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

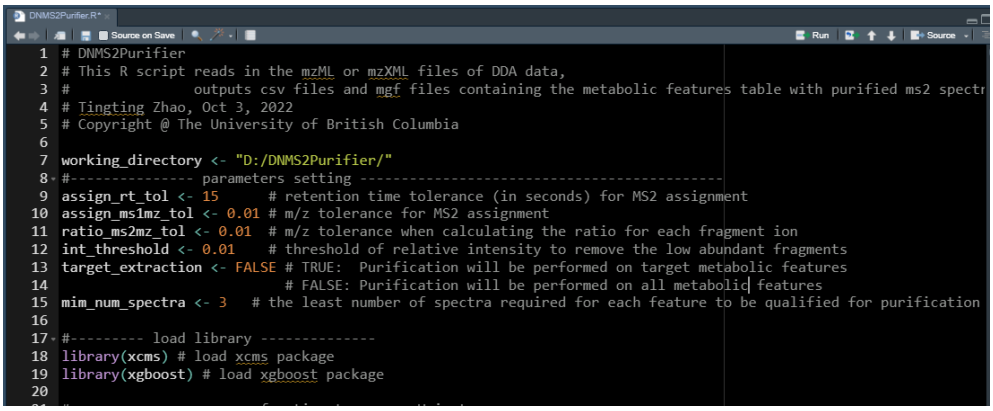
To run the above R scripts, the user needs to prepare their computer with R software, RStudio software and R packages as below:

- install the R language (www.r-project.org)
- install the RStudio (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  # DNMS2Purifier
2  # This R script reads in the mzML or mzXML files of DDA data,
3  #              outputs csv files and mgf files containing the metabolic features table with purified ms2 spectr
4  # Tingting Zhao, Oct 3, 2022
5  # Copyright @ The University of British Columbia
6
7  working_directory <- "D:/DNMS2Purifier/"
8  #-------------- parameters setting ------------------------------------------
9  assign_rt_tol <- 15      # retention time tolerance (in seconds) for MS2 assignment
10 assign_ms1mz_tol <- 0.01 # m/z tolerance for MS2 assignment
11 ratio_ms2mz_tol <- 0.01  # m/z tolerance when calculating the ratio for each fragment ion
12 int_threshold <- 0.01    # threshold of relative intensity to remove the low abundant fragments
13 target_extraction <- FALSE # TRUE:  Purification will be performed on target metabolic features
14                           # FALSE: Purification will be performed on all metabolic features
15 mim_num_spectra <- 3   # the least number of spectra required for each feature to be qualified for purification
16
17 #-------- load library -------------
18 library(xcms) # load xcms package
19 library(xgboost) # load xgboost package
20
```

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

```
7  working_directory <- "D:/DNMS2Purifier/"
```

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

Name	Date modified	Type	Size
CIR5.mzXML	4/7/2022 1:28 PM	MZXML File	11,704 KB
CIR57.mzXML	4/7/2022 1:28 PM	MZXML File	10,991 KB
CIR58.mzXML	4/7/2022 1:29 PM	MZXML File	11,490 KB
CIR59.mzXML	4/8/2022 11:22 AM	MZXML File	10,781 KB
CIR60.mzXML	4/8/2022 11:24 AM	MZXML File	11,630 KB
XGBoost_model.RDS	5/29/2022 7:57 PM	RDS File	138 KB

3) Set the parameters in lines 9-15.

