

metID: A R package for Automatable Compound Annotation for LC–MS-based Data

Xiaotao Shen PhD

Stanford University School of Medicine
Department of Genetics
Dr. Snyder lab

November 3th, 2021 @ Philadelphia ASMS 2021



shenxt@stanford.edu



shenxt.info



@xiaotaoshen1990

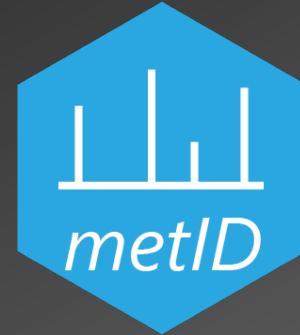


github.com/jaspershen

</> **metID: Compound Annotation for LC-MS Data**

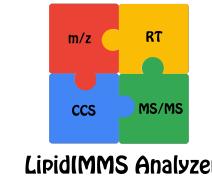
- In-house Database Construction
- Databases Provided in metID
- Database Sharing
- Case Study

</> **TidyMass: A Computational Framework for LC-MS Data Processing and Analysis**



metID: Compound Annotation for LC-MS Data

► Background



Metabolite identification and Dysregulated Network Analysis



METCCS

CHOLINE ADENOSINE TRIPHOSPHATE CHOLESTEROL **TESTOSTERONE** GLUTAMIC ACID UREA GALACTOSE CHOLINE ACYL CARNITINE THYMINE
TRYPTOPHAN PHOSPHOCHOLINE CYANINE CHOLINE AMINO ACID CHOLINE CHOLINE GLYCEROL
PYRUVIC ACID **TESTOSTERONE GLUCOSE** **GLUCOSE** **CHOLESTEROL** **Nicotinamide Adenine Dinucleotide** **Oxaloacetic Acid** **Oxaloacetate** **Acetyl Coenzyme A**
GLUCOSE **CHOLESTEROL** **OXALOSUCINIC ACID** **ACETYL FUMARATE** **CHOLINE** **CHOLINE GLYCEROL**
Nicotinamide Adenine Dinucleotide **Oxaloacetic Acid** **Oxaloacetate** **Acetyl Coenzyme A** **CHOLINE** **CHOLINE GLYCEROL**
SERINE TRYPTOPHAN PHOSPHOCHOLINE **ACYL CARNITINE** **THYMINE** **GLYCEROL** **CHOLINE** **CHOLINE GLYCEROL**
PYRUVIC ACID **UREA** **GALACTOSE** **PHOSPHOCHOLINE** **ACYL CARNITINE** **CHOLINE** **CHOLINE GLYCEROL**
TESTOSTERONE **GLUCOSE** **CHOLESTEROL** **OXALOSUCINIC ACID** **OXALOSUCINIC ACID** **CHOLINE** **CHOLINE GLYCEROL**
GLUCOSE **CHOLESTEROL** **ADENINE** **DINUCLEOTIDE** **CHOLINE** **CHOLINE GLYCEROL** **CHOLINE** **CHOLINE GLYCEROL**
Nicotinamide Adenine Dinucleotide **PhosphoCholine** **AcylCarnitine** **Thymine** **Glycerol**
SERINE TRYPTOPHAN

Background



1. User-friendly,
2. Simple.

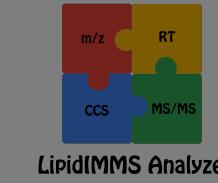


GNPS

METCCS



Metabolite identification and
Dysregulated Network Analysis



CHOLINE ADENOSINE TRIPHOSPHATE CHOLESTEROL TESTOSTERONE GLUCOSE
SERINE TRYPTOPHAN PHOSPHOCHOLINE CYLCARNITINE THYMINE GLUCOSE
PYRUVIC ACID UREA GALACTOSE CHOLINE PHOSPHATE CHOLINE GLYCEROL
TESTOSTERONE GLUCOSE CHOLINE CHOLINE CHOLINE CHOLINE CHOLINE
PYRUVIC ACID UREA CHOLINE CHOLINE CHOLINE CHOLINE CHOLINE
GLUCOSE CHOLESTEROL OXALOSUCINIC ACID GALACTOSE CHOLINE
NICOTINAMIDE ADENINE DINUCLEOTIDE OXALOSUCINIC ACID GALACTOSE CHOLINE
SERINE TRYPTOPHAN PHOSPHOCHOLINE CYLCARNITINE CHOLINE MALIC ACID
PYRUVIC ACID UREA GALACTOSE CHOLINE ADENOSINE CHOLINE MALIC ACID
TESTOSTERONE GLUCOSE CHOLESTEROL OXALOSUCINIC ACID GALACTOSE CHOLINE
GLUCOSE CHOLESTEROL OXALOSUCINIC ACID GALACTOSE CHOLINE GLYCEROL
NICOTINAMIDE ADENINE DINUCLEOTIDE OXALOSUCINIC ACID GALACTOSE CHOLINE
SERINE TRYPTOPHAN PHOSPHOCHOLINE CYLCARNITINE THYMINE GLUCOSE



Background



1. User-friendly,
2. Simple.

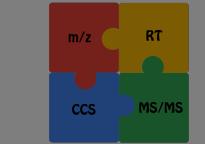


GNPS

METCCS



1. For GUI software, most of them only support Windows,
2. Not flexible.



LipidIMMS Analyzer



CHOLINE ADENOSINE TRIPHOSPHATE CHOLESTEROL TESTOSTERONE GLUCOSE
SERINE TRYPTOPHAN PHOSPHOCHOLINE CYCLICARNITINE THYMIDYLIC ACID
PYRUVIC ACID UREA GALACTOSE CHOLINE PHOSPHATE CHOLESTERYL GLYCEROL
TESTOSTERONE GLUCOSE CHOLINE LYSINE CHOLINE CHOLINE CHOLINE CHOLINE
GLUCOSE CHOLESTEROL OXALOSUCINIC ACID KETONURIC AMINO ACID
NICOTINAMIDE ADENINE DINUCLEOTIDE OXALOSUCINIC ACID GALACTOSYLCERAMIDE
SERINE TRYPTOPHAN PHOSPHOCHOLINE CYCLICARNITINE CHOLINE MALEIC ACID
PYRUVIC ACID UREA GALACTOSE CHOLINE PHOSPHATE CHOLESTEROOL GLYCEROL
TESTOSTERONE GLUCOSE CHOLESTEROL OXALOSUCINIC ACID GALACTOSYLCERAMIDE
GLUCOSE CHOLESTEROL ADENINE DINUCLEOTIDE OXALOSUCINIC ACID GLYCEROL
NICOTINAMIDE ADENINE DINUCLEOTIDE PHOSPHOCHOLINE CYCLICARNITINE THYMIDYLIC ACID
SERINE TRYPTOPHAN PHOSPHOCHOLINE CYCLICARNITINE THYMIDYLIC ACID

