the sample files (in .mzML or .mzXML format) and the trained XGBoost model (.RDS file).



1. Set the parameters in lines 9-15.

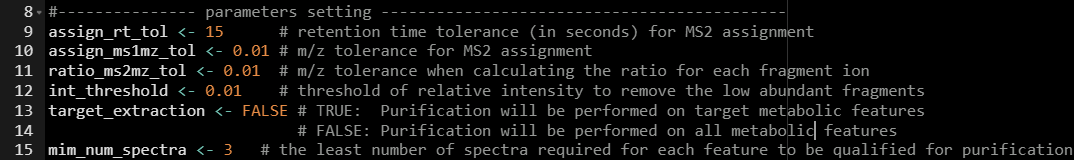
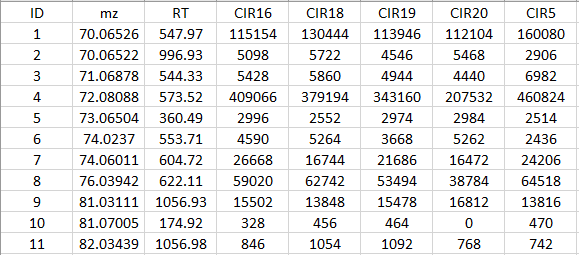


Table. Parameter settings.

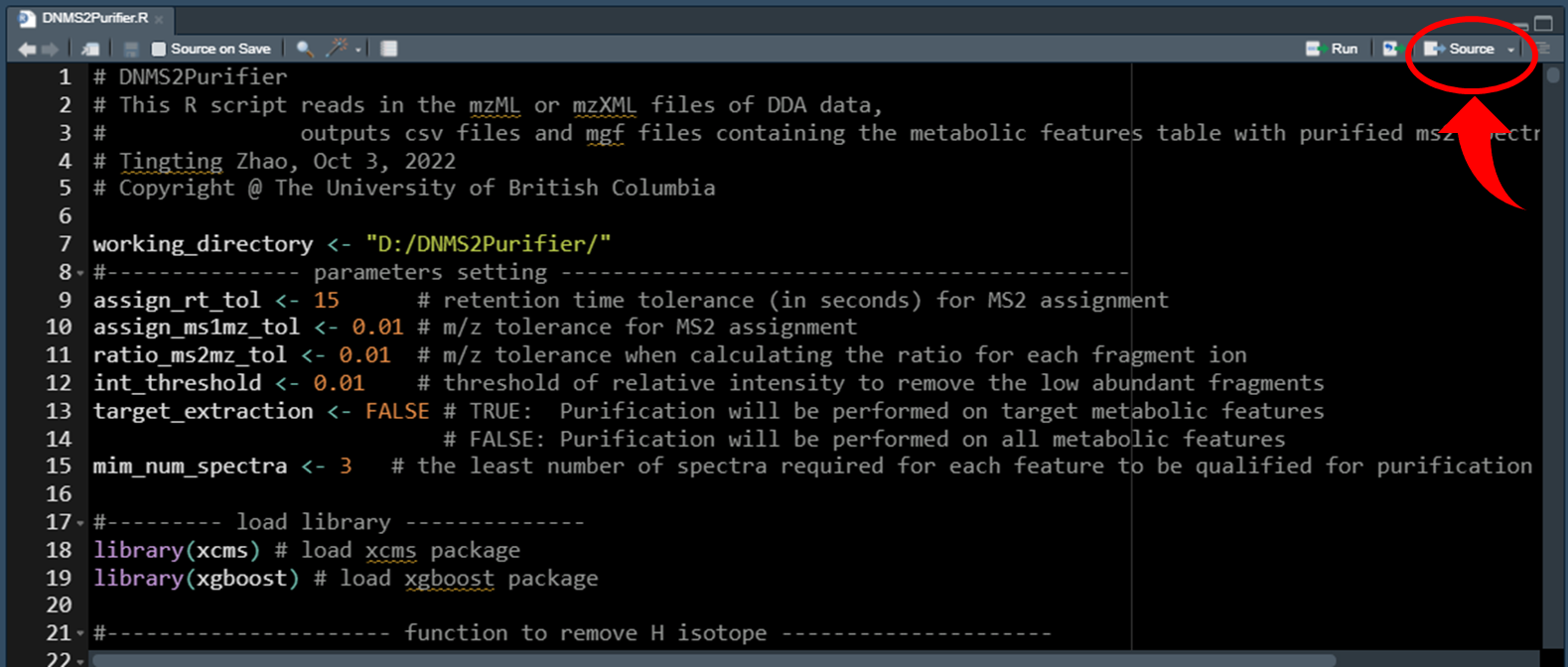
|  |  |
| --- | --- |
| **Parameters** | **Function** |
| assign\_rt\_tol | Numeric, retention time tolerance (in seconds) for MS/MS assignment. Default: 15 |
| assign\_ms1mz\_tol | Numeric, precursor *m/z* tolerance for MS/MS assignment.  Default: 0.01 |
| ratio\_ms2mz\_tol | Numeric, MS/MS tolerance for fragment ion alignment among different spectra. Default: 0.01 |
| int\_threshold | Numeric, relative intensity threshold of reserved fragment ions in MS/MS spectra. Default: 0.01 |
| target\_extraction | Logical. TRUE if MS/MS spectral purification is performed on target metabolic features in a customized feature table (“target\_table.csv”) is needed\*; FALSE if purification is applied on all metabolic features. Default: FALSE. |
| mim\_num\_spectra | Integer, the minimum number of MS/MS spectra required for each metabolic feature to be purified. Default: 3 |

