User's guideline

Table of Contents

Chapter 1. Evaluation of Citrullinated PSMs

Download

To download Citrullination analysis.ipynb:

- Go to https://github.com/Sunghyun-Huh/Citrullination-Diagnostic-Ion-Analysis.
- Click on 'Citrullination analysis.ipynb' among the listed files.
- Click 'Raw' on the top right panel.
- Press ctrl+s and type 'Citrullination analysis.ipynb' to keep the ipynb extension.

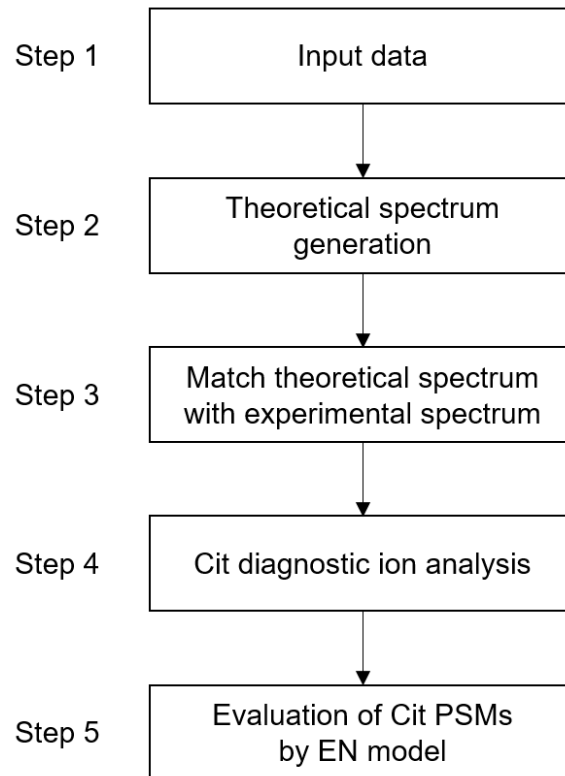## Requirements

Python requirements

- Python version 3.6 or greater
- Libraries needed: pandas, numpy, pyteomics, itertools, collections, statistics, re, os, glob

Necessary Files

- Input search result file: The input file must be a CSV file(s) containing the following three columns (see **Step 1** for the details):

'Title'	MS2 spectrum title as written in MGF file
'Peptide'	Peptide sequence with modification delta mass rounded up to third decimal places
'Charge'	Charge state of the peptide

- Spectrum file: The spectrum file must be a Mascot Generic Format (MGF) file(s) containing MS2 spectra corresponding to those matched to the PSMs in the input search result file. If MS2 spectra in the input search file and spectrum file are not equivalent, only the common MS2 spectra will be retained and subsequently processed.

Overview flowchart

Step 1	Input data
Step 2	Theoretical spectrum generation
Step 3	Match theoretical spectrum with experimental spectrum
Step 4	Cit diagnostic ion analysis
Step 5	Evaluation of Cit PSMs by EN model

Step 1: Input data

A. Setting initial parameters

The initial parameter settings:

0. Initial parameters

```python
In [ ]:  # Fragmentation method used
         Frag_method = 'HCD' # HCD, CID, ETD, ECD

         # Ion types for theoretical spectrum generation
         ion_type = ['Precur', 'y', 'b', 'a', 'z', 'c', 'INT', 'IM']

         # Generate each ion type
         if Frag_method == 'HCD' or Frag_method == 'CID':
             annot_yb     = True # y-, b-, a-ion
             annot_zc     = False # z-, c-ion
         elif Frag_method == 'ETD' or Frag_method == 'ECD':
             annot_yb     = False # y-, b-, a-ion
             annot_zc     = True # z-, c-ion
         annot_precur = True # precursor ion
         annot_INT    = True # internal ion
         annot_IM     = True # immonium ion
         annot_dict = {
             'Precur' : annot_precur,
             'y'      : annot_yb,
             'b'      : annot_yb,
             'a'      : annot_yb,
             'z'      : annot_zc,
             'c'      : annot_zc,
             'INT'    : annot_INT,
             'IM'     : annot_IM
         }

         # MS2 mass tolerance (ppm)
         ms2_ppm = 15

         # Signal-to-noise (SNR) filter for MS2 spectrum
         apply_SNR = True # Apply SNR filter
         SNR       = 2 # SNR threshold
         low       = 0.05 # Define low x% intensity as baseline noise level

         # Maximum charge state of sequence ions
         max_charge = 2 # 2, 3, ... 'max'

         # Maximum number of neutral loss from a single ion
         max_NL = 3
```

Frag_method	Fragmentation method used in the input data (value = 'HCD', 'CID', 'ETD', or 'ECD'). If set as 'HCD' or 'CID', y-ion, b-ion, and a-ion will be generated for theoretical spectrum. If set as 'ETD' or 'ECD', z-ion and c-ion will be generated for theoretical spectrum. Commonly, precursor, internal, and immonium ions will be generated for all these fragmentation methods.
ms2_ppm	MS2 level mass tolerance in ppm (default = 15 ppm).
apply_SNR	Determine whether to apply signal-to-noise filter to remove noise peaks (value = True or False; default = True).
SNR	Signal threshold level. The average intensity of noise peaks (as defined in 'low') multiplied by this signal threshold level will be the final signal-to-noise filter. If 'apply_SNR' = True, all peaks below the signal-to-noise filter will be removed (default = 2).
low	Proportion of MS2 peaks regarded as noise. If 'apply_SNR' = True, all peaks below this noise level will be treated as noise (default = 0.05).
max_charge	Maximum charge state of fragment ions (default = 2).
max_NL	Maximum number of neutral losses from a single ion (default = 3).

Step 1: Input data

B. Loading input data

Example input search result file:

A	B	C
Title	Peptide	Charge
20160312_02_A1.10012.10012.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1001	YETSGIGEAR+0.984VK	2
20160312_02_A1.10045.10045.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1004	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10116.10116.3 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1011	NIVTPR+0.984TPPPSQGK	3
20160312_02_A1.10222.10222.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1022	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10334.10334.3 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1033	NIVTPR+0.984TPPPSQGK	3
20160312_02_A1.10362.10362.3 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1036	TPSTAHLR+0.984VPK	3
20160312_02_A1.10418.10418.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1041	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10479.10479.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1047	AQSR+0.984EQLAALK	2
20160312_02_A1.1054.1054.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1054"	DSR+0.984SGSPM+15.995AR	2
20160312_02_A1.10602.10602.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1060	NIVTPR+0.984TPPPSQGK	2
20160312_02_A1.10646.10646.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1064	Q+0.984KR+0.984LQ+0.984AM+15.995Q+0.984K	2
20160312_02_A1.10671.10671.3 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1067	SGSEAGSPRR+0.984PRRQR	3
20160312_02_A1.1073.1073.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1073"	R+0.984GGGGGRR+0.984SK	2
20160312_02_A1.10764.10764.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1076	R+0.984FIN+0.984DMVK	2
20160312_02_A1.10769.10769.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1076	NIVTPR+0.984TPPPSQ+0.984GK	2
20160312_02_A1.10874.10874.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1087	MAR+0.984EAEFEAEQER	2
20160312_02_A1.11026.11026.3 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1102	RGR+0.984PPKDEK	3
20160312_02_A1.11286.11286.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1128	N+0.984R+0.984Q+0.984VIC+57.021VTLK	2
20160312_02_A1.11398.11398.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1139	GGTSR+0.984ALAAASSVK	2
20160312_02_A1.11489.11489.3 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1148	EEFER+0.984Q+0.984N+0.984KQLR	3
20160312_02_A1.11557.11557.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1155	TVEMR+0.984DGEVIK	2
20160312_02_A1.11735.11735.2 File:"20160312_02_A1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1173	Q+0.984Q+0.984IADLR+0.984EDLKR	2

search_result_file

Format for peptide sequences should follow that of MS-GF+ search result. Specifically, modification delta masses should be rounded up to third decimal places. Currently allowed modifications are as follows:

Modification	Mod on peptide
Carbamidomethyl Cys	C+57.021
Oxidation Met	M+15.995
Deamidated Asn	N+0.984
Deamidated Gln	Q+0.984
Citrullinated Arg	R+0.984
Pyro-Glu from Glu	E-17.027
Pyro-Glu from Gln	Q-18.011
iTRAQ 4plex Lys	K+144.102
iTRAQ 8plex Lys	K+304.205
TMT Lys	K+229.163
iTRAQ 4plex N-term	+144.102
iTRAQ 8plex N-term	+304.205
TMT N-term	+229.163
Acetyl N-term	+42.011

Step 1: Input data

B. Loading input data

A snapshot of codes for loading input files:

1. Input files

```
In [ ]:  # Set current working directory
         PATH = "F:/Project/"
         os.chdir(PATH)
```

```
In [ ]:  # Input files
         spec_files   = glob.glob('spectrum_file.mgf') # MGF file(s)
         search_files = glob.glob('search_result_file.csv') # Search result file(s)
```

Users can upload local input files via the following steps:

- Set the directory in which the input files are located.