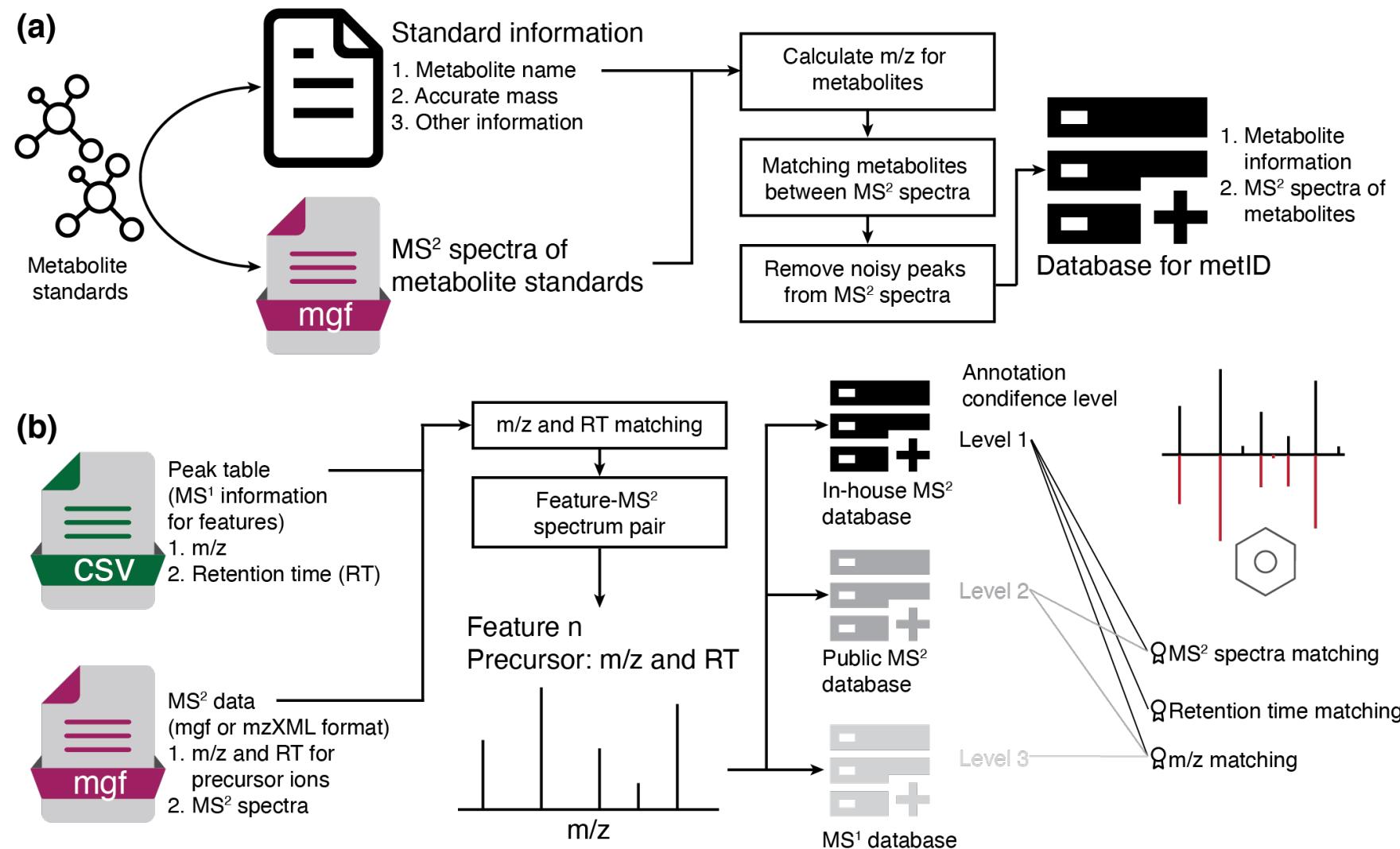


➤ metID: Design and Overview

1. R-based: can be installed any platform (Windows, Mac and Linux).
2. Universal format database.
3. Support multiple database format from other common tools (GNPS, MoNA and so on).
4. Automatable metabolite annotation functions.

metID: Design and Overview

1. R-based: can be installed any platform (Windows, Mac and Linux).
2. Universal format database.
3. Support multiple database format from other common tools (GNPS, MoNA and so on).
4. Automatable metabolite annotation functions.



➤ In-house Database Construction

A lot of labs have compound standard information (retention time and MS/MS spectra).



➤ In-house Database Construction

A lot of labs have compound standard information (retention time and MS/MS spectra).



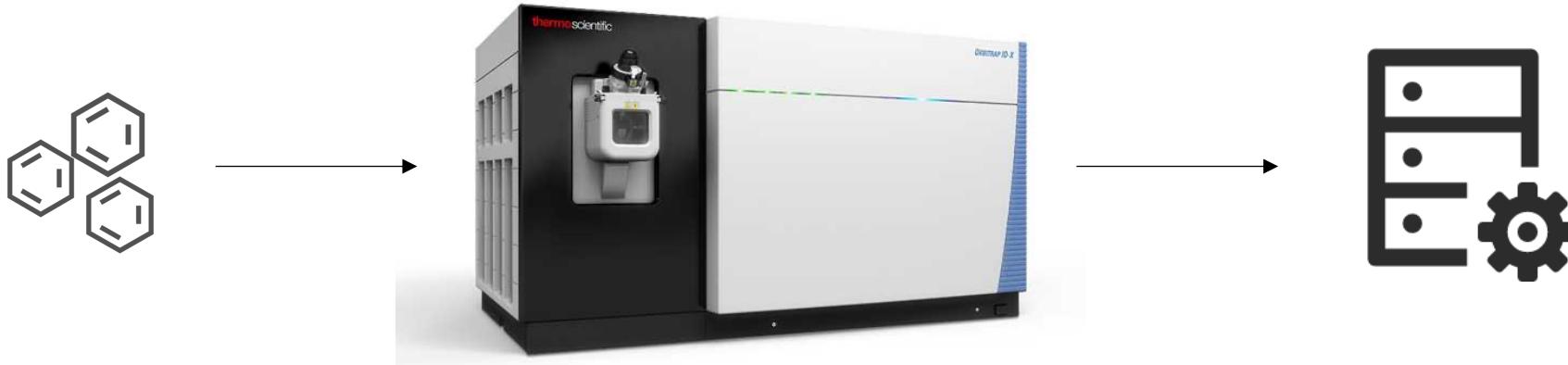
Standard information
1. Metabolite name
2. Accurate mass
3. Other information



MS² spectra of
metabolite standards

➤ In-house Database Construction

A lot of labs have compound standard information (retention time and MS/MS spectra).



