TIDD: Tool-Independent and Data-Dependent PSM rescoring method

Outline:

Introduction

A Tool-Independent and Data-Dependent PSM rescoring method.

TIDD is a universal post-processing tool which supports confident peptideidentifications regardless of the search engine adopted. TIDD can work for any(including newly developed search engines because it calculates universalfeatures that assess peptide-spectrum match quality while it allows additional featuresprovided by search engines (or users) as well.

Here, we support two types of TIDD version -- a simple GUI based TIDD and command version of TIDD.
GUI
1. Data Management: extract features and check feature densities"),
2. Model fitting: fit SVM models"),

*Computational resources supported by free version of \"R studio cloud\" is limited, so it takes a little time for feature extraction."),

*Sample file description:
1.test_required.tsv: contain 6 required features for feature extraction
2.*_feature.stv: contain common features except calculated Xcorr
3.*_xcorrColAdded.tsv: add calculated Xcorr feature
4.*_ModelFitting_Results.tsv: add rescoring score feature

Required:
1. R (v4.1 or above) packages: shiny, shinythemes, shinydashboard,e1071,ROCR
2. java jre/jdk v1.8 or above

Download:
1.Download code from github: https://github.com/HanyangBISLab/TIDD.git

2.Download/Test code from R studio cloud:
https://rstudio.cloud/spaces/178915/project/2994889
3.Test GUI from shinyapps.io: https://honglan-li.shinyapps.io/project/

Install and run

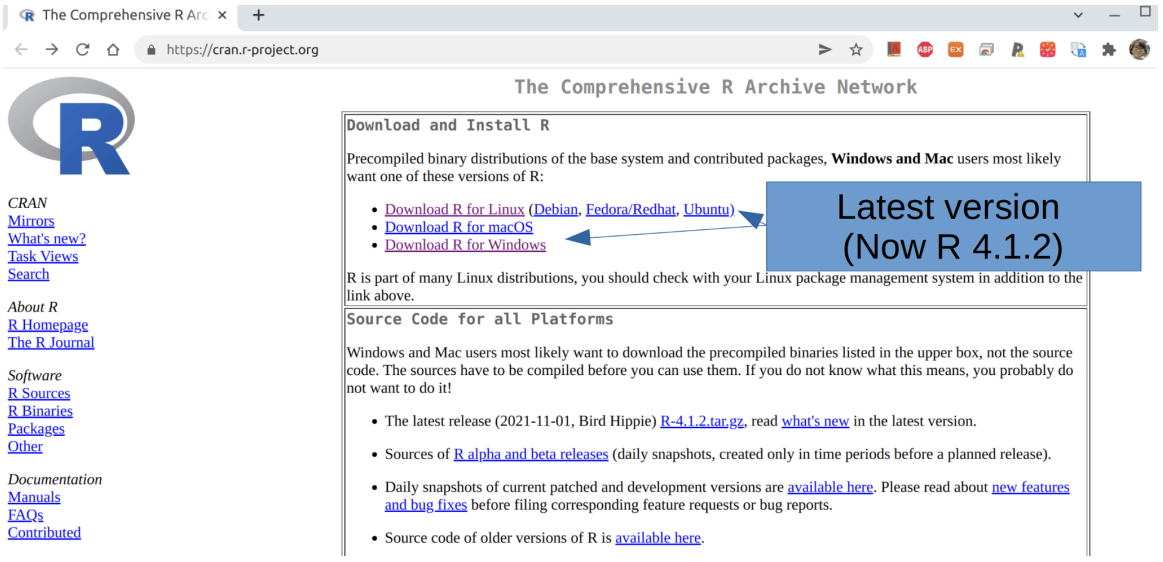
It is based on R and java programming.

So it required:

- Install java jdk (17 or above) and R(4.1.2 or above)
- For R, the following packages are needed to install.
  install.packages("shiny") install.packages("shinythemes")
  install.packages("shinyFiles") install.packages("shinydashboard")
  install.packages("ROCR") install.packages("e1071")
- Run shinyApp

Detail: R and Java based program

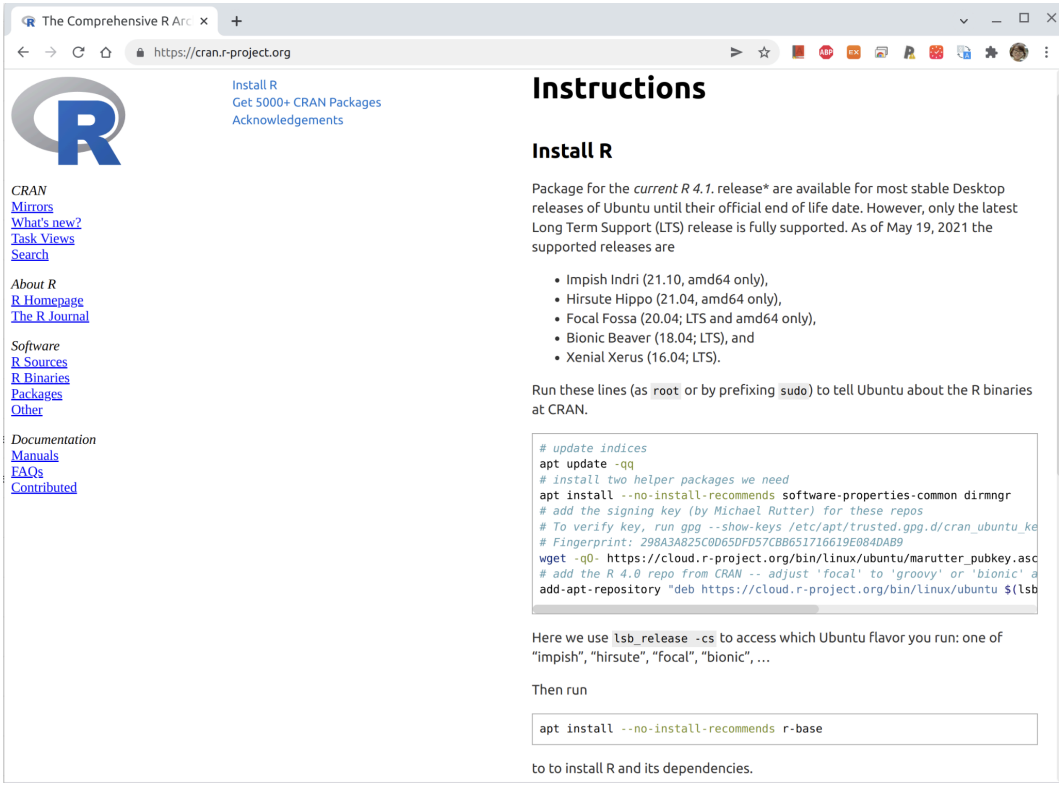
1. Install R version 4.1.2 (or above)
   link: https://cran.r-project.org/
   E.g. download R-4.1.2* ver