- Type the input filenames. In case of multiple MGF or search result files in the same directory, type in '*.mgf' or '*.csv'.

A snapshot of loaded MGF file:

```
In [38]:  df_exp
```

Out[38]:

	Title	m/z array	intensity array
0	20160312_02_A1.93.93.2 File:"20160312_02_A1.ra...	[113.0715067, 113.6931047, 114.1196269, 115.08...	[215.7922515869, 159.3930053711, 169.514968872...
1	20160312_02_A1.206.206.2 File:"20160312_02_A1....	[110.0715323, 113.0712549, 114.0551851, 115.08...	[171.9085998535, 2270.2216796875, 213.05671691...
2	20160312_02_A1.323.323.2 File:"20160312_02_A1....	[112.3516646, 113.0712827, 114.0551846, 115.08...	[162.815612793, 3220.46875, 206.597442627, 867...
3	20160312_02_A1.441.441.2 File:"20160312_02_A1....	[113.0715639, 114.0551846, 114.1746251, 115.05...	[3046.6838378906, 329.9189758301, 160.89303588...
4	20160312_02_A1.559.559.2 File:"20160312_02_A1....	[113.0714846, 114.0555112, 114.3805672, 114.46...	[3265.0134277344, 383.2307434082, 174.38641357...
...	...	...	...
390	20160312_04_A3.27367.27367.4 File:"20160312_04...	[110.071535, 111.0363426, 112.087122, 120.0813...	[2331.1821289063, 355.9834289551, 505.66485595...
391	20160312_04_A3.28271.28271.4 File:"20160312_04...	[110.0715594, 110.5183569, 111.074985, 114.091...	[4666.2412109375, 348.894744873, 435.891632080...
392	20160312_04_A3.29399.29399.3 File:"20160312_04...	[110.0716261, 111.0749855, 114.0918189, 116.04...	[9665.826171875, 666.3602294922, 485.172882080...
393	20160312_04_A3.29644.29644.3 File:"20160312_04...	[110.071646, 111.0375996, 112.2654049, 112.639...	[1841.1348876953, 159.2207641602, 165.47415161...
394	20160312_04_A3.30248.30248.2 File:"20160312_04...	[110.0715359, 115.0868826, 115.2165788, 116.03...	[2713.2185058594, 989.6865844727, 373.98999023...

395 rows × 3 columns

A snapshot of loaded search result file:

```
In [44]:  df
```

Out[44]:

	Title	Peptide	Charge
0	20160312_02_A1.10012.10012.2 File:"20160312_02...	YETSGIGEAR+0.984VK	2
1	20160312_02_A1.10045.10045.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	2
2	20160312_02_A1.10116.10116.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	3
3	20160312_02_A1.10222.10222.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	2
4	20160312_02_A1.10334.10334.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	3
...	...	...	...
94	20160312_02_A1.323.323.2 File:"20160312_02_A1....	DSR+0.984SGSPM+15.995AR	2
95	20160312_02_A1.32378.32378.4 File:"20160312_02...	GHIEWPDFLSHESLLLLQ+0.984Q+0.984LR+0.984PQNSLLR	4
96	20160312_02_A1.3280.3280.2 File:"20160312_02_A...	VSR+0.984VASPK	2
97	20160312_02_A1.33227.33227.2 File:"20160312_02...	AR+0.984HRKTGQ+0.984KVALK	2
98	20160312_02_A1.33883.33883.2 File:"20160312_02...	RERVR+0.984QLR	2

99 rows × 3 columns

Step 2: Theoretical spectrum generation

Theoretical spectrum is generated for each Cit PSM in the following order:

- Singly charged sequence and internal ions (e.g., *y*- and *b*-ions for sequence or internal ions).

- Neutral loss variants of precursor, sequence, and internal ions.

- Multiply charged sequence ions.

- Immonium ions.

A snapshot of input data:

```
In [20]: df_pep
```

Out[20]:

	Title	Peptide	mod_Peptide	seq_Peptid
0	20160312_02_A1.10012.10012.2 File:"20160312_02...	YETSGIGEAR+0.984VK	YETSGIGEArVK	YETSGIGEArV
1	20160312_02_A1.10045.10045.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
2	20160312_02_A1.10116.10116.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
3	20160312_02_A1.10222.10222.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
4	20160312_02_A1.10334.10334.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
...	...	...	...	
94	20160312_02_A1.323.323.2 File:"20160312_02_A1....	DSR+0.984SGSPM+15.995AR	DSrSGSPmAR	DSrSGSPmA
95	20160312_02_A1.32378.32378.4 File:"20160312_02...	GHIEWPDFLSHESLLLLQ+0.984Q+0.984LR+0.984PQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLL
96	20160312_02_A1.3280.3280.2 File:"20160312_02_A...	VSR+0.984VASPK	VSrVASPK	VSrVASF
97	20160312_02_A1.33227.33227.2 File:"20160312_02...	AR+0.984HRKTGQ+0.984KVALK	ArHRKTGqKVALK	ArHRKTGqKVAL
98	20160312_02_A1.33883.33883.2 File:"20160312_02...	RERVR+0.984QLR	RERVrQLR	RERVrQL

99 rows × 8 columns

A snapshot of output data:

In [35]: df_pep_mz

Out[35]:

	Title	Peptide	mod_Peptide	seq_Pepti
0	20160312_02_A1.10012.10012.2 File:"20160312_02...	YETSGIGEAR+0.984VK	YETSGIGEArVK	YETSGIGEAr\
1	20160312_02_A1.10045.10045.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
2	20160312_02_A1.10116.10116.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
3	20160312_02_A1.10222.10222.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
4	20160312_02_A1.10334.10334.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
...	...	...	...	
94	20160312_02_A1.323.323.2 File:"20160312_02_A1....	DSR+0.984SGSPM+15.995AR	DSrSGSPmAR	DSrSGSPmA
95	20160312_02_A1.32378.32378.4 File:"20160312_02...	GHIEWPDFLSHESLLLLQ+0.984Q+0.984LR+0.984PQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLL
96	20160312_02_A1.3280.3280.2 File:"20160312_02_A...	VSR+0.984VASPK	VSrVASPK	VSrVASF
97	20160312_02_A1.33227.33227.2 File:"20160312_02...	AR+0.984HRKTGQ+0.984KVALK	ArHRKTGqKVALK	ArHRKTGqKVAL
98	20160312_02_A1.33883.33883.2 File:"20160312_02...	RERVR+0.984QLR	RERVrQLR	RERVrQl

99 rows × 20650 columns

In [36]: df_pep_label

Out[36]:

	Title	Peptide	mod_Peptide	seq_Pepti
0	20160312_02_A1.10012.10012.2 File:"20160312_02...	YETSGIGEAR+0.984VK	YETSGIGEArVK	YETSGIGEAr\
1	20160312_02_A1.10045.10045.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
2	20160312_02_A1.10116.10116.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
3	20160312_02_A1.10222.10222.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
4	20160312_02_A1.10334.10334.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	NIVTPrTPPPSQG
...	...	...	...	
94	20160312_02_A1.323.323.2 File:"20160312_02_A1....	DSR+0.984SGSPM+15.995AR	DSrSGSPmAR	DSrSGSPmA
95	20160312_02_A1.32378.32378.4 File:"20160312_02...	GHIEWPDFLSHESLLLLQ+0.984Q+0.984LR+0.984PQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLL
96	20160312_02_A1.3280.3280.2 File:"20160312_02_A...	VSR+0.984VASPK	VSrVASPK	VSrVASF
97	20160312_02_A1.33227.33227.2 File:"20160312_02...	AR+0.984HRKTGQ+0.984KVALK	ArHRKTGqKVALK	ArHRKTGqKVAL
98	20160312_02_A1.33883.33883.2 File:"20160312_02...	RERVR+0.984QLR	RERVrQLR	RERVrQl

99 rows × 20650 columns

Step 3: Match theoretical spectrum with experimental spectrum

Theoretical and experimental spectra are compared, and only the matched ions are retained.