```
8  #-------------- parameters setting ------------------------------------------
9  assign_rt_tol <- 15      # retention time tolerance (in seconds) for MS2 assignment
10 assign_ms1mz_tol <- 0.01 # m/z tolerance for MS2 assignment
11 ratio_ms2mz_tol <- 0.01  # m/z tolerance when calculating the ratio for each fragment ion
12 int_threshold <- 0.01    # threshold of relative intensity to remove the low abundant fragments
13 target_extraction <- FALSE # TRUE:  Purification will be performed on target metabolic features
14                           # FALSE: Purification will be performed on all metabolic features
15 mim_num_spectra <- 3   # the least number of spectra required for each feature to be qualified for purification
```

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

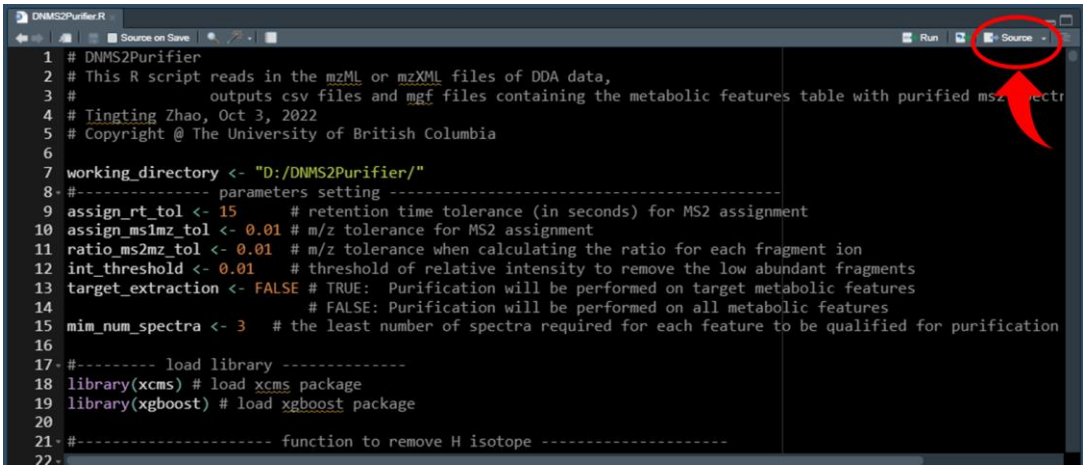
ratio_ms2mz_tol	Numeric, MS/MS tolerance for fragment ion alignment among different spectra. Default: 0.01
int_threshold	Numeric, relative intensity threshold of reserved fragment ions in MS/MS spectra. Default: 0.01
target_extraction	Logical. TRUE if MS/MS spectral purification is performed on target 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:

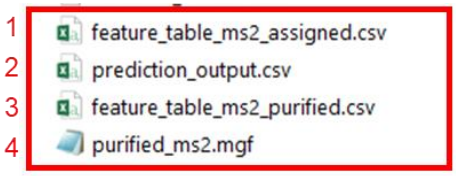
ID	mz	RT	CIR16	CIR18	CIR19	CIR20	CIR5
1	70.06526	547.97	115154	130444	113946	112104	160080
2	70.06522	996.93	5098	5722	4546	5468	2906
3	71.06878	544.33	5428	5860	4944	4440	6982
4	72.08088	573.52	409066	379194	343160	207532	460824
5	73.06504	360.49	2996	2552	2974	2984	2514
6	74.0237	553.71	4590	5264	3668	5262	2436
7	74.06011	604.72	26668	16744	21686	16472	24206
8	76.03942	622.11	59020	62742	53494	38784	64518
9	81.03111	1056.93	15502	13848	15478	16812	13816
10	81.07005	174.92	328	456	464	0	470
11	82.03439	1056.98	846	1054	1092	768	742

- 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.

4) Run the R script by clicking "Source" on the top right of the RStudio panel.



5) 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:

ID	mz	RT	CIR16	CIR18	CIR19	CIR20	CIR5	ms2mz_1	ms2mz_2	ms2mz_3	ms2mz_4	ms2mz_5	ms2Int_1	ms2Int_2	ms2Int_3	ms2Int_4	ms2Int_5
83	116.0017	770.776	1188	1154	1164	1650	1708	118.0863	NA	118.091	118.0864	118.0908	2898	NA	3030	2504	3148
84	116.0346	641.0845	6086	3230	6350	2504	6380	118.0861	58287;118.	118.0911	2384;118.0	5201;118.0	1998	1174;1618	2444	650;2514	752;1952
85	116.0708	547.905	1774738	1868842	1750546	1771954	2031782	86757;72.0	3382;71.05	736;70.33	64892362	115;71.071	15736;478	4004;2817	22894;494	16878;9162	4118;16848;3
86	116.0711	996.9865	62346	69182	55996	75278	35394	4156;116.0	6055;116.0	8891;116.0	1561;116.0	74217;116.4	828;1954	1986;2814	1488;2666	1710;2448	1314;2468;2
87	117.0743	547.3285	143490	154084	131324	139294	180832	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
88	117.0744	996.7905	4506	4438	3310	4286	2196	NA	2649;116.0	NA	NA	NA	NA	0;934;2616	NA	NA	NA
89	117.0915	143.678	2178	3042	2656	2198	2726	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
90	117.114	644.522	2242	2264	2014	2792	2842	NA	80919584	NA	NA	6034;118.0	NA	518;698;22	NA	NA	754;1968
91	118.0505	443.6935	1296	3402	1344	3806	872	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

- 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:

ID	RT	precursor_mz	precursor_int	reference_peak_mz	fragment_mz	RSD	appearance_rate	relative_intensity	prediction
35	217.197	100.075877	797012	100.0755	72.0444	0.21	0.19	0.0155	1
35	217.197	100.075877	797012	100.0755	72.0809	0.51	0.29	0.0166	0
35	217.197	100.075877	797012	100.0755	82.0647	0.25	0.52	0.0495	1
40	634.1935	102.0551814	75882	84.0483	104.1121	0.54	0.62	0.8773	0
42	581.6495	102.1279606	108580	102.1282	74.0964	0.19	0.65	0.0817	1
42	581.6495	102.1279606	108580	102.1282	104.1093	1.16	0.75	0.1253	0
43	540.2875	103.054567	17540	103.0547	77.0387	0.21	0.95	0.5404	1
43	540.2875	103.054567	17540	103.0547	95.0496	0.22	0.86	0.3539	1
43	540.2875	103.054567	17540	103.0547	105.0454	0.24	1	0.873	1
48	584.526	104.1073727	1529846	104.1071	104.1185	0.19	0.32	0.0131	0
52	315.6005	105.0339484	22168	77.0387	95.0498	0.29	0.38	0.5318	1

- 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 RSD of fragment ions
- Column of “appearance_rate”: appearance rate of 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:

ID	preMz	RT	CIR16	CIR18	CIR19	CIR20	CIR5	original_ms2_mz	original_ms2_int	purified_ms2_mz	purified_ms2_int
73	94.0652	567.082	218926	230150	0	155956	103978	45982034;92.048539	880;788;550;1140;	.0544;78.0337;79.0456	1102;3452;4880;114
79	96.04439	210.7295	34544	32766	514	19508	13752	05860834;96.043951	3892;2080	79.0415;96.044	3892;2080
80	96.04437	125.2965	93410	99848	316	92876	76452	.8;96.0446130603023	1002;8126;626	96.0446;68.0497	8126;1002
90	98.09644	406.801	16258	14996	0	7470	6934	51919969;98.096341	1424;2176	98.0963;70.0647	2176;1424
92	98.98431	634.475	126498	94024	3932	41568	14430	.86340936;98.984165	1702;3724	98.9842;80.9735	3724;1702
98	100.0757	170.305	306276	239314	393750	195756	171578	;82.0653383445308;1	580;1954;14040;158	72.0441;72.0811;8	14040;310;580;1954
99	100.0757	214.651	199492	159146	1022	111366	82314	3;100.0754279347 2;1	346;6190;392	100.0754;82.0653	6190;346
111	102.0914	502.692	38098	1808	1752	11796	12596	17267063;85.088390	756;12806;6278;284	0807;72.0804;102.09	12806;1756;2844
112	102.1278	520.747	68682	79362	100472	88604	83058	16269826;102.12787	434;1350	102.1279;74.0971	1350;434
121	104.0709	444.3715	161694	228524	2294	176802	111698	44607495;104.07102	326;256;1036;732	104.071	1036

- 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:

```
BEGIN IONS
TITLE=73
RTINSECONDS=567.082
PEPMASS=94.0652
CHARGE=1+
94.0651 39502
65.0385 756
67.0544 1102
78.0337 3452
79.0417 4880
93.0576 1140
96.0446 1510
END IONS

BEGIN IONS
TITLE=79
RTINSECONDS=210.7295
PEPMASS=96.04439
CHARGE=1+
79.0415 3892
96.044 2080
END IONS
```

Part II: instructions for retraining XGBoost model

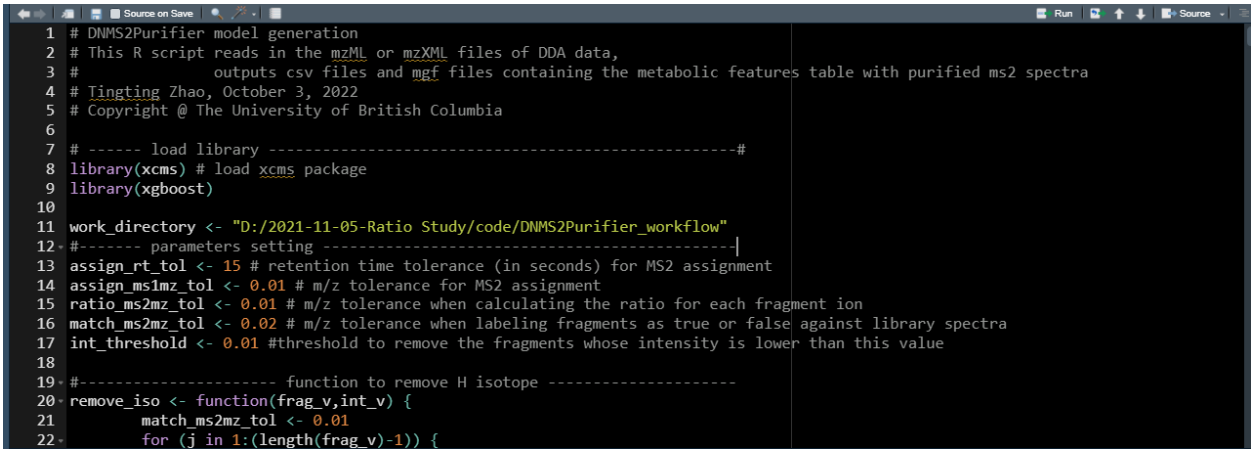
1) Data preparation

   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:

name	mz	RT	fragments
.gamma.-Aminobutyric acid	104.0706	562.20	69.0333;86.0599;87.0439;104.0706;104.0705
Choline cation	104.1072	594.00	56.0491;58.0647;60.0804;86.096;88.0752;104.1066
Cytosine	112.0505	483.60	52.018;67.0288;68.0128;69.0445;71.0237;85.0394;94.0398;95.0238;112.0505

   - 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).

2) Download and open the R script of "DNMS2Purifier_model_generation.r" in RStudio.

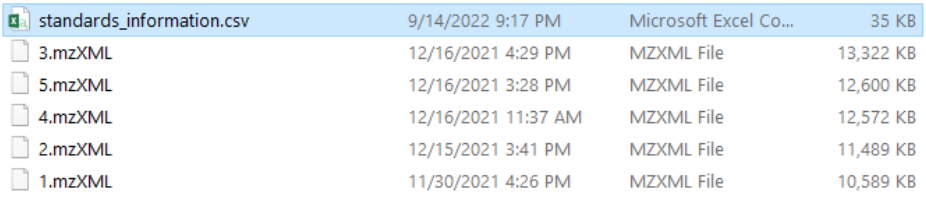