Downloaded file: window→ R-4.1.2-win.exe → “double click” to install R

ubuntu → click “ubuntu” → You can see the guideline for installing

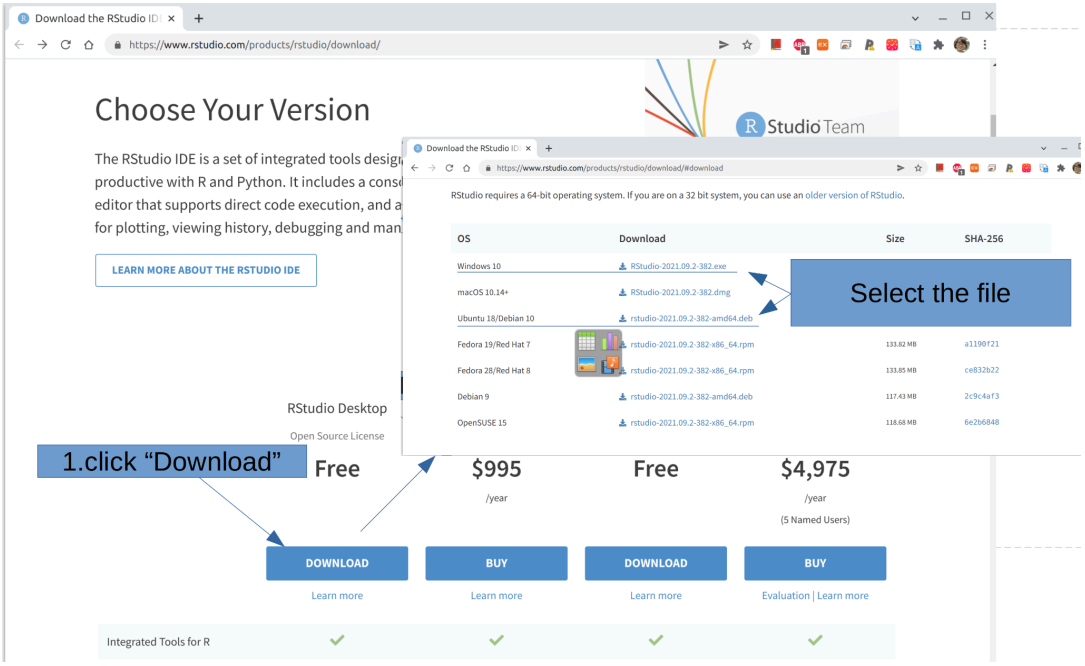
<captures below is from https://cran.r-project.org/>



2. Install R studio

link: https://www.rstudio.com/products/rstudio/download/

E.g. download Rstudio-2021.09.1*



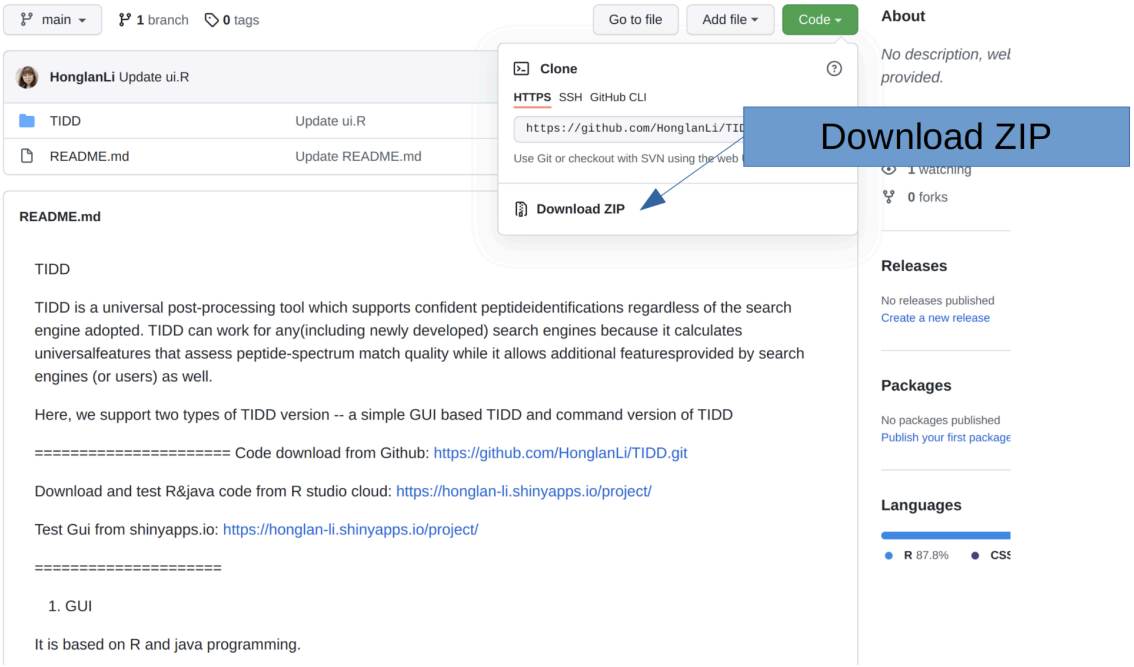
Install: window → double click “RStudio 202109*.exe”

ubuntu → follow the guidline provided by the link below:
https://linuxconfig.org/how-to-install-rstudio-on-ubuntu-20-04-focal-fossa-linux