A snapshot of input data:

```
In [53]: df_exp_mz

Out[53]: 0      [115.0868788, 129.1025438, 130.050427, 148.604...
         1      [113.0712549, 115.0870324, 129.1025438, 130.05...
         2      [113.0712827, 115.0869625, 129.1025433, 130.05...
         3      [113.0715639, 115.0507616, 115.0869447, 116.09...
         4      [113.0714846, 115.0507612, 115.0872086, 129.10...
                                     ...
         390    [110.071535, 120.0813037, 121.0843465, 129.102...
         391    [110.0715594, 115.0868819, 120.0813009, 121.08...
         392    [110.0716261, 111.0749855, 120.0813701, 143.11...
         393    [110.071646, 120.0808807, 129.1025478, 130.086...
         394    [110.0715359, 120.0813789, 121.0842901, 129.06...
         Name: m/z array, Length: 395, dtype: object
```

```
In [50]: df_theo_mz
```

Out[50]:

	mz_Precursor	mz_y_1	mz_y_2	mz_y_3	mz_y_4	mz_y_5	mz_y_6	mz_y_7	mz_y_8	mz_y_9	...	mz_IM-NH3_21	mz_IM-NH3_22
0	655.833519	147.113353	246.181767	403.266894	474.304008	603.346601	660.368065	773.452129	830.473593	917.505621	...	NaN	NaN
1	746.910097	147.113353	204.134817	332.193395	419.225423	516.278187	613.330951	710.383715	811.431394	968.516521	...	NaN	NaN
2	498.276006	147.113353	204.134817	332.193395	419.225423	516.278187	613.330951	710.383715	811.431394	968.516521	...	NaN	NaN
3	746.910097	147.113353	204.134817	332.193395	419.225423	516.278187	613.330951	710.383715	811.431394	968.516521	...	NaN	NaN
4	498.276006	147.113353	204.134817	332.193395	419.225423	516.278187	613.330951	710.383715	811.431394	968.516521	...	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
94	540.741111	175.119501	246.156615	393.192015	490.244779	577.276807	634.298271	721.330299	878.415426	965.447454	...	NaN	NaN
95	836.443032	175.119501	288.203565	401.287629	488.319657	602.362584	730.421162	827.473926	984.559053	1097.643117	...	113.071488	53.039125 8
96	422.748533	147.113353	244.166117	331.198145	402.235259	501.303673	658.388800	745.420828	NaN	NaN	...	NaN	NaN
97	747.947348	147.113353	260.197417	331.234531	430.302945	558.397908	687.440502	744.461966	845.509645	973.604608	...	NaN	NaN
98	557.334162	175.119501	288.203565	416.262143	573.347270	672.415684	828.516795	957.559388	NaN	NaN	...	NaN	NaN

99 rows × 20178 columns

```
In [51]: df_theo_label
```

Out[51]:

	mz_Precursor	mz_y_1	mz_y_2	mz_y_3	mz_y_4	mz_y_5	mz_y_6	mz_y_7	mz_y_8	mz_y_9	...	mz_IM-NH3_21	mz_IM-NH3_22	mz_IM-NH3_23	mz_IM-NH3_24	mz_IM-NH3_25	mz_IM-NH3_26
0	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	NaN	NaN	NaN	NaN	NaN	NaN
1	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	NaN	NaN	NaN	NaN	NaN	NaN
2	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	NaN	NaN	NaN	NaN	NaN	NaN
3	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	NaN	NaN	NaN	NaN	NaN	NaN
4	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
94	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	NaN	NaN	NaN	NaN	NaN	NaN
95	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	IM(r)-NH3	IM(P)-NH3	IM(Q)-NH3	IM(N)-NH3	IM(S)-NH3	IM(L)-NH3
96	Precursor	y1	y2	y3	y4	y5	y6	y7	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
97	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	NaN	NaN	NaN	NaN	NaN	NaN
98	Precursor	y1	y2	y3	y4	y5	y6	y7	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN

99 rows × 20178 columns

A snapshot of output data (matched ions):

```
In [39]: df_mz_label
```

Out[39]:

	Title	Peptide	mod_Peptide	Pep_length	Charge	mz_Precurs
0	20160312_02_A1.10012.10012.2 File:"20160312_02...	YETSGIGEAR+0.984VK	YETSGIGEArVK	12	2	655.8335
1	20160312_02_A1.10045.10045.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	2	746.9100
2	20160312_02_A1.10116.10116.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	3	498.2760
3	20160312_02_A1.10222.10222.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	2	746.9100
4	20160312_02_A1.10334.10334.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	3	498.2760
...	...	...	...	...	...	
94	20160312_02_A1.323.323.2 File:"20160312_02_A1....	DSR+0.984SGSPM+15.995AR	DSrSGSPmAR	10	2	540.7411
95	20160312_02_A1.32378.32378.4 File:"20160312_02...	GHIEWPDFLSHESLLLLQ+0.984Q+0.984LR+0.984PQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLLR	28	4	836.4430
96	20160312_02_A1.3280.3280.2 File:"20160312_02_A...	VSR+0.984VASPK	VSrVASPK	8	2	422.7485
97	20160312_02_A1.33227.33227.2 File:"20160312_02...	AR+0.984HRKTGQ+0.984KVALK	ArHRKTGqKVALK	13	2	747.9473
98	20160312_02_A1.33883.33883.2 File:"20160312_02...	RERVR+0.984QLR	RERVrQLR	8	2	557.3341

99 rows × 289 columns

Step 4: Cit diagnostic ion analysis

Annotations and occurrence numbers of citrullination diagnostic ions are reported for each Cit PSM.

A snapshot of input data:

In [41]: `df_mz_label_uniq`

Out[41]:

	Title	Peptide	mod_Peptide	Pep_length	Charge	mz_Precurs
0	20160312_02_A1.10012.10012.2 File:"20160312_02...	YETSGIGEAR+0.984VK	YETSGIGEArVK	12	2	655.8335
1	20160312_02_A1.10045.10045.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	2	746.9100
2	20160312_02_A1.10116.10116.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	3	498.2760
3	20160312_02_A1.10222.10222.2 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	2	746.9100
4	20160312_02_A1.10334.10334.3 File:"20160312_02...	NIVTPR+0.984TPPPSQGK	NIVTPrTPPPSQGK	14	3	498.2760
...	...	...	...	...	...	
94	20160312_02_A1.323.323.2 File:"20160312_02_A1....	DSR+0.984SGSPM+15.995AR	DSrSGSPmAR	10	2	540.7411
95	20160312_02_A1.32378.32378.4 File:"20160312_02...	GHIEWPDFLSHESLLLLQ+0.984Q+0.984LR+0.984PQNSLLR	GHIEWPDFLSHESLLLLqqLrPQNSLLR	28	4	836.4430
96	20160312_02_A1.3280.3280.2 File:"20160312_02_A...	VSR+0.984VASPK	VSrVASPK	8	2	422.7485
97	20160312_02_A1.33227.33227.2 File:"20160312_02...	AR+0.984HRKTGQ+0.984KVALK	ArHRKTGqKVALK	13	2	747.9473
98	20160312_02_A1.33883.33883.2 File:"20160312_02...	RERVR+0.984QLR	RERVrQLR	8	2	557.3341

99 rows × 289 columns

A snapshot of output data (binary occurrence for the immonium ion and occurrence numbers for the other diagnostic ions):

In [66]: `Final_result`

Out[66]:

el	precNL_label	seqNL_label	...	Total_NL_count	precNL_count	seqNL_count	intNL_count	Total_INT_count	Dipeptide_count	Tripeptide_count	IM_NH3_count
3- r- r- ...		y5-43-NH3++,y3-43,y4-43,y10-43-H2O-H2O++,y10-4...	...	48	0	23	25	10	4	6	0
T- r- ...		y10-43-H2O++,y10-43++,y11-43-H2O++,y11-43++,y1...	...	38	0	15	23	14	8	6	0
T- r- ...		b7-43-H2O	...	20	0	1	19	15	9	6	0
T- r- ...		y9-43-H2O++,y9-43++,y10-43-H2O++,y10-43++,y11-...	...	56	0	23	33	17	11	6	0
T- r- ...		b7-43-H2O	...	26	0	1	25	14	7	7	0
...		...	...	...	...	...	...	...	...	...	...