\* If the user chooses to purify MS/MS spectra of targeted features, they need to provide a feature table named “target\_table.csv” in the working directory. The feature table should contain the information of metabolic features of interest, formatted as below:



* Column of “ID”: feature index
* Column of “mz”: *m/z* value of metabolic features
* Column of “RT”: retention times of metabolic features in seconds
* Columns from “CIR16” to “CIR5”: feature intensities in the corresponding samples. These column names are consistent with the sample files.

1. Run the R script by clicking “Source” on the top right of the RStudio panel.

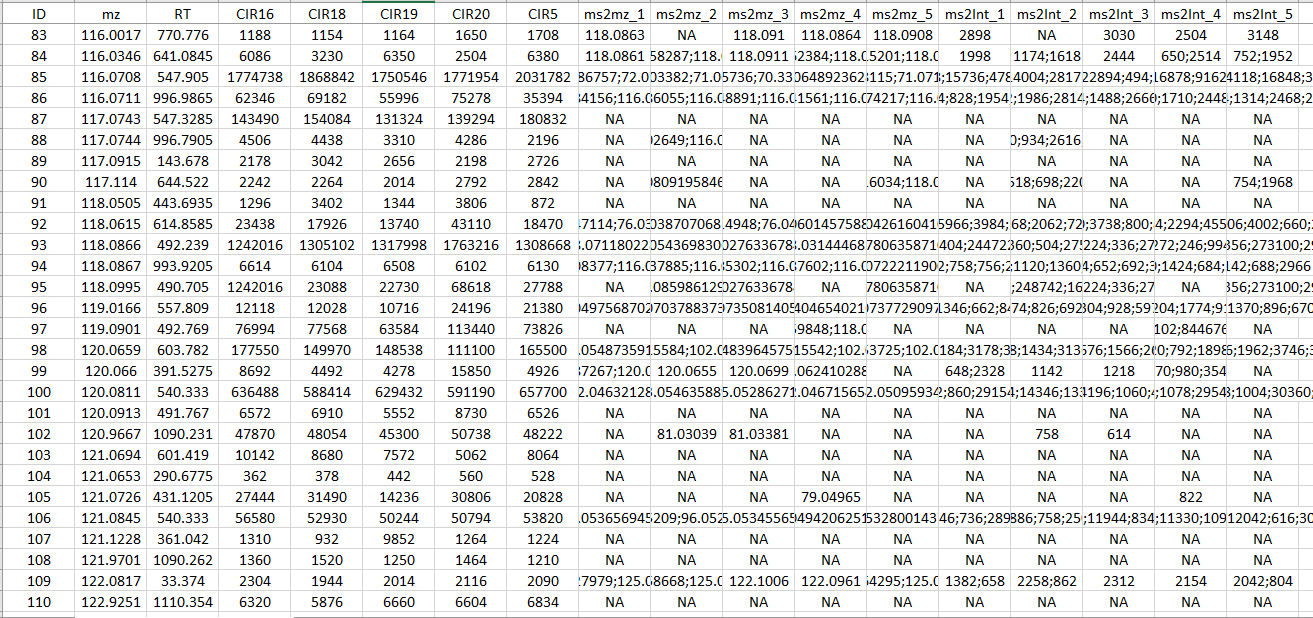


1. Three .csv files and one .mgf file will be output in the working directory as shown below:



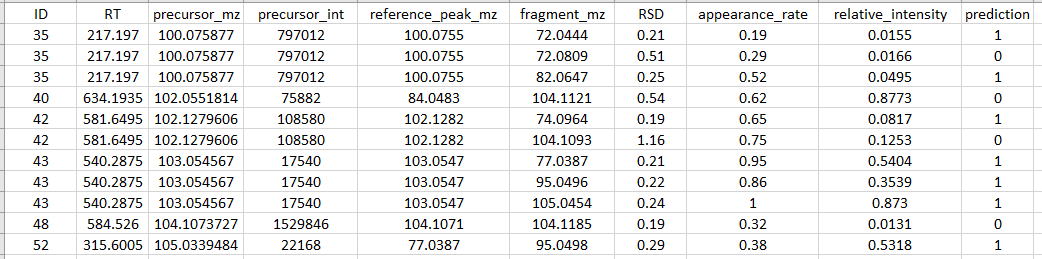