3. Install java 8
   download: https://www.oracle.com/java/technologies/downloads/

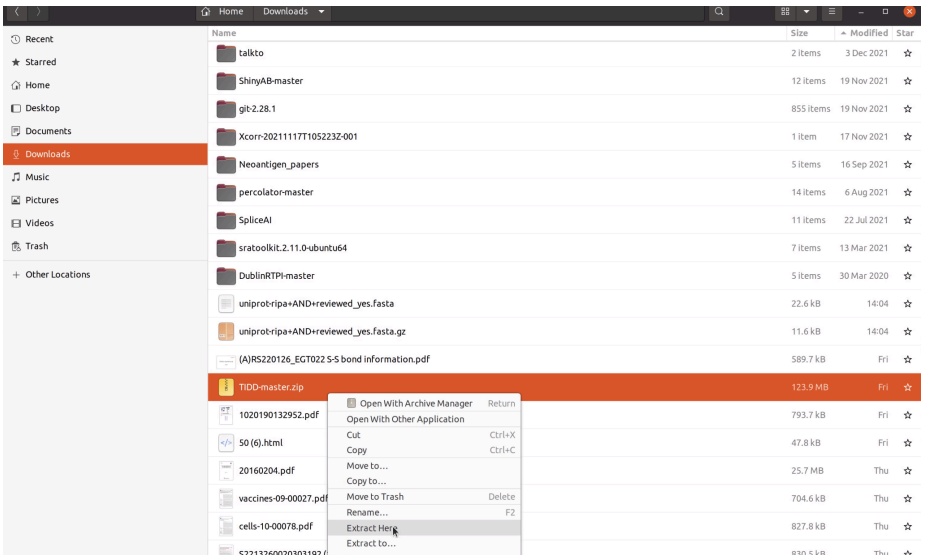
   Installation guide:
   https://docs.oracle.com/javase/10/install/installation-jdk-and-jre-microsoft-windows-platforms.htm#JSJIG-GUID-29333CFD-E7A6-498B-9317-97700C81D928
   E.g. jdk-17.01*