Step 5: Evaluation of Cit PSMs by EN model

The EN model developed in the study is applied to evaluate the validity of each Cit PSM.

A snapshot of input data:

In [66]: Final_result

Out[66]:

el	precNL_label	seqNL_label	...	Total_NL_count	precNL_count	seqNL_count	intNL_count	Total_INT_count	Dipeptide_count	Tripeptide_count	IM_NH3_count
3- ır- ır- ...		y5-43- NH3++,y3- 43,y4- 43,y10-43- H2O- H2O++,y10- 4...	...	48	0	23	25	10	4	6	0
T- 'r- :...		y10-43- H2O++,y10- 43++,y11-43- H2O++,y11- 43++,y1...	...	38	0	15	23	14	8	6	0
T- 'r- ...		b7-43-H2O	...	20	0	1	19	15	9	6	0
T- 'r- ...		y9-43- H2O++,y9- 43++,y10- 43- H2O++,y10- 43++,y11-...	...	56	0	23	33	17	11	6	0
T- 'r- ...		b7-43-H2O	...	26	0	1	25	14	7	7	0
...		...	...	...	...	...	...	...	...	...	...

A snapshot of output data (Cit_probability, probability that a PSM contains Cit; and Cit_prediction = 1 when Cit_probability > 0.5):

In [65]: Final_result

Out[65]:

...	Total_NL_count	precNL_count	seqNL_count	intNL_count	Total_INT_count	Dipeptide_count	Tripeptide_count	IM_NH3_count	Cit_probability	Cit_prediction
...	48	0	23	25	10	4	6	0	1.000000	1
...	38	0	15	23	14	8	6	0	1.000000	1
...	20	0	1	19	15	9	6	0	1.000000	1
...	56	0	23	33	17	11	6	0	1.000000	1
...	26	0	1	25	14	7	7	0	1.000000	1
...	...	...	...	...	...	...	...	...	...	...

Step 5: Evaluation of Cit PSMs by EN model

Example output CSV file including the prediction results (Cit_probabilty and Cit_prediction):

Descriptions of the columns in the output CSV file:

Column	Description
Title	MS2 spectrum title
Peptide	Original peptide sequence
mod_peptide	Simplified peptide sequence with a predefined set of symbols for modifications
Pep_length	Peptide length
mz_Precursor	Theoretical precursor m/z
Cit_Count	Number of citrullinated sites
Total_NL_label	Annotations of all diagnostic neutral loss ions
precNL_label	Annotations of precursor neutral losses
seqNL_label	Annotations of sequence ion neutral losses
intNL_label	Annotations of internal ion neutral losses
Total_INT_label	Annotations of all diagnostic internal ions
Dipeptide_label	Annotations of diagnostic dipeptides
Tripeptide_label	Annotations of diagnostic tripeptides
IM_NH3_label	Annotation of IM(Cit)-NH$_3$
Total_NL_count	Number of all diagnostic neutral loss ions
precNL_count	Number of precursor neutral losses
seqNL_count	Number of sequence ion neutral losses
intNL_count	Number of internal ion neutral losses
Total_INT_count	Number of all diagnostic internal ions
Dipeptide_count	Number of diagnostic dipeptides
Tripeptide_count	Number of diagnostic tripeptides
IM_NH3_count	Number of IM(Cit)-NH$_3$
Cit_probability	Probability (P) of citrullination status calculated by the EN model (HCD data only)
Cit_prediction	Classification of citrullination status using a P cutoff >0.5 (HCD data only)

Chapter 2. Prediction of False Negative Citrullinated PSMs

Download

To download Citrullination analysis.ipynb:

- Go to https://github.com/Sunghyun-Huh/Citrullination-Diagnostic-Ion-Analysis.
- Click on 'False negative prediction.ipynb' among the listed files.
- Click 'Raw' on the top right panel.
- Press ctrl+s and type 'False negative prediction.ipynb' to keep the ipynb extension.

Requirements

Python requirements

- Python version 3.6 or greater
- Libraries needed: pandas, numpy, pyteomics, itertools, collections, statistics, re, os, glob, scipy.stats

Necessary Files

- Unmodified peptide file: The input file must be a CSV file containing the following column (see **Step 1** for the details):

  'Peptide'   Peptide sequence with modification delta mass rounded up to third decimal places

- Total PSM file: The input file must be a CSV file containing the following column (see **Step 1** for the details):

  'Title'   MS2 spectrum title as written in MGF file

- Spectrum file: The spectrum file must be a Mascot Generic Format (MGF) file(s) containing either total MS2 spectra or MS2 spectra unassigned by database searching only. For the former case, the codes will automatically search for unassigned MS2 spectra.

Overview flowchart
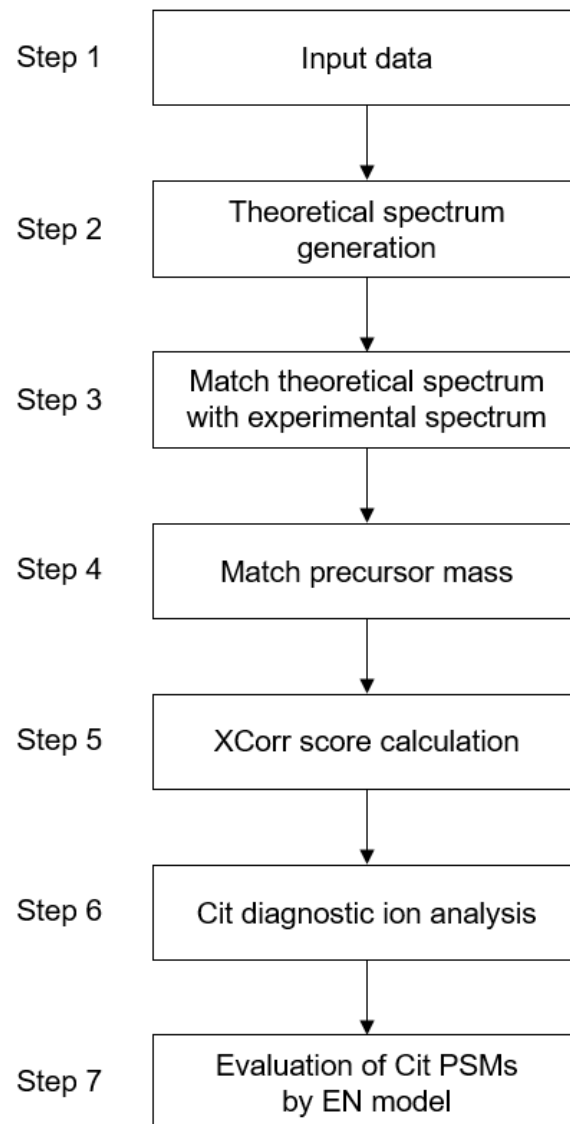
Step 1	Input data
Step 2	Theoretical spectrum generation
Step 3	Match theoretical spectrum with experimental spectrum
Step 4	Match precursor mass
Step 5	XCorr score calculation
Step 6	Cit diagnostic ion analysis
Step 7	Evaluation of Cit PSMs by EN model

Step 1: Input data

A. Setting initial parameters

The initial parameter settings:

0. Initial parameters

```
In [95]:  # Fragmentation method used
          Frag_method = 'HCD' # HCD, CID, ETD, ECD

          # Ion types for theoretical spectrum generation
          ion_type = ['Precur', 'y', 'b', 'a', 'z', 'c', 'INT', 'IM']

          # Generate each ion type
          if Frag_method == 'HCD' or Frag_method == 'CID':
              annot_yb    = True # y-, b-, a-ion
              annot_zc    = False # z-, c-ion
          elif Frag_method == 'ETD' or Frag_method == 'ECD':
              annot_yb    = False # y-, b-, a-ion
              annot_zc    = True # z-, c-ion
          annot_precur = True # precursor ion
          annot_INT    = True # internal ion
          annot_IM     = True # immonium ion
          annot_dict = {
              'Precur' : annot_precur,
              'y'      : annot_yb,
              'b'      : annot_yb,
              'a'      : annot_yb,
              'z'      : annot_zc,
              'c'      : annot_zc,
              'INT'    : annot_INT,
              'IM'     : annot_IM
          }

          # Mass tolerance (ppm)
          ms1_ppm = 10 # MS1 level
          ms2_ppm = 15 # MS2 level

          # Signal-to-noise (SNR) filter for MS2 spectrum
          apply_SNR = True # Apply SNR filter
          SNR       = 2 # SNR threshold
          low       = 0.05 # Define low x% intensity as baseline noise level

          # Maximum charge state of sequence ions
          max_charge = 2

          # Maximum number of neutral loss from a single ion
          max_NL = 3

          # Threshold for intensity coverage filter
          intensity_coverage = 0.2 # Filter out spectra with sum of annotated intensities < x% of sum of total intensities
```