Detailed information about the above output files:

5.1 The “feature\_table\_ms2\_assigned.csv” refers to the feature table with original MS/MS spectra. An example is shown below:



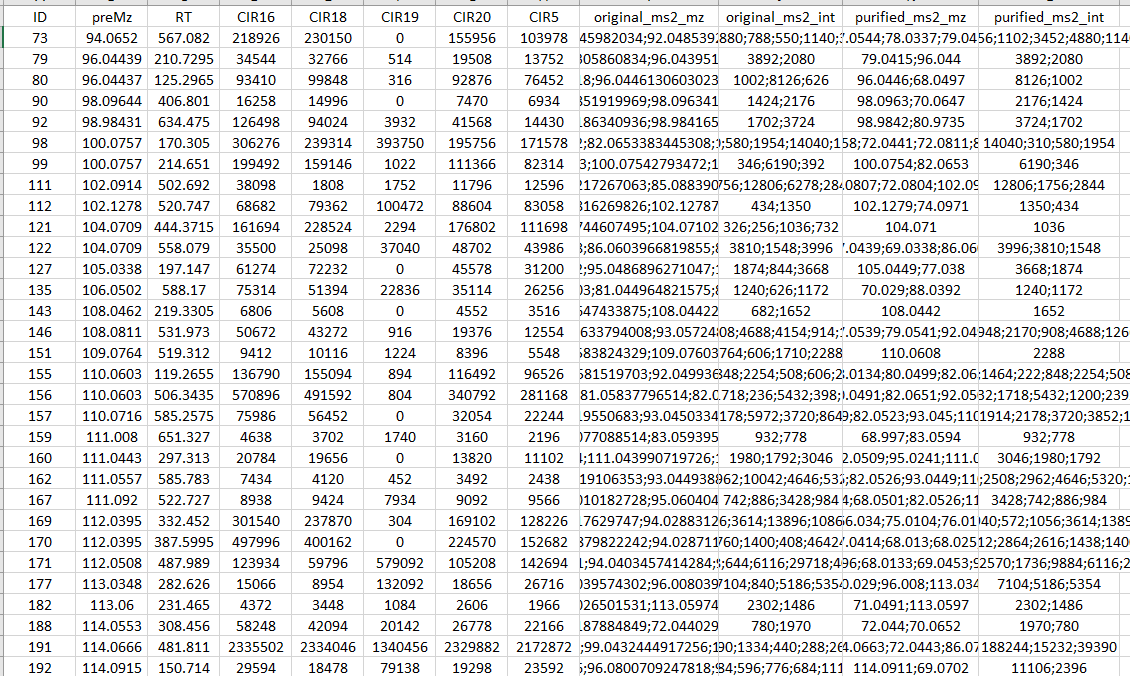
* Column of “ID”: feature index
* Column of “mz”: *m/z* value of metabolic features
* Column of “RT”: retention times of metabolic features in min
* Columns of “CIR16” to “CIR5”: intensity of metabolic features in the corresponding samples
* Columns of “ms2mz\_1” to “ms2mz\_5”: the *m/z* of fragment ions in the original MS/MS spectra from sample “CIR16” to “CIR5”, respectively, separated by semicolons
* Columns of “ms2Int\_1” to “ms2Int\_5”: the fragment intensity in the original MS/MS spectra from sample “CIR16” to “CIR5”, respectively, separated by semicolons.

