User Manual for "SteroidXtract.py"

(Version 1.0, Aug 9, 2021)

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 script performs steroid MS^2 spectra extraction from untargeted metabolomics datasets.
- The program is written in Python and is publicly available at https://github.com/HuanLab/SteroidXtract.
- Please see below for the code instructions.

1) Install the Python language and following packages: pyteomics, numpy, pandas, os, math, tensorflow.
2) Download and open the Python script "SteroidXtract.py" in a Python IDE (Integrated Development Environment) such as PyCharm or Visual Studio Code.
3) Change the working directories in the Python script. The model directory should contain files "SteroidXtract_model.json" and "SteroidXtract_model.h5". Use "/" instead of "\" in the directory.

```
# Working directories
input_dir = 'E:/SteroidXtract_2020/mzXML files' # input data path for SteroidXtract
output_dir = 'E:/SteroidXtract_2020/output' # output data path for SteroidXtract
model_dir = 'E:/SteroidXtract_2020/model' # data path for SteroidXtract models
```

4) Set the parameters in the following table.

```
# Parameter setting
ms1_tol = 0.005  # MS1 mass tolerance
ms2_tol = 0.01  # MS2 mass tolerance
rt_threshold = 23  # valid retention time in minute, MS2 after rt_threshold are discarded
pre_int_threshold = 1000  # precursor intensity threshold
                          # (MS2 with precursor intensities lower than the threshold are discarded)
```

Table. Parameter settings.

Parameter name	Function
ms1_tol	Mass tolerance for MS^1 scans, default 0.005 *m/z*.
ms2_tol	Mass tolerance for MS^1 scans, default 0.01 *m/z*.
rt_threshold	Retention time threshold (in minute). MS^2 generated after the rt_threshold will be discarded.
pre_int_threshold	Precursor intensity threshold. MS2 with precursor intensities lower than pre_int_threshold are discarded.

5) Run the Python script. For each sample, a CSV file containing the prediction results of all the MS^2 spectra and a MGF file containing the steroid MS^2 spectra will be generated.

Note:

For the file conversion process, please refer to the following website.

https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/

Please also use 64-bit binary encoding precision (MSConvert).