Frag_method	Fragmentation method used in the input data (value = 'HCD', 'CID', 'ETD', or 'ECD'). If set as 'HCD' or 'CID', y-ion, b-ion, and a-ion will be generated for theoretical spectrum. If set as 'ETD' or 'ECD', z-ion and c-ion will be generated for theoretical spectrum. Commonly, precursor, internal, and immonium ions will be generated for all these fragmentation methods.
ms1_ppm	MS1 level mass tolerance in ppm (default = 10 ppm).
ms2_ppm	MS2 level mass tolerance in ppm (default = 15 ppm).
apply_SNR	Determine whether to apply signal-to-noise filter to remove noise peaks (value = True or False; default = True).
SNR	Signal threshold level. The average intensity of noise peaks (as defined in 'low') multiplied by this signal threshold level will be the final signal-to-noise filter. If 'apply_SNR' = True, all peaks below the signal-to-noise filter will be removed (default = 2).
low	Proportion of MS2 peaks regarded as noise. If 'apply_SNR' = True, all peaks below this noise level will be treated as noise (default = 0.05).
max_charge	Maximum charge state of fragment ions (default = 2).
max_NL	Maximum number of neutral losses from a single ion (default = 3).
intensity_coverage	Threshold for intensity coverage filter. Spectra with [sum of annotated intensities / sum of total intensities] smaller than this threshold will be filtered out (default = 0.2).

Step 1: Input data

B. Loading input data

Example Unmodified peptide file:

A
Peptide
TQEEAIVK
TSFADGK
KYEGDIK
KYEGDIK
KYEGDIK
SKDQGATYQK
TQEEAIVK
TQEEAIVK
YTPVEEK
TSFADGK
SKDQGATYQK
NGAAQAVTAENK
LYENTQDYDK
ITTAQEM+15.995YDK

Format for peptide sequences should follow that of MS-GF+ search result. Specifically, modification delta masses should be rounded up to third decimal places. Currently allowed modifications are as follows:

Modification	Mod on peptide
Carbamidomethyl Cys	C+57.021
Oxidation Met	M+15.995
Deamidated Asn	N+0.984
Deamidated Gln	Q+0.984
Citrullinated Arg	R+0.984
Pyro-Glu from Glu	E-17.027
Pyro-Glu from Gln	Q-18.011
iTRAQ 4plex Lys	K+144.102
iTRAQ 8plex Lys	K+304.205
TMT Lys	K+229.163
iTRAQ 4plex N-term	+144.102
iTRAQ 8plex N-term	+304.205
TMT N-term	+229.163
Acetyl N-term	+42.011

Step 1: Input data

B. Loading input data

Example Total PSM file:

A
Title
QEHF1_10122_DNL.4368.4368.6 File:"QEHF1_10122_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4368"
QEHF1_10122_DNL.4396.4396.6 File:"QEHF1_10122_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4396"
QEHF1_10123_DNL.4596.4596.6 File:"QEHF1_10123_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4596"
QEHF1_10123_DNL.4434.4434.6 File:"QEHF1_10123_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4434"
QEHF1_10124_DNL.4668.4668.6 File:"QEHF1_10124_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4668"
QEHF1_10125_DNL.4356.4356.6 File:"QEHF1_10125_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4356"
QEHF1_10125_DNL.4383.4383.6 File:"QEHF1_10125_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4383"
QEHF1_10125_DNL.4601.4601.6 File:"QEHF1_10125_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4601"
QEHF1_10247_DNL.4239.4239.6 File:"QEHF1_10247_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4239"
QEHF1_10250_DNL.4480.4480.6 File:"QEHF1_10250_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4480"
QEHF1_10250_DNL.4272.4272.6 File:"QEHF1_10250_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4272"
QEHF1_10275_DNL.3539.3539.6 File:"QEHF1_10275_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=3539"
QEHF1_10280_DNL.4501.4501.6 File:"QEHF1_10280_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4501"
QEHF1_10280_DNL.4474.4474.6 File:"QEHF1_10280_DNL.raw", NativeID:"controllerType=0 controllerNumber=1 scan=4474"

Step 1: Input data

B. Loading input data

A snapshot of codes for loading input files:

1. Input files

```
In [3]: # Set current working directory
        PATH = "F:/Project/"
        os.chdir(PATH)
```

```
In [4]: # Input files
        spec_files = glob.glob('*.mgf')[0:5] # MGF file(s)
        df_UnmodR  = pd.read_csv("UnmodR_peptide.csv") # UnmodR peptides w/o Cit counterpart peptides
        df_total   = pd.read_csv("Total_peptide.csv") # Total identified peptides
```

Users can upload local input files via the following steps:

- Set the directory in which the input files are located.

- Type the input filenames. In case of multiple MGF files in the same directory, type in '*.mgf'.

A snapshot of loaded MGF file:

```
In [105]: df_exp

Out[105]:
```

	m/z array	intensity array	Title	exp_Precursor	Charge
0	[108.9069509, 108.9075611, 108.9081714, 111.03...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 18287.58203125,...	QEHF1_10098_DNL.2.2.1 File:"QEHF1_10098_DNL.ra...	371.101932	1
1	[108.9068349, 108.9074451, 108.9080554, 110.10...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1604.9095458984...	QEHF1_10098_DNL.3.3.1 File:"QEHF1_10098_DNL.ra...	445.121060	1
2	[108.9068088, 108.907419, 108.9080293, 121.956...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2023.1558837891...	QEHF1_10098_DNL.4.4.1 File:"QEHF1_10098_DNL.ra...	536.166752	1
3	[108.9067553, 108.9073656, 108.9079758, 149.03...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 61633.01953125,...	QEHF1_10098_DNL.5.5.1 File:"QEHF1_10098_DNL.ra...	462.147507	1
4	[108.9068558, 108.907466, 108.9080763, 116.332...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1553.5465087891...	QEHF1_10098_DNL.6.6.1 File:"QEHF1_10098_DNL.ra...	593.159294	1
...	...	...	...	...	...
34469	[108.9067848, 108.907395, 108.9080053, 110.335...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2188.8542480469...	QEHF1_10102_DNL.8264.8264.1 File:"QEHF1_10102_...	751.511179	1
34470	[108.9067921, 108.9074023, 108.9080126, 112.55...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 3034.580078125,...	QEHF1_10102_DNL.8265.8265.1 File:"QEHF1_10102_...	1143.789214	1
34471	[108.9068589, 108.9074691, 108.9080794, 111.03...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2754.2233886719...	QEHF1_10102_DNL.8266.8266.1 File:"QEHF1_10102_...	359.314999	1
34472	[108.9067433, 108.9073536, 108.9079638, 121.02...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1498.1771240234...	QEHF1_10102_DNL.8267.8267.1 File:"QEHF1_10102_...	407.300978	1
34473	[108.9066313, 108.9072416, 108.9078518, 111.03...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2158.365234375,...	QEHF1_10102_DNL.8268.8268.1 File:"QEHF1_10102_...	387.192683	1

34474 rows × 5 columns

A snapshot of loaded Unmodified peptide file:

```
In [108]: df_UnmodR
```

Out[108]:

	Peptide
0	TQEEAIVK
1	TSFADGK
2	KYEGDIK
3	KYEGDIK
4	KYEGDIK
...	...
994	THNISEGM+15.995MGYDTYPK
995	THNISEGMMGYDTYPK
996	GAEVSVVNMEALQAER
997	EADGSVTHTFVIK
998	THNISEGMM+15.995GYDTYPK

999 rows × 1 columns

A snapshot of loaded Total PSM file:

```
df_total
```

	Title
0	QEHF1_10122_DNL.4368.4368.6 File:"QEHF1_10122_...
1	QEHF1_10122_DNL.4396.4396.6 File:"QEHF1_10122_...
2	QEHF1_10123_DNL.4596.4596.6 File:"QEHF1_10123_...
3	QEHF1_10123_DNL.4434.4434.6 File:"QEHF1_10123_...
4	QEHF1_10124_DNL.4668.4668.6 File:"QEHF1_10124_...
...	...
31391	QEHF3_06513_DNL.9237.9237.1 File:"QEHF3_06513_...
31392	QEHF3_06513_DNL.6988.6988.1 File:"QEHF3_06513_...
31393	QEHF3_06513_DNL.9195.9195.1 File:"QEHF3_06513_...
31394	QEHF3_06513_DNL.3945.3945.1 File:"QEHF3_06513_...
31395	QEHF3_06513_DNL.6917.6917.1 File:"QEHF3_06513_...