5.2 The “prediction\_output.csv” refers to the binary classification prediction results of individual fragment ions. An example is shown below:



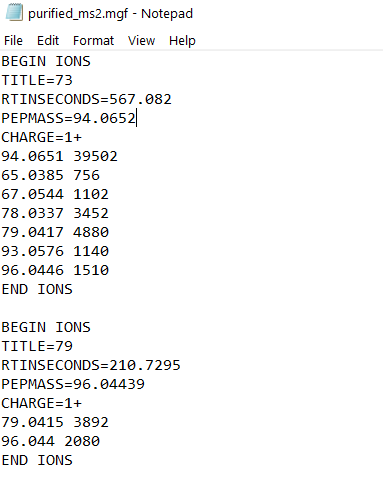
* Column of “ID”: feature index
* Column of “RT”: retention times of the metabolic features
* Column of “precursor\_mz”: *m/z* of the metabolic feature
* Column of “precursor\_int”: feature intensities
* Column of “reference\_peak\_mz”: *m/z* of reference fragment ion
* Column of “fragment\_mz”: *m/z* of fragment ions
* Column of “RSD”: ratio RSDof fragment ions
* Column of “appearance\_rate”: appearance rateof fragment ions
* Column of “relative\_intensity”: relative intensity of fragment ions
* Column of “prediction”: “1” for predicted as true fragment, “0” for false

5.3 The “feature\_table\_ms2\_purified.csv” refers to the feature table with both original and purified MS/MS spectra. An example is shown below:



* Column of “ID”: feature index
* Column of “preMz”: *m/z* of metabolic features
* Column of “RT”: retention times of metabolic features
* Columns of “CIR16” to “CIR5”: feature intensities in different samples
* Columns of “original\_ms2\_mz” and “original\_ms2\_int”: *m/z* and intensities of fragment ions in the original MS/MS spectra from the sample of highest feature intensity
* Columns of “purified\_ms2\_mz” and “purified\_ms2\_int”: *m/z* and intensities of fragment ions in the purified MS/MS spectra.

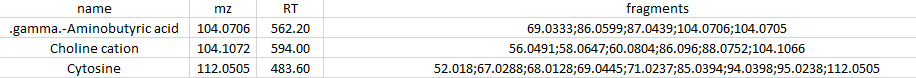
5.4 The “purified\_ms2.mgf” contains the purified MS/MS spectra for metabolic features. An example is shown below:



**Part II: instructions for retraining XGBoost model**

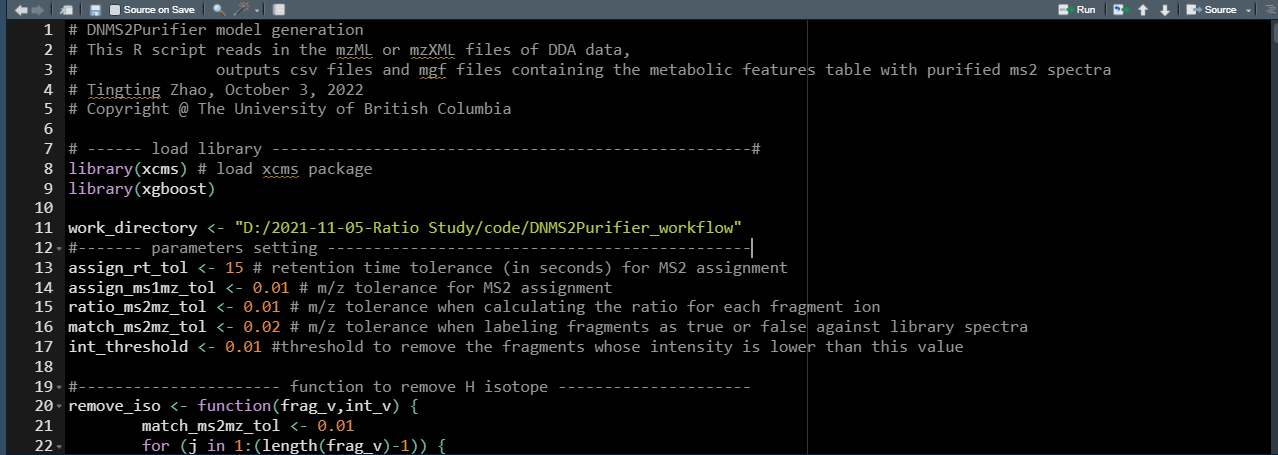
1. Data preparation

A set of LC-MS/MS data of samples containing chemical standards is needed for model training. Samples can be prepared by mixing a standard pool with a biological matrix (e.g. urine) at different ratios (e.g. 4:1, 2:1, 1:1, 1:2, 1:4). Following LC-MS/MS analysis of above samples, the obtained raw data need to be converted to either .mzML or .mzXML files. Besides file conversion, the information of chemical standards needs to be saved in a file (“standards\_information.csv”) as shown below:



* Column of “name”: name of chemical standards
* Column of “mz”: *m/z* of chemical standards
* Column of “RT”: retention time (in seconds) of metabolic features
* Columns of “fragments”: *m/z* of all the possible fragment ions from the chemical standards. This can be obtained by referring to the high-quality spectral database (e.g. NIST20).

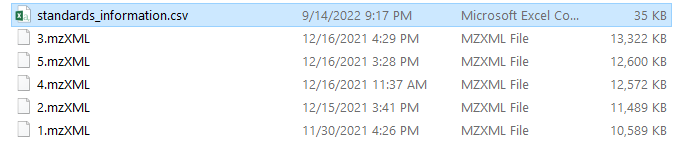
1. Download and open the R script of “DNMS2Purifier\_model\_generation.r” in RStudio.



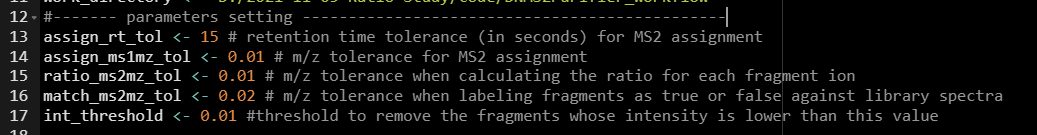
1. Change the working directory in the R script in line 11.



Notably, the working directory should contain all sample files and the .csv file named “standards\_information.csv”.