Standard information
1. Metabolite name
2. Accurate mass
3. Other information



MS² spectra of
metabolite standards



Example Code

```
> test.database <-  
construct_database( path = new.path,  
version = "0.0.1",  
metabolite.info.name = "standard.information.csv",  
source = "Michael Snyder lab",  
link = "http://snyderlab.stanford.edu/",  
creater = "Xiaotao Shen",  
email = "shenxt1990@163.com",  
rt = TRUE,  
mz.tol = 15,  
rt.tol = 30,  
threads = 5 )
```

› Database format for metID

Compounds information

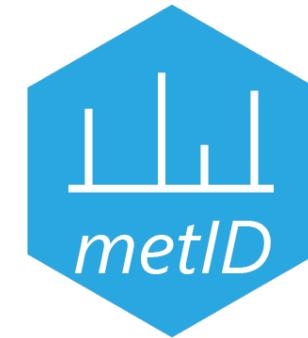
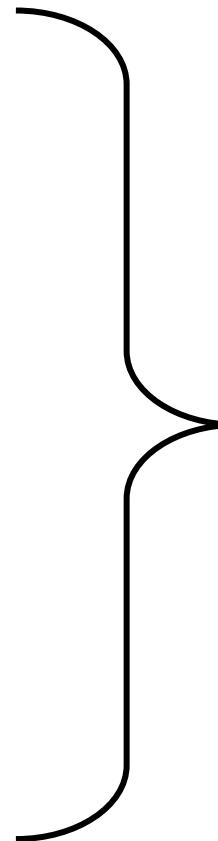
ID	Name	Formula	Accurate mass	RT
C1	xxx	xxx	xxx	xxx
C2	xxx	xxx	xxx	xxx
...

Spectrum information

ID	Polarity	MS2 spectrum
C1	xxx	xxx
C1	xxx	xxx
...

```
> test.database
-----Base information-----
# Version: 0.0.1
# Source: MS
# Link: http://snyderlab.stanford.edu/
# Creator: Xiaotao Shen ( shenxt1990@163.com )
# With RT information
# -----Spectral information-----
# There are 14 items of metabolites in database:
# Lab.ID; Compound.name; mz; RT; CAS.ID; HMDB.ID; KEGG.ID; Formula; mz.pos; mz.neg; Submitter; Family;
# Sub.pathway; Note
# There are 833 metabolites in total
# There are 356 metabolites in positive mode
# There are 534 metabolites in negative mode
# Collision energy in positive mode:
# NCE25; NCE50
# Collision energy in negative mode:
# NCE25; NCE50
```

› Public Databases



Database format for metID

› Public Databases



MassBank
High Quality Mass Spectral Database

MoNA

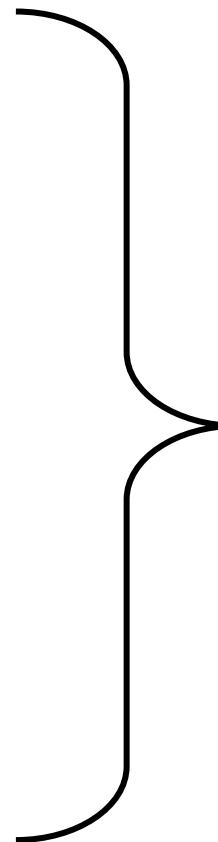
UCDAVIS
Fiehn Lab

DRUGBANK

GNPS



T3DB



</> Example Code

Read databases from other tools

```
> massbank_database <-  
> construct_mona_database(file = "MassBank.msp",  
  source = "MassBank")  
  
> mona_database <-  
  construct_mona_database(file = "MoNA.msp",  
  source = "MoNA")  
  
> gnps_database <-  
  construct_gnps_database(file = "GNPS.msp",  
  source = "GNPS")
```

Output databases for other tools

```
> write_msp_mona(database = database, path = ".")  
> write_msp_massbank(database = database, path = ".")  
> write_msp_gnps(database = database, path = ".")  
  
> write_mgf_mona(database = database, path = ".")  
> write_mgf_massbank(database = database, path = ".")  
> write_mgf_gnps(database = database, path = ".")
```

› Databases Provided in metID



Database	Compound number	Spectrum number	Information	Level	Source
msDatabase_rplc0.0.2	917	2,047	mz/RT/MS ²	1	Snyder lab
msDatabase_hilic0.0.2	846	2,570	mz/RT/MS ²	1	Snyder lab
hmdbDatabase0.0.2	5,646	22,331	mz/MS ²	2	https://hmdb.ca/downloads
massbankDatabase0.0.2	18,311	91,148	mz/MS ²	2	https://massbank.eu/MassBank
monaDatabase0.0.2	19,537	40,315	mz/MS ²	2	https://mona.fiehnlab.ucdavis.edu
orbitrapDatabase0.0.1 ³	8,360	23,227	mz/MS ²	2	https://mona.fiehnlab.ucdavis.edu
fiehn_hilic_database0.0.1	3,042	3,042	mz/MS ²	2	https://fiehnlab.ucdavis.edu/staff/kind/publications