31396 rows × 1 columns

Step 2: Theoretical spectrum generation

Theoretical spectrum is generated for each Cit peptide as described in **Chapter 1**:

A snapshot of input data:

`df_pep`

	Peptide	mod_Peptide	seq_Peptide	Charge	Pep_length	mz_Precursor	Cit_Count
0	VPAPVDGER+0.984	VPAPVDGEr	VPAPVDGEr	1	9	940.474	1
1	NHNLYIAR+0.984	NHNLYIAr	NHNLYIAr	1	8	1001.52	1
2	NHNLYIAR+0.984	NHNLYIAr	NHNLYIAr	2	8	501.262	1
3	TAIIR+0.984YNYASGK	TAIIrYNYASGK	TAIIrYNYASGK	1	12	1357.71	1
4	TAIIR+0.984YNYASGK	TAIIrYNYASGK	TAIIrYNYASGK	2	12	679.36	1
...	...	...	...	...	...	...	...
204	YTFTMRR+0.984	YTFTMRr	YTFTMRr	2	7	488.24	1
205	TLEDNVALR+0.984ER	TLEDNVALrER	TLEDNVALrER	1	11	1316.68	1
206	TLEDNVALR+0.984ER	TLEDNVALrER	TLEDNVALrER	2	11	658.844	1
207	TLEDNVALRER+0.984	TLEDNVALREr	TLEDNVALREr	1	11	1316.68	1
208	TLEDNVALRER+0.984	TLEDNVALREr	TLEDNVALREr	2	11	658.844	1

209 rows × 7 columns

A snapshot of output data:

In [97]: `df_pep_mz`

Out[97]:

	Peptide	mod_Peptide	seq_Peptide	Charge	Pep_length	mz_Precursor	Cit_Count	seq_y_1	seq_y_2	seq_y_3	...	mz_IM-NH3_14	mz_IM-NH3_15	mk NH
0	VPAPVDGER+0.984	VPAPVDGEr	VPAPVDGEr	1	9	940.473987	1	r	rE	rEG	...	NaN	NaN	
1	NHNLYIAR+0.984	NHNLYIAr	NHNLYIAr	1	8	1001.516845	1	r	rA	rAI	...	NaN	NaN	
2	NHNLYIAR+0.984	NHNLYIAr	NHNLYIAr	2	8	501.262335	1	r	rA	rAI	...	NaN	NaN	
3	TAIIR+0.984YNYASGK	TAIIrYNYASGK	TAIIrYNYASGK	1	12	1357.711574	1	K	KG	KGS	...	NaN	NaN	
4	TAIIR+0.984YNYASGK	TAIIrYNYASGK	TAIIrYNYASGK	2	12	679.359699	1	K	KG	KGS	...	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
204	YTFTMRR+0.984	YTFTMRr	YTFTMRr	2	7	488.240015	1	r	rR	rRM	...	NaN	NaN	
205	TLEDNVALR+0.984ER	TLEDNVALrER	TLEDNVALrER	1	11	1316.681019	1	R	RE	REr	...	NaN	NaN	
206	TLEDNVALR+0.984ER	TLEDNVALrER	TLEDNVALrER	2	11	658.844422	1	R	RE	REr	...	NaN	NaN	
207	TLEDNVALRER+0.984	TLEDNVALREr	TLEDNVALREr	1	11	1316.681019	1	r	rE	rER	...	NaN	NaN	
208	TLEDNVALRER+0.984	TLEDNVALREr	TLEDNVALREr	2	11	658.844422	1	r	rE	rER	...	NaN	NaN	

209 rows × 6560 columns

In [98]: `df_pep_label`

Out[98]:

	Peptide	mod_Peptide	seq_Peptide	Charge	Pep_length	mz_Precursor	Cit_Count	seq_y_1	seq_y_2	seq_y_3	...	mz_IM-NH3_14	mz_IM-NH3_15	mk NH
0	VPAPVDGER+0.984	VPAPVDGEr	VPAPVDGEr	1	9	Precursor	1	y1	y2	y3	...	NaN	NaN	
1	NHNLYIAR+0.984	NHNLYIAr	NHNLYIAr	1	8	Precursor	1	y1	y2	y3	...	NaN	NaN	
2	NHNLYIAR+0.984	NHNLYIAr	NHNLYIAr	2	8	Precursor	1	y1	y2	y3	...	NaN	NaN	
3	TAIIR+0.984YNYASGK	TAIIrYNYASGK	TAIIrYNYASGK	1	12	Precursor	1	y1	y2	y3	...	NaN	NaN	
4	TAIIR+0.984YNYASGK	TAIIrYNYASGK	TAIIrYNYASGK	2	12	Precursor	1	y1	y2	y3	...	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
204	YTFTMRR+0.984	YTFTMRr	YTFTMRr	2	7	Precursor	1	y1	y2	y3	...	NaN	NaN	
205	TLEDNVALR+0.984ER	TLEDNVALrER	TLEDNVALrER	1	11	Precursor	1	y1	y2	y3	...	NaN	NaN	
206	TLEDNVALR+0.984ER	TLEDNVALrER	TLEDNVALrER	2	11	Precursor	1	y1	y2	y3	...	NaN	NaN	
207	TLEDNVALRER+0.984	TLEDNVALREr	TLEDNVALREr	1	11	Precursor	1	y1	y2	y3	...	NaN	NaN	
208	TLEDNVALRER+0.984	TLEDNVALREr	TLEDNVALREr	2	11	Precursor	1	y1	y2	y3	...	NaN	NaN	

209 rows × 6560 columns

Step 3: Match precursor mass

Precursor masses of theoretical and experimental spectra are compared, and only the matched spectra are retained.

A snapshot of input data:

```
In [113]: Mass_theo
Out[113]:
```

	0
0	940.473987
1	1001.516845
2	1002.524670
3	1357.711574
4	1358.719399
...	...
204	976.480030
205	1316.681019
206	1317.688844
207	1316.681019
208	1317.688844

209 rows × 1 columns

```
In [114]: Mass_exp
Out[114]:
```

	0
0	371.101932
1	445.121060
2	536.166752
3	462.147507
4	593.159294
...	...
33123	751.511179
33124	1143.789214
33125	359.314999
33126	407.300978
33127	387.192683

33128 rows × 1 columns

A snapshot of output data:

```
In [119]: df_unID_scan['Mass_Check']
Out[119]: 0        0
          1        0
          2        0
          3        0
          4        0
                  ..
          33123    0
          33124    0
          33125    0
          33126    0
          33127    0
          Name: Mass_Check, Length: 33128, dtype: int64
```

Step 4: XCorr score calculation

For the matching experimental spectrum and theoretical spectrum (or spectra), XCorr score is calculated and only the theoretical spectrum with the greatest XCorr score is retained.