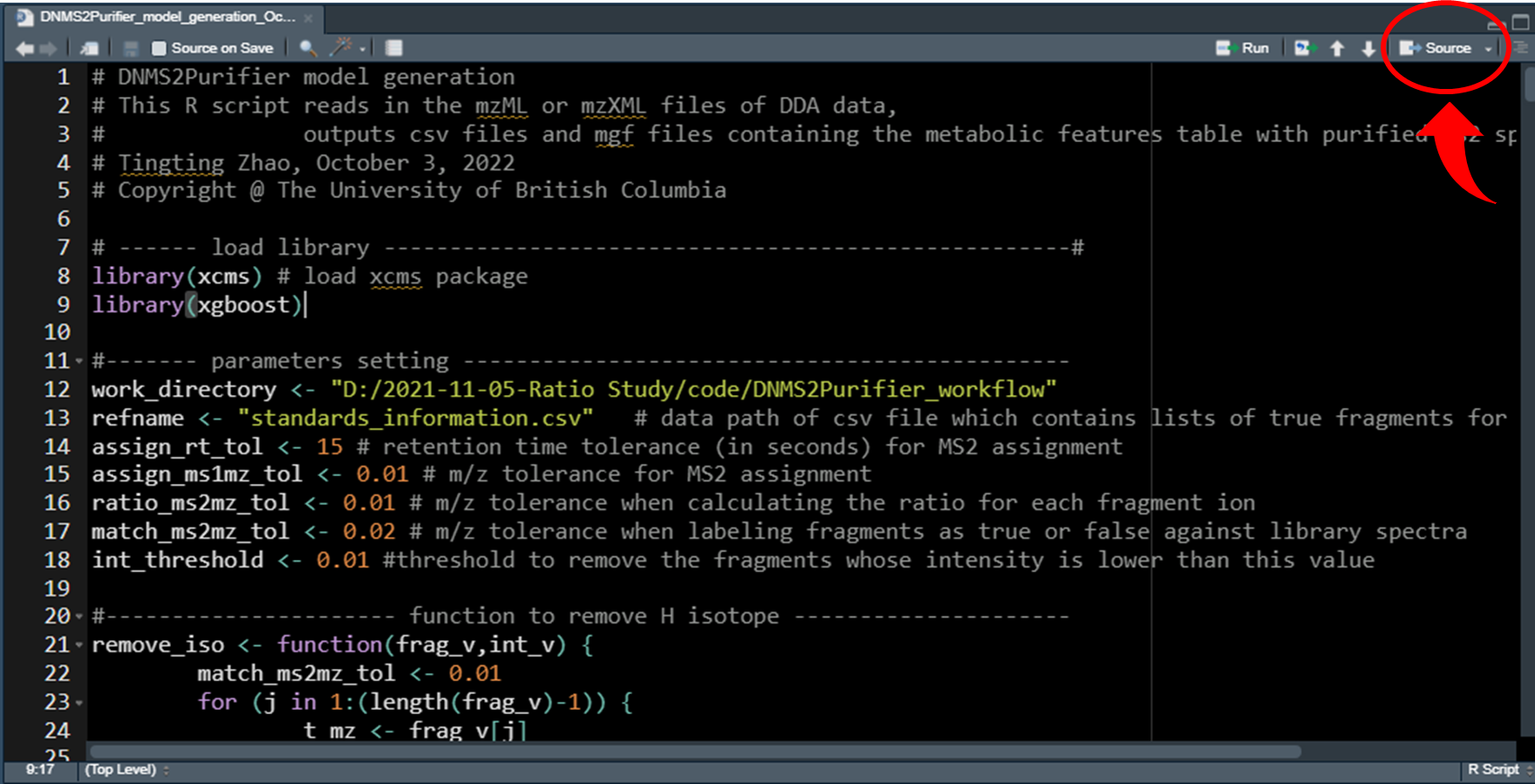
1. Set the parameters in the R script in line 13-17.



Table**.** Parameter settings.

|  |  |
| --- | --- |
| **Parameters** | **Function** |
| assign\_rt\_tol | Numeric, retention time tolerance (in seconds) for MS/MS assignment. Default: 15 |
| assign\_ms1mz\_tol | Numeric, precursor *m/z* tolerance for MS/MS assignment.  Default: 0.01 |
| ratio\_ms2mz\_tol | Numeric, MS/MS tolerance for fragment ion alignment among different spectra. Default: 0.01 |
| match\_ms2mz\_tol | Numeric, MS/MS tolerance threshold for alignment of fragment ion in the reference spectra. Default: 0.02 |
| int\_threshold | Relative intensity threshold to keep the fragment ions in MS/MS spectra.  Default: 0.01 |

1. Run the R script by clicking “Source” on the top right of the RStudio panel.



1. The newly trained XGBoost model will be output as “XGBoost.RDS” in the working directory as shown below.