```
1  # DNMS2Purifier model generation
2  # This R script reads in the mzML or mzXML files of DDA data,
3  #              outputs csv files and mgf files containing the metabolic features table with purified ms2 spectra
4  # Tingting Zhao, October 3, 2022
5  # Copyright @ The University of British Columbia
6
7  # ------ load library ----------------------------------------------------#
8  library(xcms) # load xcms package
9  library(xgboost)
10
11 work_directory <- "D:/2021-11-05-Ratio Study/code/DNMS2Purifier_workflow"
12 #------- parameters setting -----------------------------------------------
13 assign_rt_tol <- 15 # retention time tolerance (in seconds) for MS2 assignment
14 assign_ms1mz_tol <- 0.01 # m/z tolerance for MS2 assignment
15 ratio_ms2mz_tol <- 0.01 # m/z tolerance when calculating the ratio for each fragment ion
16 match_ms2mz_tol <- 0.02 # m/z tolerance when labeling fragments as true or false against library spectra
17 int_threshold <- 0.01 #threshold to remove the fragments whose intensity is lower than this value
18
19 #---------------------- function to remove H isotope ---------------------
20 remove_iso <- function(frag_v,int_v) {
21         match_ms2mz_tol <- 0.01
22         for (j in 1:(length(frag_v)-1)) {
```

3) Change the working directory in the R script in line 11.

```
11 work_directory <- "D:/2021-11-05-Ratio Study/code/DNMS2Purifier_workflow"
```

   Notably, the working directory should contain all sample files and the .csv file named "standards_information.csv".



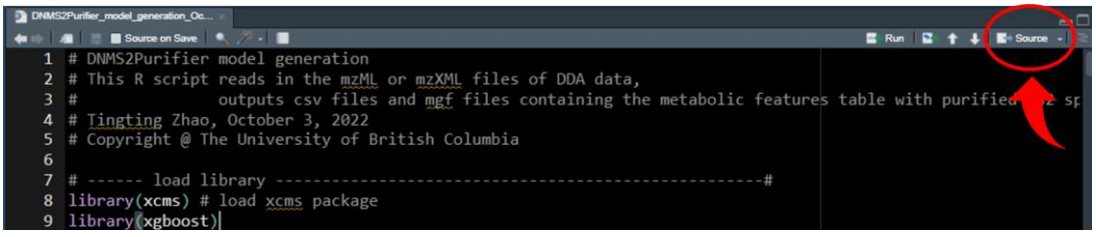
4) Set the parameters in the R script in line 13-17.

```
12 #------- parameters setting -------------------------------------------------
13 assign_rt_tol <- 15 # retention time tolerance (in seconds) for MS2 assignment
14 assign_ms1mz_tol <- 0.01 # m/z tolerance for MS2 assignment
15 ratio_ms2mz_tol <- 0.01 # m/z tolerance when calculating the ratio for each fragment ion
16 match_ms2mz_tol <- 0.02 # m/z tolerance when labeling fragments as true or false against library spectra
17 int_threshold <- 0.01 #threshold to remove the fragments whose intensity is lower than this value
```

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

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



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