› Databases Provided in metID



DRUGBANK



The Blood Exposome Database

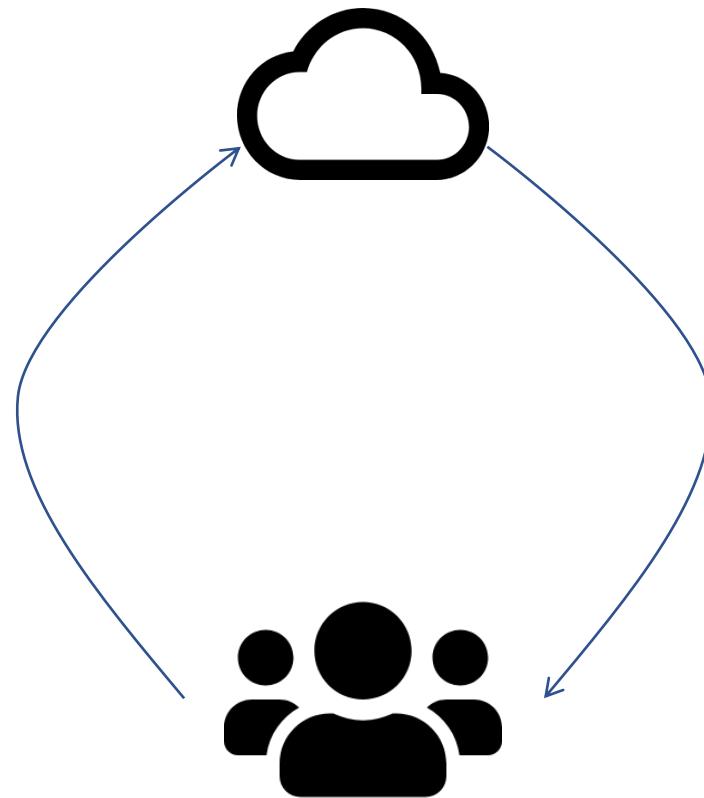
Database	Compound number	Spectrum number	Information	Level	Source
hmdbMS1Database0.0.1	114,004	0	mz	3	https://hmdb.ca/downloads
keggMS1Database_1.0	16,409	0	mz	3	https://www.genome.jp/kegg/compound/
drugbankMS1Database5.1.8	11,174	0	mz	3	https://go.drugbank.com/releases/latest
T3DBMS1Database_1.0	3,533	0	mz	3	http://www.t3db.ca/downloads
bloodExposomeMS1Database_1.0	65,957	0	mz	3	https://bloodexposome.org/#/download

› Databases Provided in metID



The Blood Exposome Database

Database	Compound number	Spectrum number	Information	Level	Source
hmdbMS1Database0.0.1	114,004	0	mz	3	https://hmdb.ca/downloads
keggMS1Database_1.0	16,409	0	mz	3	https://www.genome.jp/kegg/compound/
54,896 compound and 180,063 spectra in total (Only the databases with MS/MS spectra)					
drugbankMS1Database5.1.8	11,174	0	mz	3	https://go.drugbank.com/releases/latest
T3DBMS1Database_1.0	3,533	0	mz	3	http://www.t3db.ca/downloads
bloodExposomeMS1Database_1.0	65,957	0	mz	3	https://bloodexposome.org/#/download



Database provided for metID

Xiaotao Shen PhD (<https://www.shenxt.info/>)

Si Wu PhD

School of Medicine, Stanford University

Created on 2020-03-28 and updated on 2021-05-10

Source: vignettes/public_databases.Rmd

Contents

In-house MS₂MS² databases from Michael Snyder lab

Public MS₂MS² databases

Public MS₁MS¹ databases

In-house *MS²* databases from Michael Snyder lab

1. Michael Snyder HILIC databases

- ⓘ Professor Michael Snyder's lab. More than 1,000 metabolites.
- ⬇ Download here

If you need this database, please contact shenxt@stanford.edu

2. Michael Snyder HILIC databases

- ⓘ Professor Michael Snyder's lab. More than 1,000 metabolites.
- ⬇ Download here

› Automatable Metabolite Annotation



Example Code

```
> param1 <-  
  identify_metabolites_params(ms1.match.ppm = 15, rt.match.tol = 15, polarity = "positive", ce = "all", column = "rp",  
    total.score.tol = 0.5, candidate.num = 3, threads = 3, database = "msDatabase_rplc0.0.2" )  
> param2 <-  
  identify_metabolites_params(ms1.match.ppm = 15, rt.match.tol = 15, polarity = "positive", ce = "all", column = "rp",  
    total.score.tol = 0.5, candidate.num = 3, threads = 3, database = "MoNA")  
  
> result <-  
  identify_metabolite_all(ms1.data = "ms1.peak.table.csv",  
    ms2.data = "QC1_MSMS_NCE25.mgf",  
    parameter.list = c(param1, param2))
```

```
result[[1]]  
#> -----metID version-----  
#> 0.4.1  
#> -----Identifications-----  
#> (Use get_identification_table() to get identification table)  
#> There are 100 peaks  
#> 23 peaks have MS2 spectra  
#> There are 14 metabolites are identified  
#> There are 10 peaks with identification  
#> -----Parameters-----  
#> (Use get_parameters() to get all the parameters of this processing)  
#> Polarity: positive  
#> Collision energy: all  
#> database: msDatabase_rplc0.0.2  
#> Total score cutoff: 0.5 #> Column: rp  
#> Adduct table: #> (M+H)+;(M+H-H2O)+;(M+H-2H2O)+
```

› Automatable Metabolite Annotation

annotation_table

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Save As...