4. Download TIDD code from git: https://github.com/HonglanLi/TIDD.git



5. Unzip/Uncompress the TIDD*.zip

   Default location: ~/Download

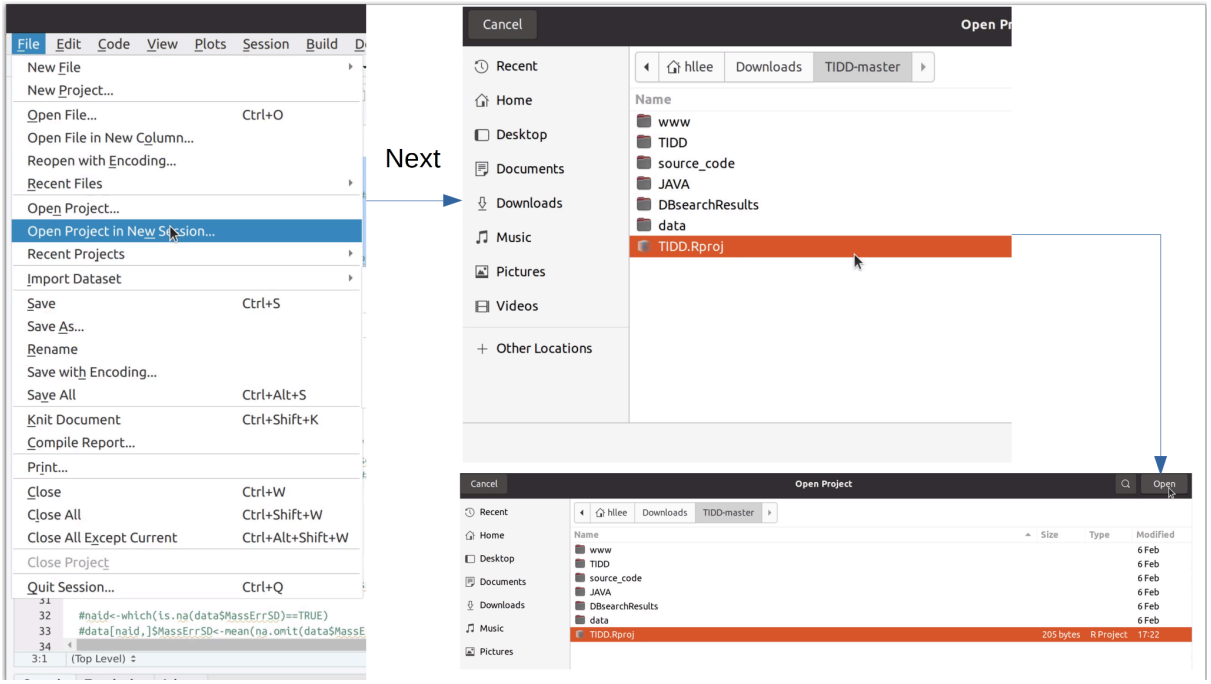


6. Open the R studio

   Import TIDD project to R studio

   File --> Open project in New Session --> select TIDD.rproj

   for example: "~/Download/TIDD-master/TIDD/TIDD.rproj

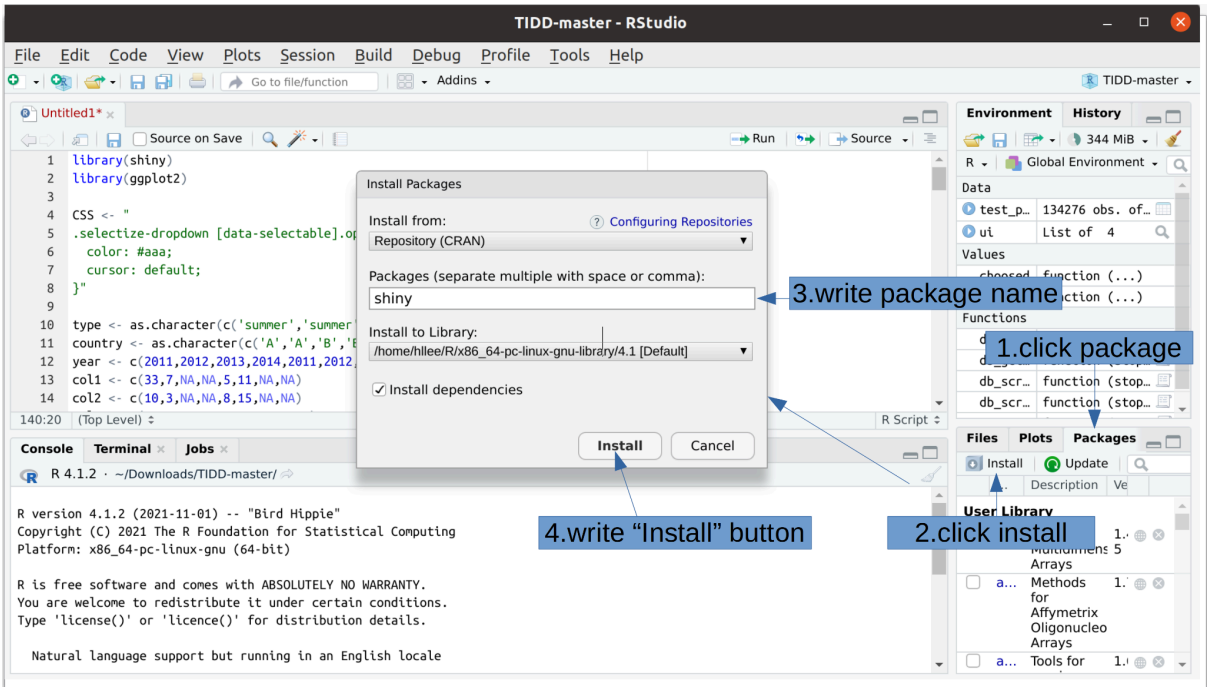


7. Install R packages

   required packages: shiny, shinythemes, shinydashboard, shinyFiles, e1071, ROCR

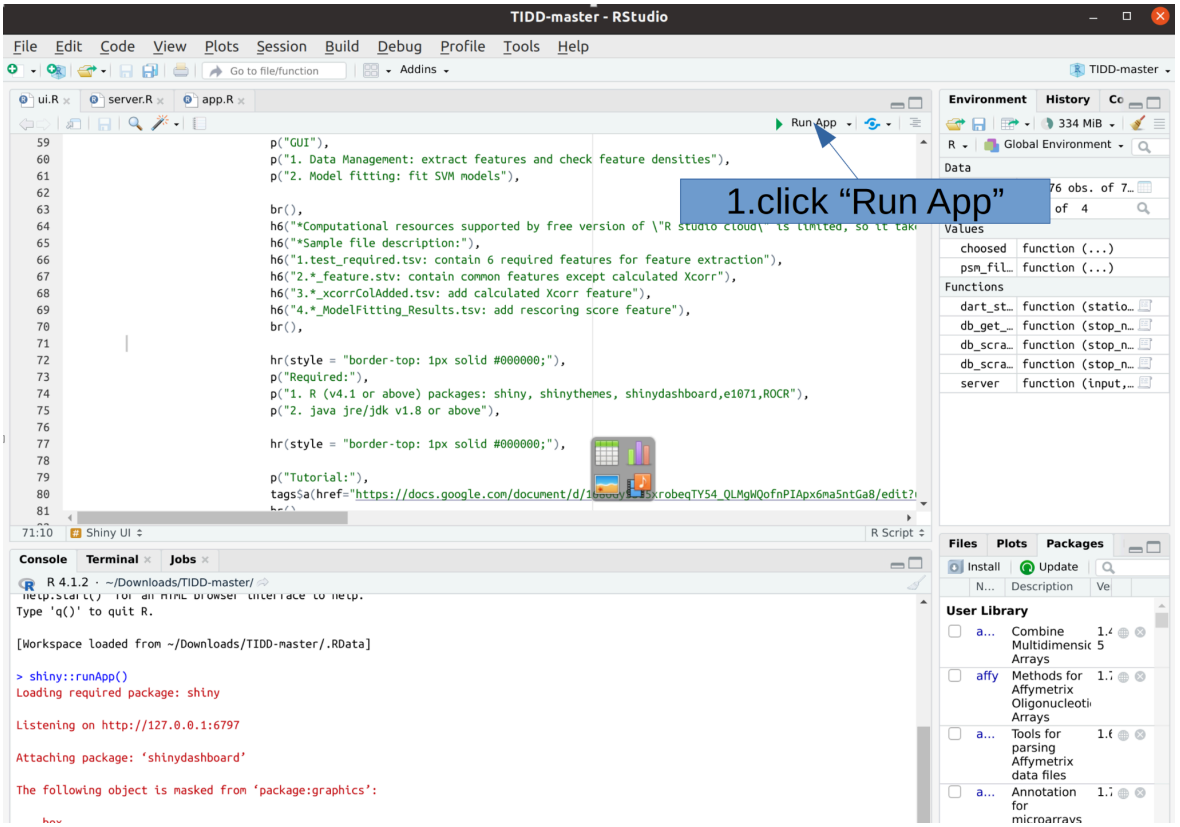
   E.g. install shiny package

   UI: click Packages --> install --> type "shiny" --> install



console: install.packages("shiny")

8. Run



File format

- TIDD input file format: PSM search result [*tsv]

This file requires 6 features:

1. Filename
2. Scan
3. Charge
4. precursor mass
5. Peptide
6. Protein

To calculate NTT value, pre and next amino acids must be included in the protein information or peptide information as below:

Example1:

Filename	Scan	Charge	PrecursorM	Peptide	Protein
test.mgf	19608	3	1449.831259	IETLDPALIRPGR	sp|P62191|PRS4_HUMAN(pre=R,next=I);sp|P62191-2|PRS4_HUMAN
test.mgf	37707	2	1685.96616	LPDGSEIPLPPILLGR	sp|Q96KP4|CNDP2_HUMAN(pre=K,next=L)
test.mgf	32201	2	2001.002903	IPEISIQDMTAQVTSPSGK	sp|P21333|FLNA_HUMAN(pre=K,next=T);sp|P21333-2|FLNA_
test.mgf	11229	2	965.506199	EVDIYTVK	sp|Q9NR22|ANM8_HUMAN(pre=K,next=T);sp|Q9NR22-2|ANM8_HUMAN;sp|Q9
test.mgf	36713	2	2752.472752	VPAATSAIITNDGIGINPAQTAGNVFLK	sp|P61960|UFM1_HUMAN(pre=K,next=H);sp|P619
test.mgf	26973	2	1268.721165	SGGLQPVPIVFR	sp|P11177|ODPB_HUMAN(pre=M,next=G);sp|P11177-2|ODPB_HUMAN
test.mgf	563	3	927.434836	AHGNSGMVR	sp|P18077|RL35A_HUMAN(pre=R,next=A)
test.mgf	9816	3	1293.700064	NVNQSLLELHK	sp|P02794|FRIH_HUMAN(pre=K,next=L)

Example2:

	SpecId	ScanNr	ExpMass	Peptide	Proteins	Charge			
2	20150911_QE7_UPLC11_DBJ_SA_HELA_3MG_50FRAC_TiO2_1.mgf					578	1166.504765	K.KM+15.9949NC+57.021464KS+79.9663FR.W	sp|Q8N428|-
	GLT16_HUMAN;sp|Q8N428-2|GLT16_HUMAN	3							
3	20150911_QE7_UPLC11_DBJ_SA_HELA_3MG_50FRAC_TiO2_1.mgf					579	1166.498119	R.C+57.021464C+57.021464TKPESER.M	CONTAMN_gi|-
	229552;CONTAMN_gi|113574	2							
4	20150911_QE7_UPLC11_DBJ_SA_HELA_3MG_50FRAC_TiO2_1.mgf					580	980.16234	K.T+79.9663DMT+79.9663EGS+79.9663.-	DECOY_sp|Q9Y230|-
	RUVB2_HUMAN;DECOY_sp|Q9Y230-2|RUVB2_HUMAN	3							
5	20150911_QE7_UPLC11_DBJ_SA_HELA_3MG_50FRAC_TiO2_1.mgf					581	765.288091	R.YGGSS+79.9663GR.P	tr|K7ESH3|K7ESH3_HUMAN 2