A snapshot of input data:

In [123]: df_unID_scan_filter

mz_intensity	Mass_Check	matched_theo_spectrum	theo_mz_array	Base_peak	norm intensity array	XCorr	Best_matched_theo_spectrum
[[110.0720135, 42529.13671875], [116.0712452, ...	1	[11]	[[110.07182241250999, 113.07148831149999, 115....	5.778410e+04	[0.7360007068337424, 0.06188990719697581, 0.58...	[0.4645181088909756]	11
[[110.0718416, 39672.3046875], [113.0715189, 6...	1	[12]	[[110.07182241250999, 113.07148831149999, 115....	6.839465e+05	[0.058004982388973406, 0.008893030278821224, 0...	[1.4735176389345732]	12
[[110.071755, 149039.875], [113.0714059, 94425...	1	[95]	[[110.07182241250999, 113.07148831149999, 117....	1.271601e+06	[0.11720645524687869, 0.07425741388072715, 0.0...	[7.076248489273531]	95
[[126.0549671, 38965.9609375], [167.0820595, 7...	1	[97]	[[113.07148831149999, 119.0496813115, 130.0980...	4.437064e+06	[0.008781925464330181, 0.0016454831627381488, ...	[0.02522510458252917]	97

A snapshot of output data:

In [116]: df_pep_label_match

Out[116]:

	Index	Peptide	mod_Peptide	Charge	Pep_length	mz_Precursor	Cit_Count	label array	Best_matched_theo_spectrum
0	11	GGTHDPLQSVR+0.984	GGTHDPLQSVr	1	11	Precursor	1	[Precursor, y1, y2, y3, y4, y5, y6, y7, y8, y9...	11
1	12	GGTHDPLQSVR+0.984	GGTHDPLQSVr	2	11	Precursor	1	[Precursor, y1, y2, y3, y4, y5, y6, y7, y8, y9...	12
2	95	GSHEPVADNSTVAGR+0.984	GSHEPVADNSTVAGr	2	15	Precursor	1	[Precursor, y1, y2, y3, y4, y5, y6, y7, y8, y9...	95
3	97	YTPVEEKQNGR+0.984	YTPVEEKQNGr	2	11	Precursor	1	[Precursor, y1, y2, y3, y4, y5, y6, y7, y8, y9...	97
4	97	YTPVEEKQNGR+0.984	YTPVEEKQNGr	2	11	Precursor	1	[Precursor, y1, y2, y3, y4, y5, y6, y7, y8, y9...	97
								[Precursor, y1,	

Step 5: Match theoretical spectrum with experimental spectrum

Theoretical and experimental spectra are compared, and only the matched ions are retained.

A snapshot of input data:

```
In [125]: df_exp_mz

Out[125]: 0     [110.0720135, 116.0712452, 128.0707251, 130.05...
          1     [110.0718416, 113.0715189, 114.0552962, 115.02...
          2     [110.071755, 113.0714059, 114.0552635, 115.087...
          3     [126.0549671, 167.0820595, 167.1178537, 181.06...
          4     [126.0550001, 167.0809289, 167.1177837, 169.13...
          5     [123.9887742, 129.1026059, 137.0712778, 178.05...
          6     [110.0719872, 129.1024763, 130.086277, 147.112...
          7     [110.0607183, 110.0718668, 113.0715497, 114.05...
          8     [110.0719087, 113.0716098, 115.0871659, 116.07...
          9     [110.0601176, 110.0719138, 112.0220986, 115.02...
          10    [110.071834, 115.0214849, 115.0507414, 116.034...
          11    [113.0713727, 114.0551653, 115.0869482, 116.07...
          12    [113.0714384, 114.0552416, 115.0870154, 116.07...
          13    [110.0719918, 116.0705729, 127.0871257, 128.09...
          14    [110.0718952, 127.0871186, 128.0897561, 129.10...
          15    [110.0719309, 127.086452, 129.1025519, 130.086...
          16    [110.0719915, 116.071178, 127.087176, 129.1024...
          17    [110.0720804, 112.0219593, 127.0872803, 128.05...
          18    [110.0719435, 127.0871958, 128.0900352, 129.10...
          19    [110.0716643, 111.0918748, 112.0759924, 113.07...
          20    [110.0717597, 111.0920319, 113.0714655, 115.08...
          21    [110.0719004, 116.0710182, 122.0719581, 123.98...
          22    [110.071825, 115.0871277, 127.0871408, 130.050...
          23    [110.0718381, 115.0870345, 116.071251, 127.086...
          24    [110.0719391, 113.0715399, 114.0553174, 115.08...
          Name: m/z array, dtype: object
```

```
In [126]: df_theo_mz
```

Out[126]:

	0	1	2	3	4	5	6	7	8	9	...	1005	1006	
0	1167.575827	176.103517	275.171931	362.203959	490.262537	603.346601	700.399365	815.426308	952.485220	1053.532899	...	NaN	NaN	
1	584.291826	176.103517	275.171931	362.203959	490.262537	603.346601	700.399365	815.426308	952.485220	1053.532899	...	NaN	NaN	
2	749.350600	176.103517	233.124981	304.162095	403.230509	504.278188	591.310216	705.353143	820.380086	891.417200	...	55.054775	27.023475	71.
3	661.323318	176.103517	233.124981	347.167908	475.226486	603.321449	732.364042	861.406635	960.475049	1057.527813	...	NaN	NaN	
4	661.323318	176.103517	233.124981	347.167908	475.226486	603.321449	732.364042	861.406635	960.475049	1057.527813	...	NaN	NaN	
5	372.521397	147.113353	244.166117	373.208711	530.293838	627.346602	790.409922	927.468834	1014.500862	102.055504	...	NaN	NaN	
6	372.521397	147.113353	244.166117	373.208711	530.293838	627.346602	790.409922	927.468834	1014.500862	102.055504	...	NaN	NaN	
7	556.776343	176.103517	277.151196	348.188310	405.209774	533.268352	632.336766	733.384445	870.443357	984.486284	...	NaN	NaN	
8	556.776343	176.103517	277.151196	348.188310	405.209774	533.268352	632.336766	733.384445	870.443357	984.486284	...	NaN	NaN	
9	861.431787	176.103517	263.135545	320.157009	449.199602	506.221066	619.305130	732.389194	130.050418	243.134482	...	NaN	NaN	
10	861.431787	176.103517	263.135545	320.157009	449.199602	506.221066	619.305130	732.389194	130.050418	243.134482	...	NaN	NaN	
11	559.788584	176.103517	291.130460	404.214524	532.273102	633.320781	761.415744	874.499808	989.526751	130.050418	...	NaN	NaN	
12	559.788584	176.103517	291.130460	404.214524	532.273102	633.320781	761.415744	874.499808	989.526751	130.050418	...	NaN	NaN	
13	928.456229	176.103517	247.140631	378.181116	435.202580	506.239694	603.292458	716.376522	813.429286	116.034768	...	NaN	NaN	
14	928.456229	176.103517	247.140631	378.181116	435.202580	506.239694	603.292458	716.376522	813.429286	116.034768	...	NaN	NaN	
15	928.456229	176.103517	247.140631	378.181116	435.202580	506.239694	603.292458	716.376522	813.429286	116.034768	...	NaN	NaN	
16	928.456229	176.103517	247.140631	378.181116	435.202580	506.239694	603.292458	716.376522	813.429286	116.034768	...	NaN	NaN	
17	928.456229	176.103517	247.140631	378.181116	435.202580	506.239694	603.292458	716.376522	813.429286	116.034768	...	NaN	NaN	
18	928.456229	176.103517	247.140631	378.181116	435.202580	506.239694	603.292458	716.376522	813.429286	116.034768	...	NaN	NaN	
19	835.434765	176.103517	307.144002	378.181116	479.228795	594.255738	722.350701	114.091889	242.186852	357.213795	...	NaN	NaN	
20	835.434765	176.103517	307.144002	378.181116	479.228795	594.255738	722.350701	114.091889	242.186852	357.213795	...	NaN	NaN	
21	1285.606441	176.103517	277.151196	348.188310	477.230903	574.283667	737.346987	794.368451	923.411044	994.448158	...	NaN	NaN	
22	1325.627201	175.119501	290.146444	403.230508	502.298922	649.334322	777.392900	940.456220	1054.499147	1168.542074	...	NaN	NaN	
23	1325.627201	176.103517	291.130460	404.214524	503.282938	650.318338	778.376916	941.440236	1055.483163	1169.526090	...	NaN	NaN	
24	655.320055	176.103517	291.130460	404.214524	503.282938	634.323423	762.382001	925.445321	1039.488248	1153.531175	...	NaN	NaN	

25 rows × 1015 columns

```
In [127]: df_theo_label
```

Out[127]:

	0	1	2	3	4	5	6	7	8	9	...	1005	1006	1007	1008	1009	1010	1011	1012	1013	1014
0	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
1	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
2	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	IM(V)-NH3	IM(A)-NH3	IM(D)-NH3	IM(N)-NH3	IM(S)-NH3	IM(T)-NH3	IM(V)-NH3	IM(A)-NH3	IM(G)-NH3	IM(r)-NH3
3	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
4	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
5	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
6	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
7	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
8	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
9	Precursor	y1	y2	y3	y4	y5	y6	y7	b1	b2	...	None	None	None	None	None	None	None	None	None	None
10	Precursor	y1	y2	y3	y4	y5	y6	y7	b1	b2	...	None	None	None	None	None	None	None	None	None	None
11	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
12	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
13	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
14	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
15	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
16	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
17	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
18	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	b1	...	None	None	None	None	None	None	None	None	None	None
19	Precursor	y1	y2	y3	y4	y5	y6	b1	b2	b3	...	None	None	None	None	None	None	None	None	None	None
20	Precursor	y1	y2	y3	y4	y5	y6	b1	b2	b3	...	None	None	None	None	None	None	None	None	None	None
21	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
22	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
23	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None
24	Precursor	y1	y2	y3	y4	y5	y6	y7	y8	y9	...	None	None	None	None	None	None	None	None	None	None

25 rows × 1015 columns

A snapshot of output data (matched ions):

```
In [121]: df_mz_label
```

Out[121]:

	unID_Title	Peptide	mod_Peptide	Charge	Pep_length	mz_Precursor	Cit_Count	peak_1
0	QEHF1_10100_DNL.2385.2385.3 File:"QEHF1_10100_...	GGTHDPLQSVR+0.984	GGTHDPLQSVr	1	11	1167.575827	1	[110.0720135, 42529.13671875, 4833 IM(H)]
1	QEHF1_10102_DNL.3984.3984.2 File:"QEHF1_10102_...	GGTHDPLQSVR+0.984	GGTHDPLQSVr	2	11	584.291826	1	[110.0718416, 39672.3046875, 6082 IM(H)]
2	QEHF1_10101_DNL.3353.3353.2 File:"QEHF1_10101_...	GSHEPVADNSTVAGR+0.984	GSHEPVADNSTVAGr	2	15	749.350600	1	[110.071755, 149039.875, IM(H)] 94
3	QEHF1_10099_DNL.5220.5220.1 File:"QEHF1_10099_...	YTPVEEKQNGR+0.984	YTPVEEKQNGr	2	11	661.323318	1	[197.1288641, 56989.97265625, 71 PV]
4	QEHF1_10100_DNL.5261.5261.1 File:"QEHF1_10100_...	YTPVEEKQNGR+0.984	YTPVEEKQNGr	2	11	661.323318	1	[169.1335899, 8058.0991210938, 506 PV-COI

Step 6: Cit diagnostic ion analysis

Annotations and occurrence numbers of citrullination diagnostic ions are reported for each Cit PSM.

A snapshot of input data:

In [124]: `df_mz_label_uniq`

Out[124]:

	unID_Title	Peptide	mod_Peptide	Charge	Pep_length	mz_Precursor	Cit_Count	peak_1
0	QEHF1_10100_DNL.2385.2385.3 File:"QEHF1_10100_...	GGTHDPLQSVR+0.984	GGTHDPLQSVr	1	11	1167.575827	1	[110.0720135, 42529.13671875, 4833 IM(H)]
1	QEHF1_10102_DNL.3984.3984.2 File:"QEHF1_10102_...	GGTHDPLQSVR+0.984	GGTHDPLQSVr	2	11	584.291826	1	[110.0718416, 39672.3046875, 6082 IM(H)]
2	QEHF1_10101_DNL.3353.3353.2 File:"QEHF1_10101_...	GSHEPVADNSTVAGR+0.984	GSHEPVADNSTVAGr	2	15	749.350600	1	[110.071755, 149039.875, IM(H)] 94
3	QEHF1_10099_DNL.5220.5220.1 File:"QEHF1_10099_...	YTPVEEKQNGR+0.984	YTPVEEKQNGr	2	11	661.323318	1	[197.1288641, 56989.97265625, 71 PV]
4	QEHF1_10100_DNL.5261.5261.1 File:"QEHF1_10100_...	YTPVEEKQNGR+0.984	YTPVEEKQNGr	2	11	661.323318	1	[169.1335899, 8058.0991210938, 506 PV-CO]

A snapshot of output data (binary occurrence for the immonium ion and occurrence numbers for the other diagnostic ions):

In [130]: `Final_result_filter`

Out[130]:

bel	precNL_label	seqNL_label	...	Total_NL_count	precNL_count	seqNL_count	intNL_count	Total_INT_count	Dipeptide_count	Tripeptide_count	IM_NH3_count
3-0-43		y1-43,y2-43,y3-43,y4-43,y10-43-NH3	...	5	0	5	0	0	0	0	1
3-4-43		y1-43,y2-43-H2O,y2-43,y4-43-H2O,y4-43	...	5	0	5	0	0	0	0	1
2O		y2-43-H2O	...	1	0	1	0	0	0	0	1
3-3-3-!...		y1-43,y2-43-H2O,y2-43,y3-43-H2O,y3-43,y7-43-H2...	...	13	0	13	0	0	0	0	1
3-3-3-l...		y1-43,y2-43-H2O,y2-43,y3-43-H2O,y3-43,y4-43-NH...	...	9	0	9	0	0	0	0	1

Step 7: Evaluation of Cit PSMs by EN model

The EN model developed in the study is applied to evaluate the validity of each Cit PSM.

A snapshot of input data:

```
In [130]: Final_result_filter
```

Out[130]:

el	precNL_label	seqNL_label	...	Total_NL_count	precNL_count	seqNL_count	intNL_count	Total_INT_count	Dipeptide_count	Tripeptide_count	IM_NH3_count
3- 0- 43		y1-43,y2- 43,y3-43,y4- 43,y10-43- NH3	...	5	0	5	0	0	0	0	1
3- 4- 43		y1-43,y2-43- H2O,y2- 43,y4-43- H2O,y4-43	...	5	0	5	0	0	0	0	1
2O		y2-43-H2O	...	1	0	1	0	0	0	0	1
3- 3- 3- !...		y1-43,y2-43- H2O,y2- 43,y3-43- H2O,y3- 43,y7-43- H2...	...	13	0	13	0	0	0	0	1
3- 3- 3- I...		y1-43,y2-43- H2O,y2- 43,y3-43- H2O,y3- 43,y4-43- NH...	...	9	0	9	0	0	0	0	1

A snapshot of output data (Cit_probability, probability that a PSM contains Cit; and Cit_prediction = 1 when Cit_probability > 0.5):

```
In [131]: Final_result_filter
```

Out[131]:

...	Total_NL_count	precNL_count	seqNL_count	intNL_count	Total_INT_count	Dipeptide_count	Tripeptide_count	IM_NH3_count	Cit_probability	Cit_prediction
...	5	0	5	0	0	0	0	1	0.996773	1
...	5	0	5	0	0	0	0	1	0.996773	1
...	1	0	1	0	0	0	0	1	0.861678	1
...	13	0	13	0	0	0	0	1	0.999999	1
...	9	0	9	0	0	0	0	1	0.999935	1

Step 7: Evaluation of Cit PSMs by EN model

Example output CSV file including the prediction results (Cit_probabilty and Cit_prediction):

Descriptions of the columns in the output CSV file:

Column	Description
unID_Title	Unassigned MS2 spectrum title
Peptide	Original peptide sequence
mod_peptide	Simplified peptide with a predefined set of symbols for modifications
Charge	Charge state of precursor ion
Pep_length	Peptide length
mz_Precursor	Theoretical precursor m/z
Cit_Count	Number of citrullinated sites
Total_NL_label	Annotations of all diagnostic neutral loss ions
precNL_label	Annotations of precursor neutral losses
seqNL_label	Annotations of sequence ion neutral losses
intNL_label	Annotations of internal ion neutral losses
Total_INT_label	Annotations of all diagnostic internal ions
Dipeptide_label	Annotations of diagnostic dipeptides
Tripeptide_label	Annotations of diagnostic tripeptides
IM_NH3_label	Annotation of IM(Cit)-NH$_3$
Total_NL_count	Number of all diagnostic neutral loss ions
precNL_count	Number of precursor neutral losses
seqNL_count	Number of sequence ion neutral losses
intNL_count	Number of internal ion neutral losses
Total_INT_count	Number of all diagnostic internal ions
Dipeptide_count	Number of diagnostic dipeptides
Tripeptide_count	Number of diagnostic tripeptides
IM_NH3_count	Number of IM(Cit)-NH$_3$
Cit_probability	Probability (P) of citrullination status calculated by the EN model (HCD data only)
Cit_prediction	Classification of citrullination status using a P cutoff >0.5 (HCD data only)