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	name	mz	rt	Compound.n	CAS.ID	HMDB.ID	KEGG.ID	Lab.ID	Adduct	mz.error	mz.match.sc	RT.error	RT.match.sc	CE	SS	Total.score	Database	Level
2	pRPLC_603	162.1125	33.746	L-Carnitine	541-15-1	HMDB00062	C00318	RPLC_406	(M+H)+	1.66789418	0.99383713	1.9743312	0.99137522	NCE25	0.60482879	0.79871748	msDatabase	1
3	pRPLC_1046	181.0720	201.182	Theophylline	511-59-6	HMDB01889	C07130	RPLC_732	(M+H)+	4.23918273	0.98665656	4.800648	0.9500755	NCE25	0.75232761	0.8603991	msDatabase	1
4	pRPLC_1112	209.0922	57.406	L-KYNURENI NA	NA	NA	NA	RPLC_252	(M+H)+	0.16269262	0.99994118	0.4264392	0.99959597	NCE25	0.64690869	0.82333865	msDatabase	1
5	pRPLC_2151	363.2198	638.691	CORTISOL	NA	NA	NA	RPLC_260	(M+H)+	7.8996867	0.87050838	5.33868	0.93862741	NCE25	0.78028335	0.8424256	msDatabase	1
6	pRPLC_3110	414.3002	545.186	Chenodeoxy C60-79-9	HMDB00637	C05466	HMDB00637	RPLC_871	(M+H+2H2O)	0.10242369	0.99976768	4.981698	0.94634357	NCE25	0.67228676	0.82272348	msDatabase	1
7	pRPLC_3959	153.066	84.24	1-N-Methyl-2 NA	HMDB04193	C05842	HMDB04193	RPLC_444	(M+H)+	1.68371582	0.99372003	4.87518841	0.94855398	NCE50	0.75418866	0.86266285	msDatabase	1
8	pRPLC_5025	195.0878	284.564	Caffein	58-08-2	HMDB01847	C07481	RPLC_439	(M+H)+	1.87017374	0.99225779	1.11113615	0.99726015	NCE25	0.72875971	0.86175935	msDatabase	1
9	pRPLC_1959	310.2011	322.275	C10:3 AC(1)	NA	NA	NA	RPLC_387	(M+H)+	0.74272449	0.99877489	10.7510749	0.77347888	NCE25	0.76187956	0.82400325	msDatabase	1
10	pRPLC_2147	166.0726	77.507	7-Methylguia 78-7	HMDB00897	C02242	HMDB00897	RPLC_435	(M+H)+	1.63462489	0.99407982	3.57981841	0.97192368	NCE25	0.69599442	0.83949808	msDatabase	1
11	pRPLC_22091	182.0813	33.42	DL-o-Tyrosin NA	NA	NA	NA	RPLC_146	(M+H)+	0.12605817	0.99966469	1.2243402	0.99667441	NCE25	0.56580409	0.78206183	msDatabase	1
12	pRPLC_3927	232.1545	77.507	Isobutyryl-L- 25518-49-4	NA	NA	NA	NO01896	(M+H)+	0.61329074	0.99916451	NA	NA	20	0.50460357	0.69006394	orbitrapData	2
13	pRPLC_4508	260.1857	222.592	Hexanoyl-Lc 22671-29-0	NA	NA	NA	NO02458	(M+H)+	0.17378276	0.99993289	NA	NA	15	0.72122198	0.82573859	orbitrapData	2
14	pRPLC_20354	352.1653	256.48	Phe-Trp	24587-41-5	NA	NA	NO04911	(M+H)+	1.28489874	0.99633791	NA	NA	35	0.62822782	0.76626909	orbitrapData	2
15	pRPLC_376	472.3032	772.908	Allocholic ac 2464-18-8	HMDB00005	C17737	HMDB00005	(M+CH3CN)+	0.28813692	0.9981552	NA	NA	0	0.99981552	hmdbMS1Da	3		
16	pRPLC_391	466.3292	746.577	LysoPA(18:0) NA	HMDB00111	NA	HMDB00111	(M+CH3CN)+	0.03145203	0.9999978	NA	NA	0	0.9999978	hmdbMS1Da	3		
17	pRPLC_629	181.072	36.36	Paraxanthin	611-59-6	HMDB00018	C13747	HMDB00018	(M+H)+	0.0153	0.9999948	NA	NA	0	0.9999948	hmdbMS1Da	3	
18	pRPLC_685	230.0701	158.205	1-[Methylsul] 132216-21-8	HMDB00330	NA	HMDB00330	(M+NH4)+	0.39101	0.9996603	NA	NA	0	0.9996603	hmdbMS1Da	3		
19	pRPLC_722	181.0721	228.305	Dihydropterri NA	HMDB00014	C05649	HMDB00014	(M+HCOO+2	0.02072	0.99999905	NA	NA	0	0.99999905	hmdbMS1Da	3		
20	pRPLC_778	289.2275	286.185	Bupivacaine	2180-92-9	HMDB00144	C07529	HMDB00144	(M+H)+	0.00282	0.9999998	NA	NA	0	0.9999998	hmdbMS1Da	3	
21	pRPLC_1148	282.875	40.947	Desflurane	57041-67-5	HMDB00153	C07519	HMDB00153	(M+2H+3K)+	0.248129	0.98641139	NA	NA	0	0.98641139	hmdbMS1Da	3	
22	pRPLC_1307	314.2326	401.848	9-Decenoylec NA	HMDB00132	NA	HMDB00132	(M+H)+	0.0347125	0.9999732	NA	NA	0	0.9999732	hmdbMS1Da	3		
23	pRPLC_1326	207.1292	406.754	2,3,5-Trimet 92233-85	HMDB00400	NA	HMDB00400	(M+H+2H2O)+	0.4956	0.93508793	NA	NA	0	0.93508793	hmdbMS1Da	3		