- TIDD input file format: MS/MS spectra [*mgf]

MS/MS spectra [*mgf] must contain title information.
The title format: filename+”.”+start scan+”.”+charge

Example:

```
BEGIN IONS
TITLE=test.2.2.2
RTINSECONDS=0.7896
PEPMASS=344.720764160156 64676.6015625
SCANS=2
CHARGE=2+
100.4279022 65.6441345215
101.0709763 204.4091339111
101.1071472 238.9379425049
102.0914917 183.1679229736
110.0713196 604.6396484375
111.0912781 213.9605712891
112.0756607 279.3070983887
113.1076431 147.8057403564
115.0876236 88.8695907593
120.0803528 154.9431304932
124.2364426 76.4476547241
```

Filename: test.mgf
Scan: 2
Charge: 2+

TITLE=test.2.2.2

- Test data: download link

Link: https://github.com/HanyangBISLab/TIDD/tree/master/data

PSM result file: ~/TIDD/data/PSM/

- Feature extraction input file: test_required.tsv
- Model fitting input file: test_required_xcorrColAdded.tsv
- Link: https://github.com/HanyangBISLab/TIDD/tree/master/data/PSM

MGF file: ~/TIDD/data/test/test.mgf

- Link: https://github.com/HanyangBISLab/TIDD/tree/master/data/MGF/test

Run: GUI

TIDD-GUI has 4 menus:

1. “About”: describes a brief introduction of TIDD, requirements for running TIDD, and a download link.
2. “File upload & download”: upload/download files to/from TIDD-GUI.
3. “Data management”: extract PSM features
4. “Model fitting”: fit model to rescore PSMs

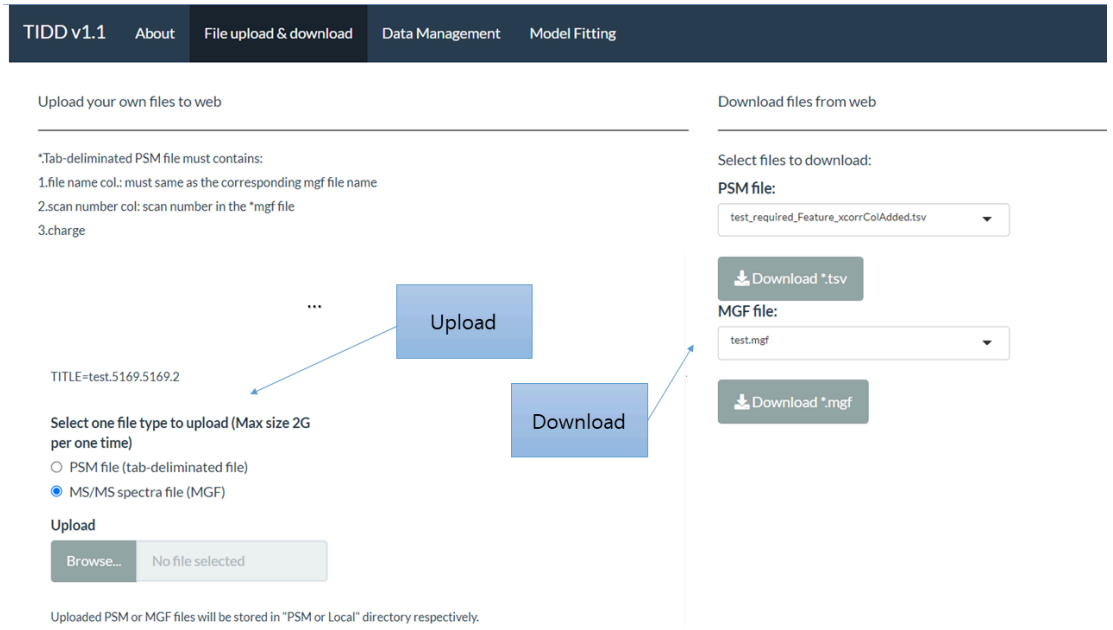
- File upload & download

Upload files:

1. Select file types <PSM file or MS/MS spectra file> for uploading
2. Click “Browse…” and select files for uploading
3. Done
    a. PSM file: located unnder the “data/PSM/” folder
    b. MGF file: located under the “data/MGF/Local/”

Download files:

1. Select one file for downloading
    a. PSM file: the files under the ‘data/PSM/’ folder
    b. MGF file: the files under the ‘data/PSM/test/’ folder
2. Click download button



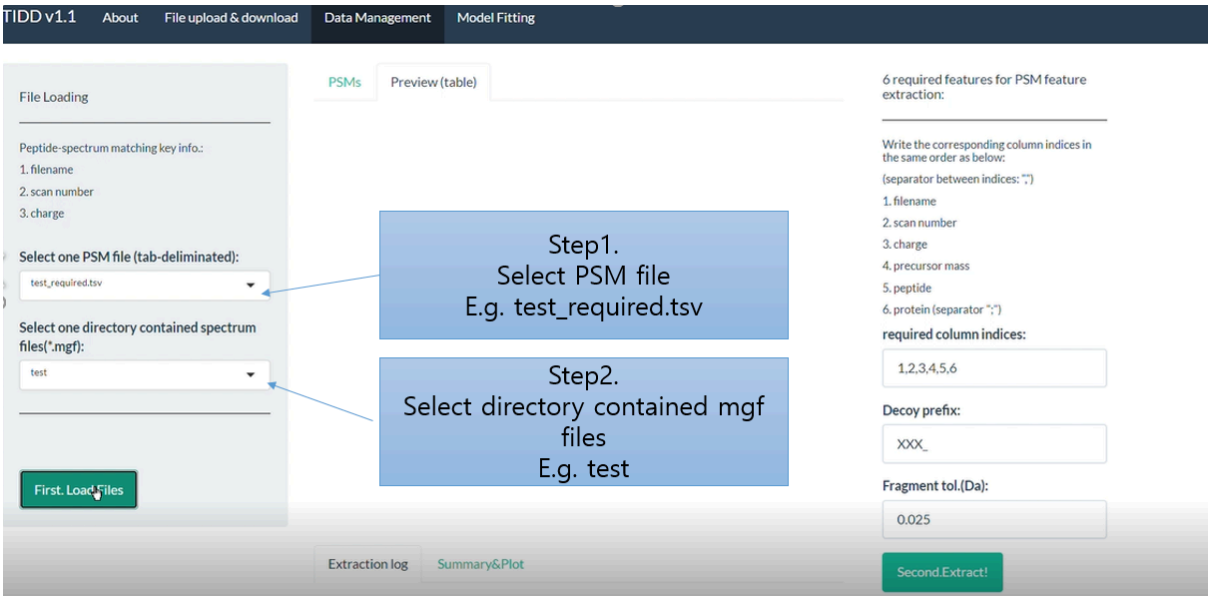
- Data management

Here, a set of PSM features can be extracted.