24	pRPLC_1560	264.0558	495.824	9-Chloroacetyl 6967-29-9	HMDB00328	NA	HMDB00328	(M+K)+	0.290635	0.9963052	NA	NA	0	0.9963052	hmdbMS1Da	3		
25	pRPLC_1799	699.3153	564.203	3,4,5-trihydri NA	HMDB01275	NA	HMDB01275	(M+H)+	0.56702888	0.99928576	NA	NA	0	0.99928576	hmdbMS1Da	3		
26	pRPLC_1835	351.2139	572.258	Corchorifatty 95341-44-9	HMDB00359	NA	HMDB00359	(M+Na)+	0.956835	0.99796755	NA	NA	0	0.99796755	hmdbMS1Da	3		
27	pRPLC_1860	249.1849	579.437	Dimethylben 891781-90-1	HMDB00322	NA	HMDB00322	(M+H)+	0.03803	0.99999679	NA	NA	0	0.99999679	hmdbMS1Da	3		
28	pRPLC_2064	419.2343	621.835	2-(3-Phenyl 3208-40-0	HMDB00361	NA	HMDB00361	(2M+K)+	0.93025776	0.99807878	NA	NA	0	0.99807878	hmdbMS1Da	3		
29	pRPLC_2065	660.4051	622.159	Astaxanthin	472-61-7	HMDB00022	C08580	HMDB00022	(M+CH3CN)+	4.19052773	0.96172822	NA	NA	0	0.96172822	hmdbMS1Da	3	
30	pRPLC_2171	568.3405	641.92	LysoPC(22:6) NA	HMDB00104	C04230	HMDB00104	(M+H)+	1.26179463	0.99646819	NA	NA	0	0.99646819	hmdbMS1Da	3		
31	pRPLC_2295	280.2627	669.073	6,10,14-Trim 762-29-8	HMDB00344	NA	HMDB00344	(M+NH4)+	0.298795	0.99980162	NA	NA	0	0.99980162	hmdbMS1Da	3		
32	pRPLC_2386	659.3539	699.573	3-[2-(3,7-din NA	HMDB01349	NA	HMDB01349	(2M+Na)+	0.45040686	0.98674451	NA	NA	0	0.98674451	hmdbMS1Da	3		
33	pRPLC_2615	508.3395	823.863	1-eicosanoyl NA	HMDB00623	NA	HMDB00623	(M+CH3CN)+	0.40219378	0.9996406	NA	NA	0	0.9996406	hmdbMS1Da	3		
34	pRPLC_3138	286.1439	537.015	Eugenyl benz 531-26-0	HMDB00320	NA	HMDB00320	(M+NH4)+	0.216305	0.99989603	NA	NA	0	0.99989603	hmdbMS1Da	3		
35	pRPLC_3633	180.962	40.947	Paranabic Ac 120-89-8	HMDB00628	NA	HMDB00628	(M+2H+3Na)+	0.543692	0.93642194	NA	NA	0	0.93642194	hmdbMS1Da	3		
36	pRPLC_3839	141.0297	63.436	4,5-Dihydro 155-54-4	HMDB00005	C00337	HMDB00005	(M+H+2H2O)+	0.490025	0.99946653	NA	NA	0	0.99946653	hmdbMS1Da	3		
37	pRPLC_3968	244.1545	87.115	Allixin	125263-70-9	HMDB00407	NA	HMDB00407	(M+NH4)+	0.24684	0.99986461	NA	NA	0	0.99986461	hmdbMS1Da	3	
38	pRPLC_4027	797.7993	99.964	Guanosine 3 NA	HMDB00604	C04494	HMDB00604	(M+2H+3K)+	1.16520415	0.99698743	NA	NA	0	0.99698743	hmdbMS1Da	3		
39	pRPLC_4250	473.1718	171.891	Arnamiol	102092-23-9	HMDB00350	NA	HMDB00350	(M+Na)+	0.28542211	0.97629873	NA	NA	0	0.97629873	hmdbMS1Da	3	
40	pRPLC_4633	810.8984	240.055	25-Methyl-1 75382-95-5	HMDB00299	NA	HMDB00299	(2M+NH4)+	2.11682248	0.99009177	NA	NA	0	0.99009177	hmdbMS1Da	3		
41	pRPLC_4673	379.2229	243.748	Neotame	165450-17-9	HMDB00345	NA	HMDB00345	(M+H)+	0.31905	0.99977382	NA	NA	0	0.99977382	hmdbMS1Da	3	
42	pRPLC_5044	286.2014	286.511	Prostaglandi 26054-67-1	HMDB02402	NA	HMDB02402	(M+NH4)+	0.0011325	1	NA	NA	0	1	hmdbMS1Da	3		
43	pRPLC_5461	310.2011	337.952	Prostaglandi 26054-67-1	HMDB02402	NA	HMDB02402	(M+CH3CN)+	0.3368825	0.99974783	NA	NA	0	0.99974783	hmdbMS1Da	3		
44	pRPLC_5721	517.1417	375.093	b-D-Glucuron NA	HMDB00397	NA	HMDB00397	(M+H)+	0.38926448	0.97479614	NA	NA	0	0.97479614	hmdbMS1Da	3		
45	pRPLC_5726	125.06	376.112	Benzene	71-43-2	HMDB00015	C01407	HMDB00015	(M+HCOO+2	0.51952	0.9994004	NA	NA	0	0.9994004	hmdbMS1Da	3	
46	pRPLC_7083	229.0973	516.577	Mukonal	20323-67-5	HMDB00302	NA	HMDB00302	(M+NH4)+	0.0869075	0.99998322	NA	NA	0	0.99998322	hmdbMS1Da	3	
47	pRPLC_7291	175.0968	536.688	1,3-Diacetyl NA	HMDB00291	NA	HMDB00291	(M+HCOO+2	0.59768	0.99920649	NA	NA	0	0.99920649	hmdbMS1Da	3		
48	pRPLC_7637	563.1516	563.24	Dukunolide C 99343-74-5	HMDB00352	NA	HMDB00352	(M+Na)+	1.46184073	0.99526242	NA	NA	0	0.99526242	hmdbMS1Da	3		

› Other Functions and Website of metID

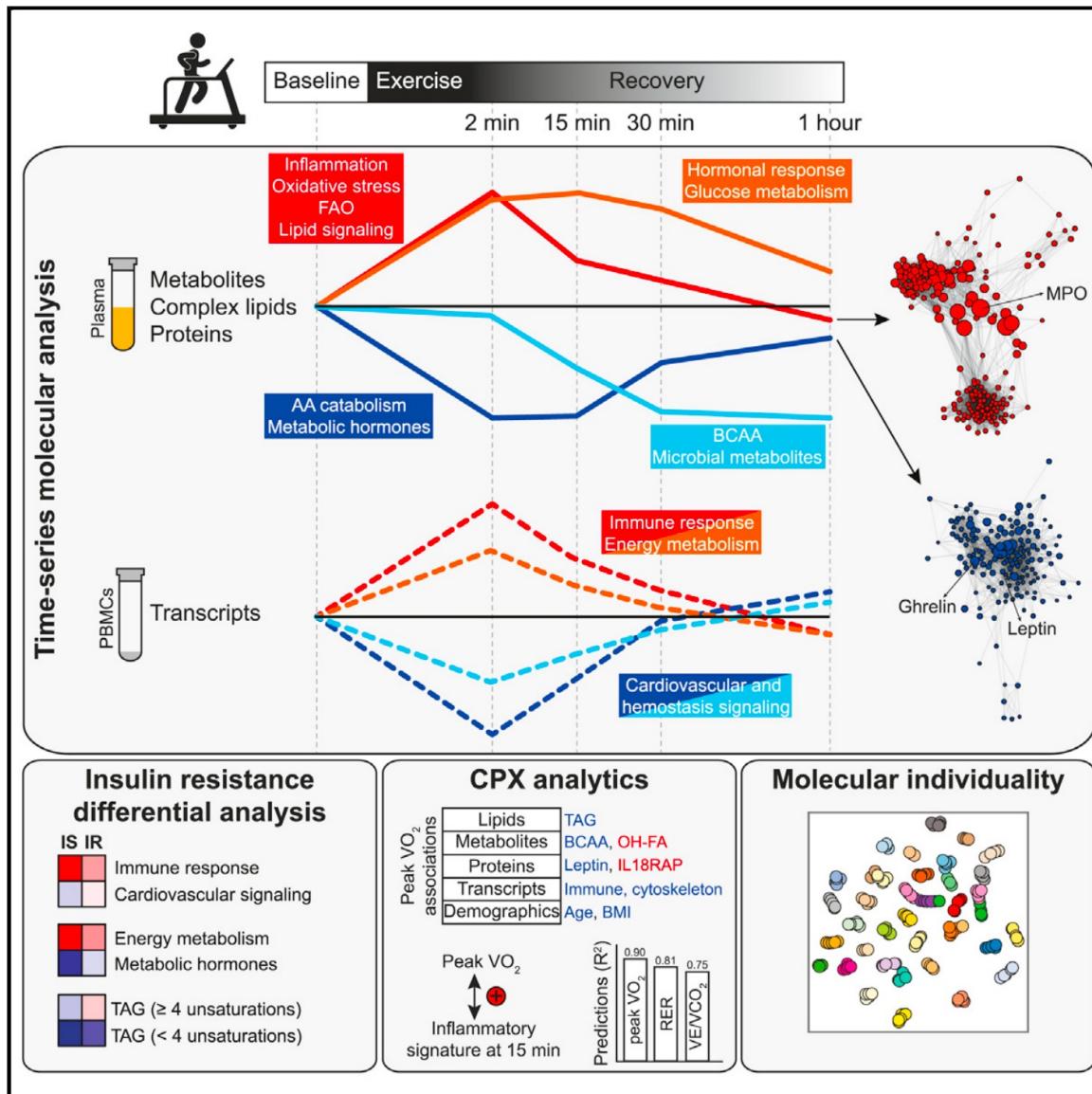
<https://jaspershen.github.io/metID/>

The screenshot shows a dark-themed website for 'metID 1.1.0'. At the top, there's a navigation bar with links for 'Help documents', 'News', 'Public databases', 'Reference', and 'Get started'. Below the navigation, a dropdown menu is open under 'Set parameter I'. The menu items are:

- Brief introduction
- Brief introduction
- In-house and public database construction using metID
- In-house database construction
- Public database construction
- Output metID database to mgf/msp for other tools
- Correct retention times in database according to internal standards
- Metabolite identification
- Metabolite identification using MS1 database
- Metabolite identification using MS2 database
- Metabolite identification using multiple databases** (This item is highlighted with a dark background)
- Identify single peak
- Others
- Other useful tools
- Test

The main content area of the website has a heading 'Set parameter I' and some explanatory text about using the 'identify' function with R code examples.

Case Study

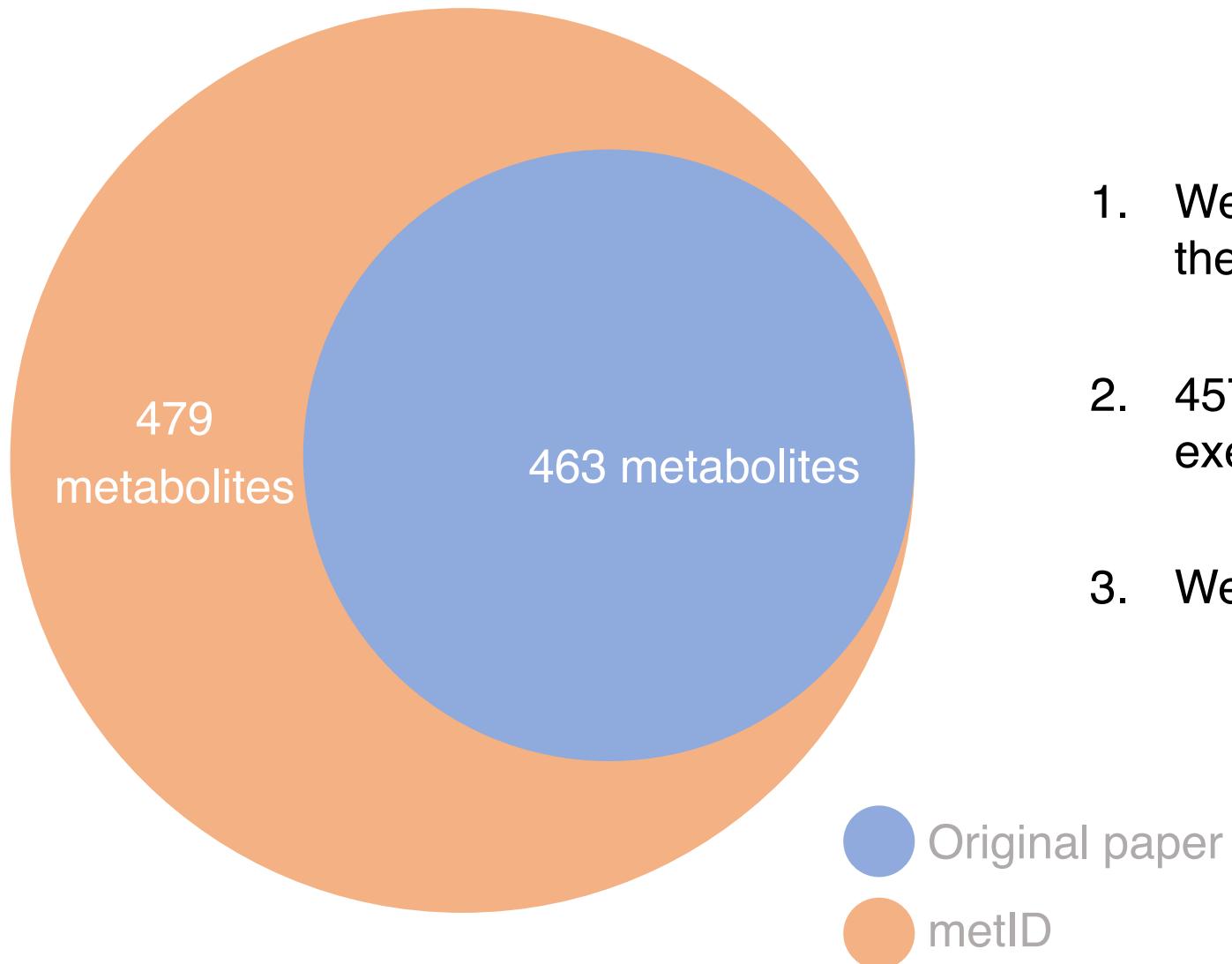


Variable	Information
Sample type	Plasma
Subject sample number	310
QC sample number	36
Data acquisition	Untargeted metabolomics
Instrument platform	LC-MS
Chromatographic condition	RPLC + HILIC
Polarity mode	Positive + negative