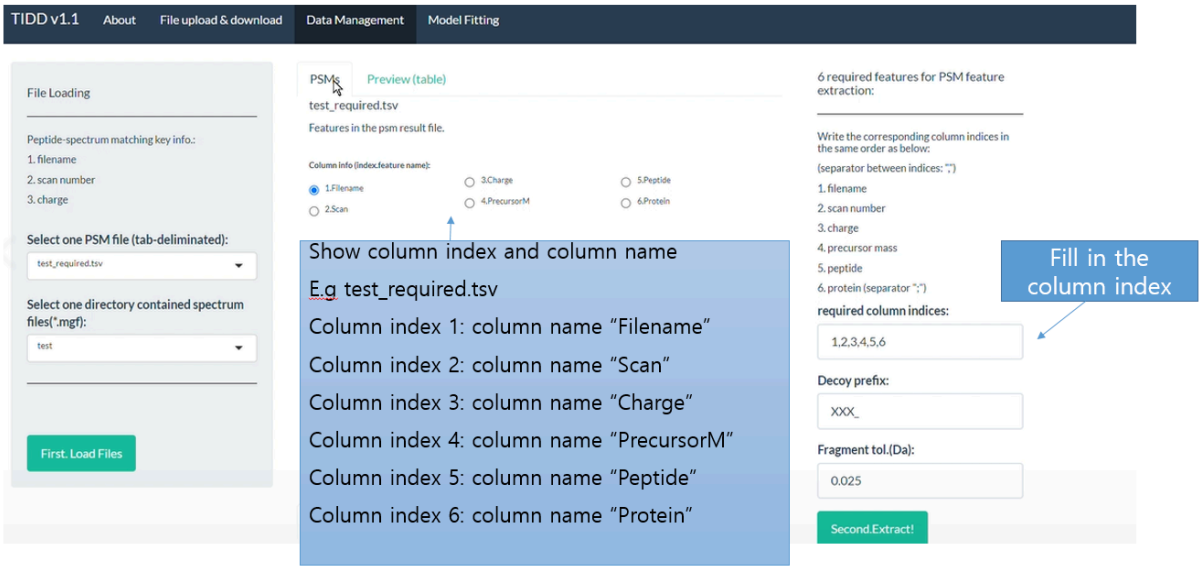
Video tutorial: [link]

Feature extraction:

1. Select files and click the

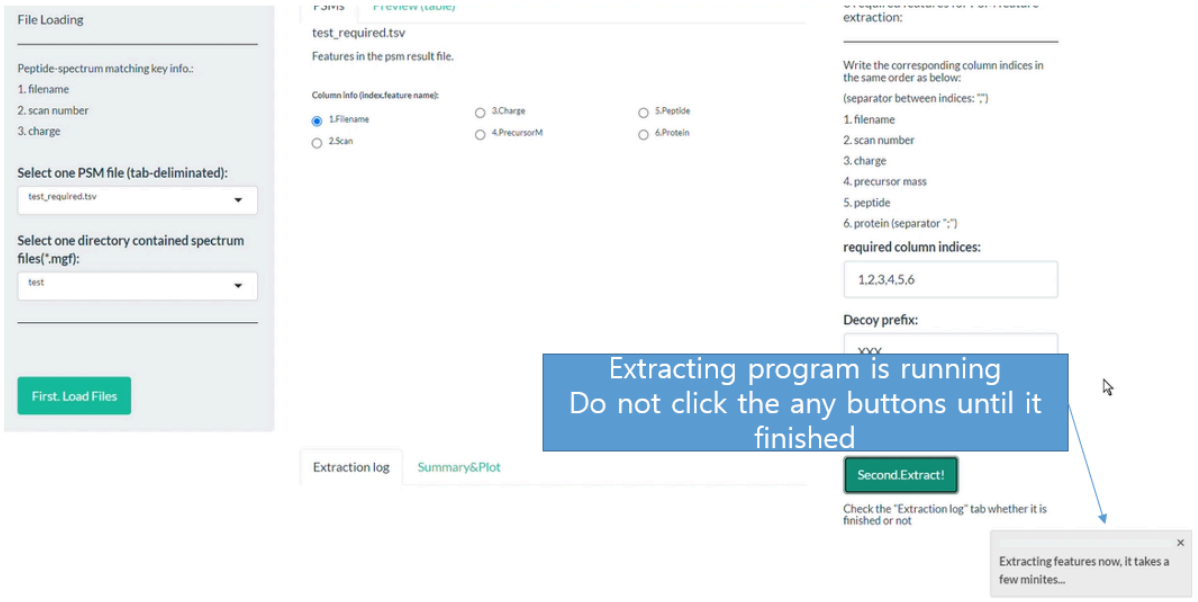


2. Fill in the 6 required information column indexes

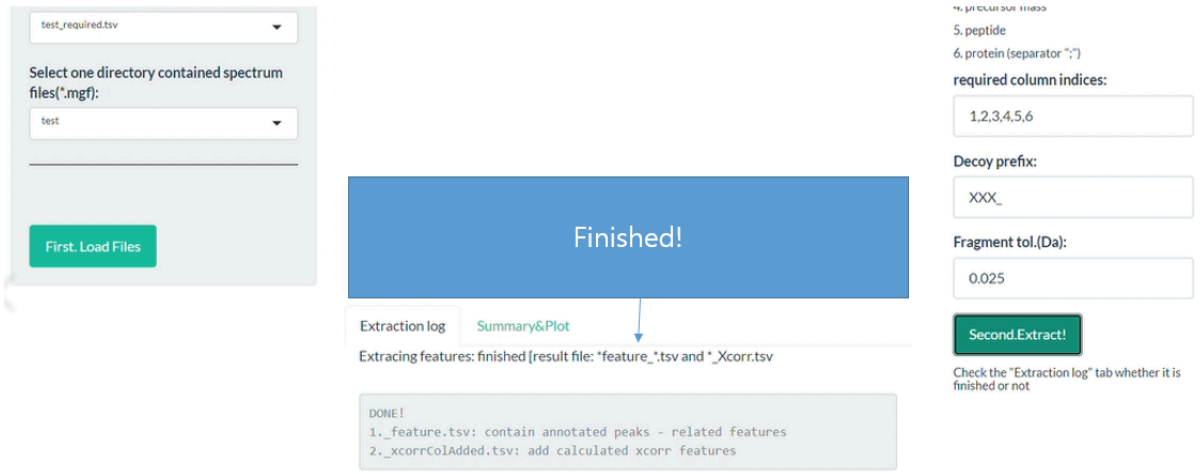


3. Click "Second Extract" button

   -Show the state of running



4. Done

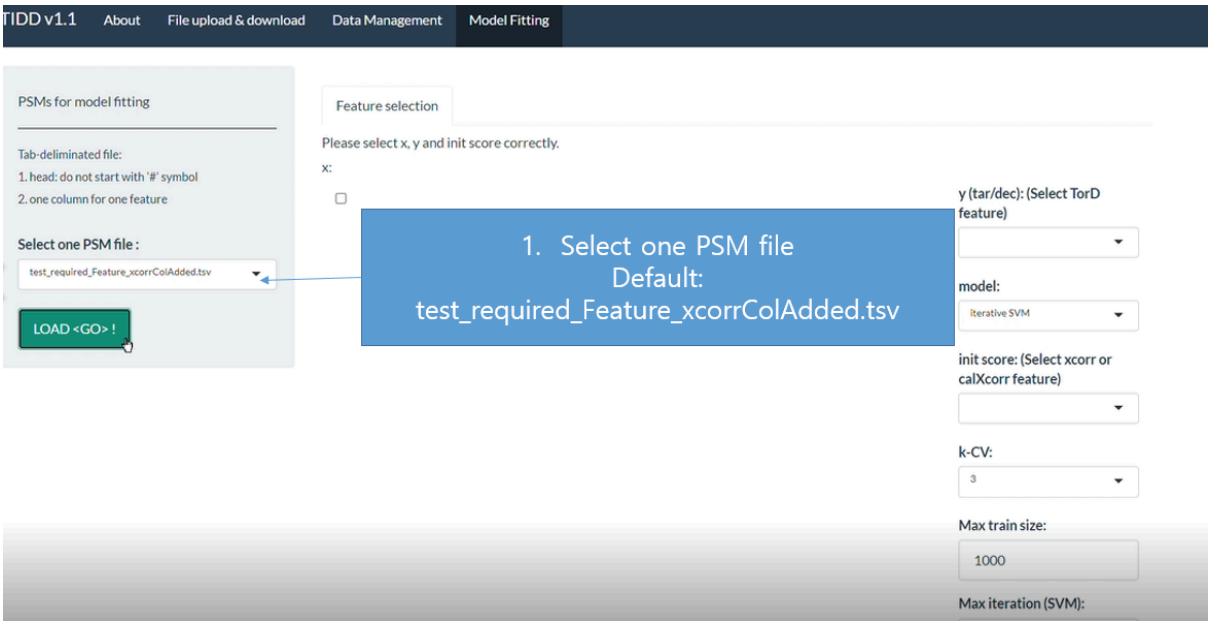


- Model Fitting

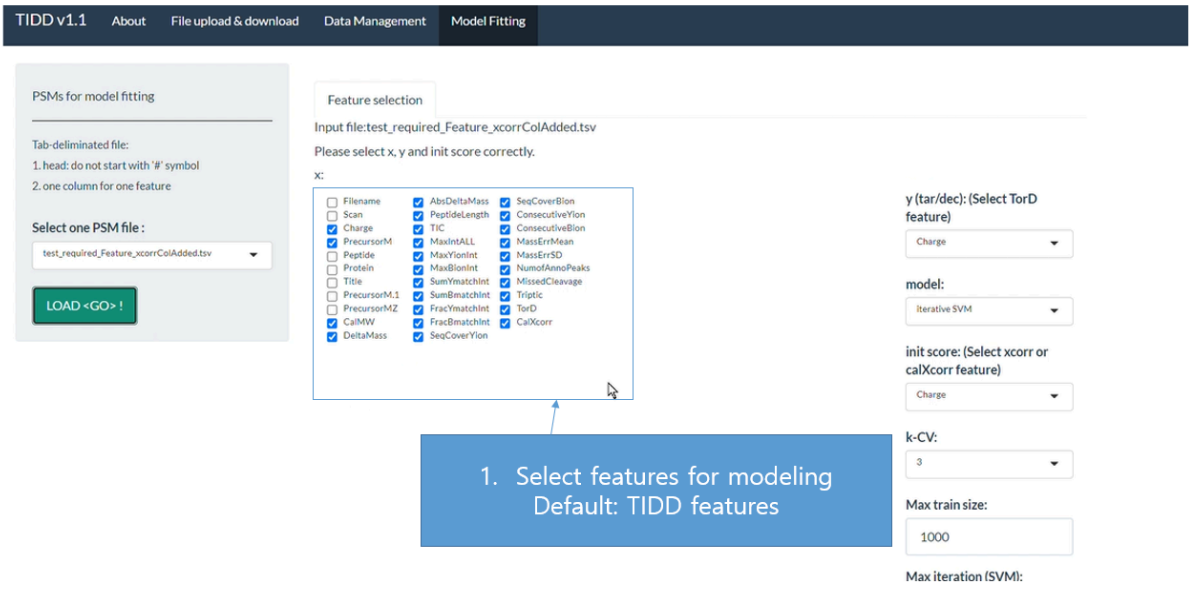
Video tutorial: [link]

Steps:

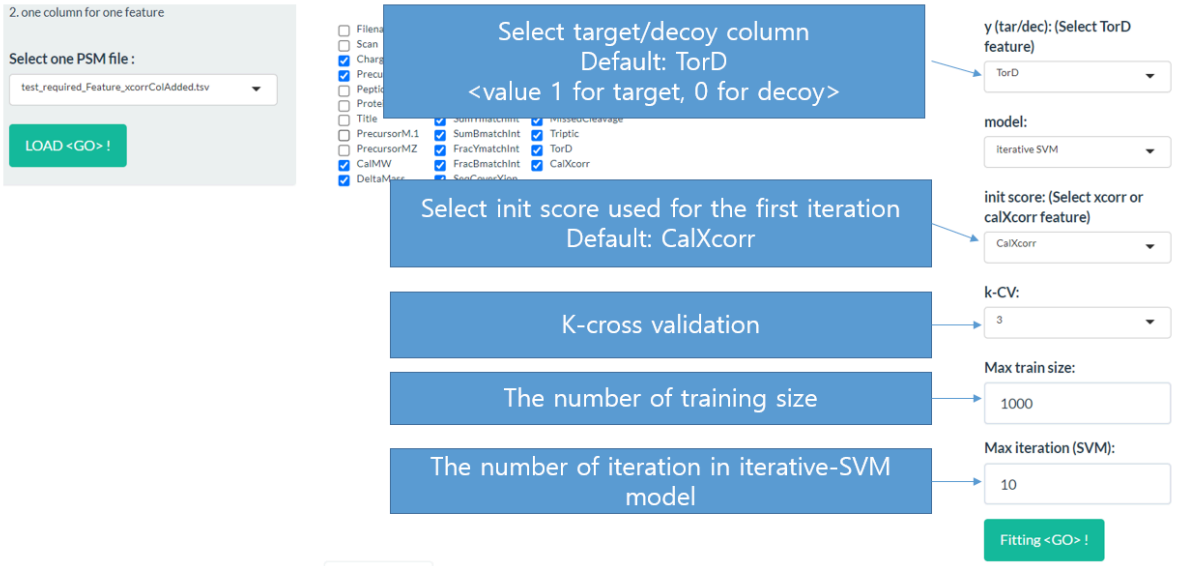
1. Select one PSM file → “LOAD <GO!>”



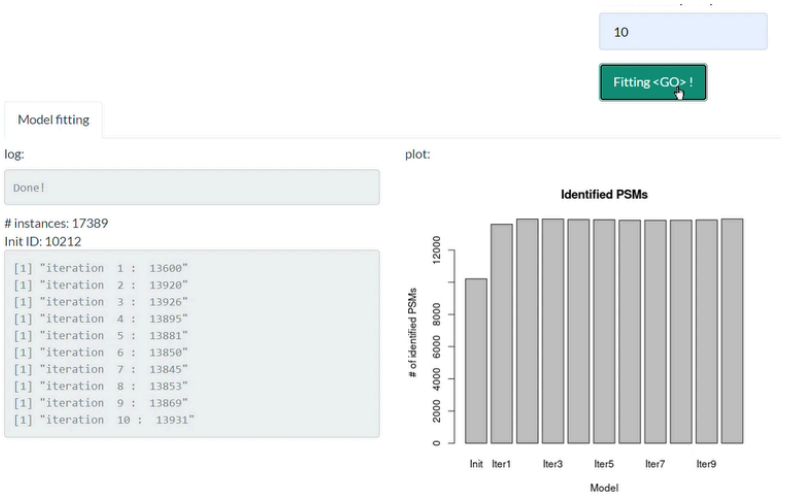
2. Select features for modeling



3. Select parameters and click “Fitting <GO!>” button



4. Show the identification results per iteration



Run: command line

It contains 3 steps:

1. Extract PSM features
2. Calcuated XCorr scores for PSMs
3. Fitting models and re-ranking PSMs

- PSM feature extraction

Two types of codes are uploaded in the github (https://github.com/HanyangBISLab/TIDD.git) The first one is a runable “TIDD.jar” file, which located in the “TIDD/JAVA”, and the other is “FeatureExtraction” source codes located in the “TIDD/source_code”

Run TIDD.jar
-Help: “java -jar TIDD.jar -h”
-Run: java -jar <options> <attributes>
-Options:

-i <filename> : PSM search result file [Required]
: tab-deliminated tsv file
-dir <filename> : MS/MS data file used for search [Required]
: Available formats [mgf]
-o <output_filename>: Output file path for extracted features [Optional]
: If not specified, it's automatically generated

-id <list of index with seperator "," >: list of column index in the tab-deleminated PSM result file [required]
: SpectrumFile column index,ScanNum column index, charge column index, precursorMZ column index,peptide column index,Protein column index

-dec <String>: decoy prefix [required]

-ft <double>: fragment tolerance (Unit:Da) [required]

-phospho <boolean>: true for phospho data [optional]
: default values is false

- Example: java -jar TIDD.jar -i psm_test.tsv -d C:/Test/ -id 1,2,3,4,5,11 -dec XXX_ -ft 0.025

- Calculate XCorr score

Two types of codes are uploaded in the github (https://github.com/HanyangBISLab/TIDD.git) The first one is a runable “Xcorr_v2.jar” file, which located in the “TIDD/JAVA”, and the other is “Xcorr” source codes located in the “TIDD/source_code”

Run Xcorr_v2.jar

-Help: “java -jar Xcorr_v2.jar -h”

-Run: “java -jar xcorr.jar spectrumDir resultFile binWidth binOffset titleCol peptideCol precursorMZCol chargeCol mgffilenameCol header”

- “spectrumDir” should be a mgf directory
- “resultFile” should be a tsv file
- column number is a zero-based number
- binWidth is default is 1.0005079 for low resolution. 0.02 is recommended for high resolution
- binOffset defulat is 0.4 for low resolution. 0.0 is fine for high resolution

-header value can be either true (result file has a header) or false (no header in the result file)

- Example: java -jar Xcorr_v2.jar /Test/ test.tsv 0.02 0.0 5,6,2,3,1 1

- Fitting models

For model fitting, we uploaded one “TIDD_ModelFit.R” file under the “TIDD/source_code” (https://github.com/HanyangBISLab/TIDD.git)

Run TIDD_ModelFit.R :

- Change the parameters inside the R code first.

<Change line 32 to 40>

TIDD_ModelFit.R

```
32  inputfile<-"/home/hllee/Documents/TIDD/modif/phosphp/AML/AML_phospho_comet/AML_phospho_format_Feature.tsv"
33  outputfile<-"/home/hllee/Documents/TIDD/modif/phosphp/AML/AML_phospho_comet/AML_phospho_format_Feature_MF_impute.tsv"
34  MaxNum_iteration<-10
35  minPLEN<-6
36  trainsize<-10000
37  k_cross<-5
38  out_ycol<-"TorD"
39  init_score<-"Xcorr"
40  features=c("TorD","Xcorr","AbsDeltaMass","Charge","PeptideLength","Triptic","MissedCleavage","PrecursorM","DeltaMass","CalMW","TIC","SumYmatchInt"
41  #features=c("TorD","CalXcorr","NumofAnnoPeaks","ConsecutiveYion","SeqCoverYion")
```

- Run “Rscript TIDD_ModelFit.R”