**463 manually annotate metabolites.
(Golden standard)**

Case Study

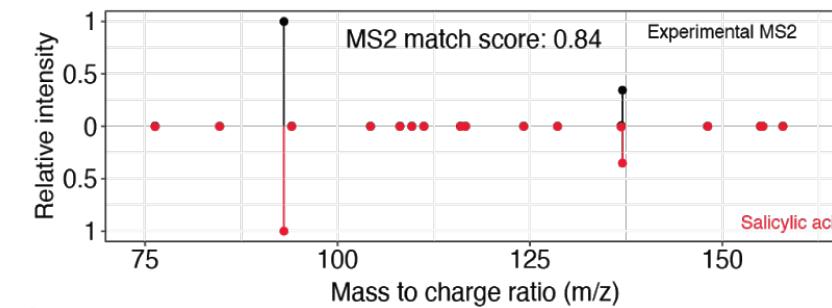
Same dataset processed by metID



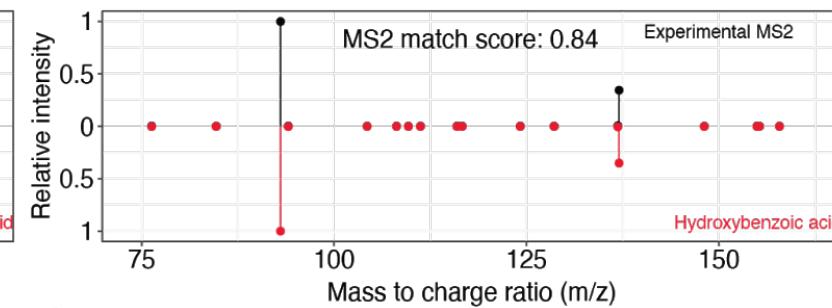
1. We retrieved all the 463 metabolites annotated in the original paper.
2. 457 out of 463 annotation are consistent with exercise paper.
3. We additionally annotate 479 more metabolites.

Non-consistent Annotations

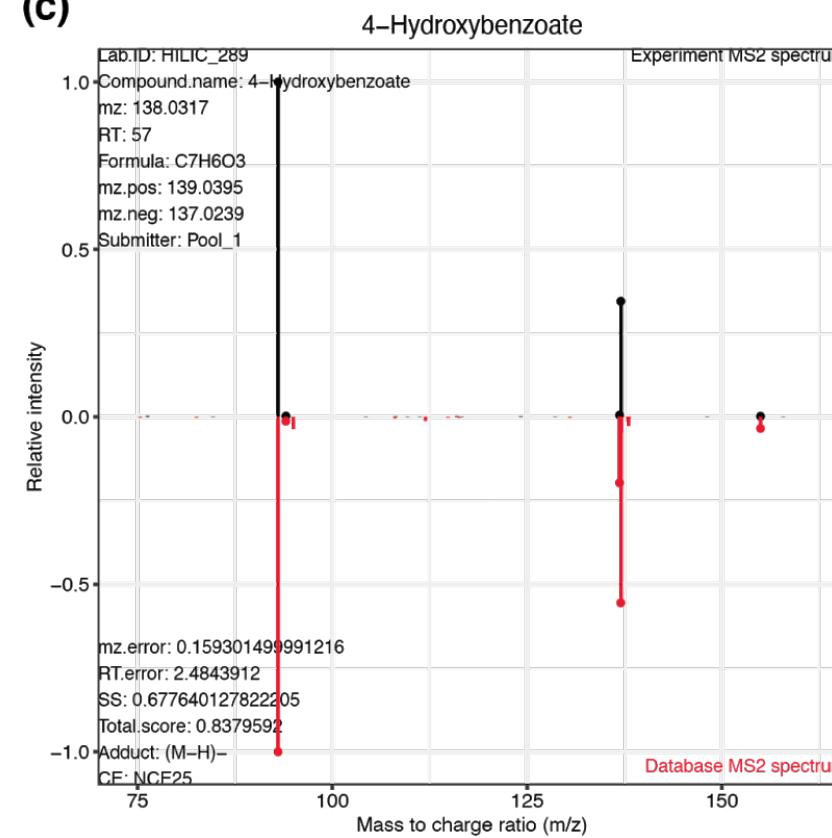
(a)



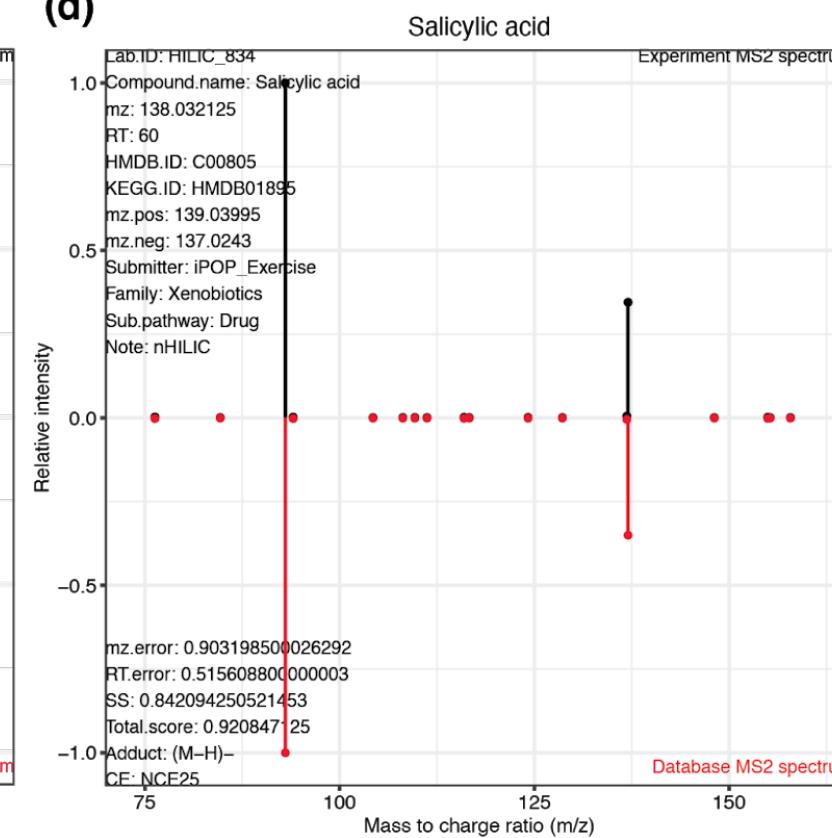
(b)



(c)



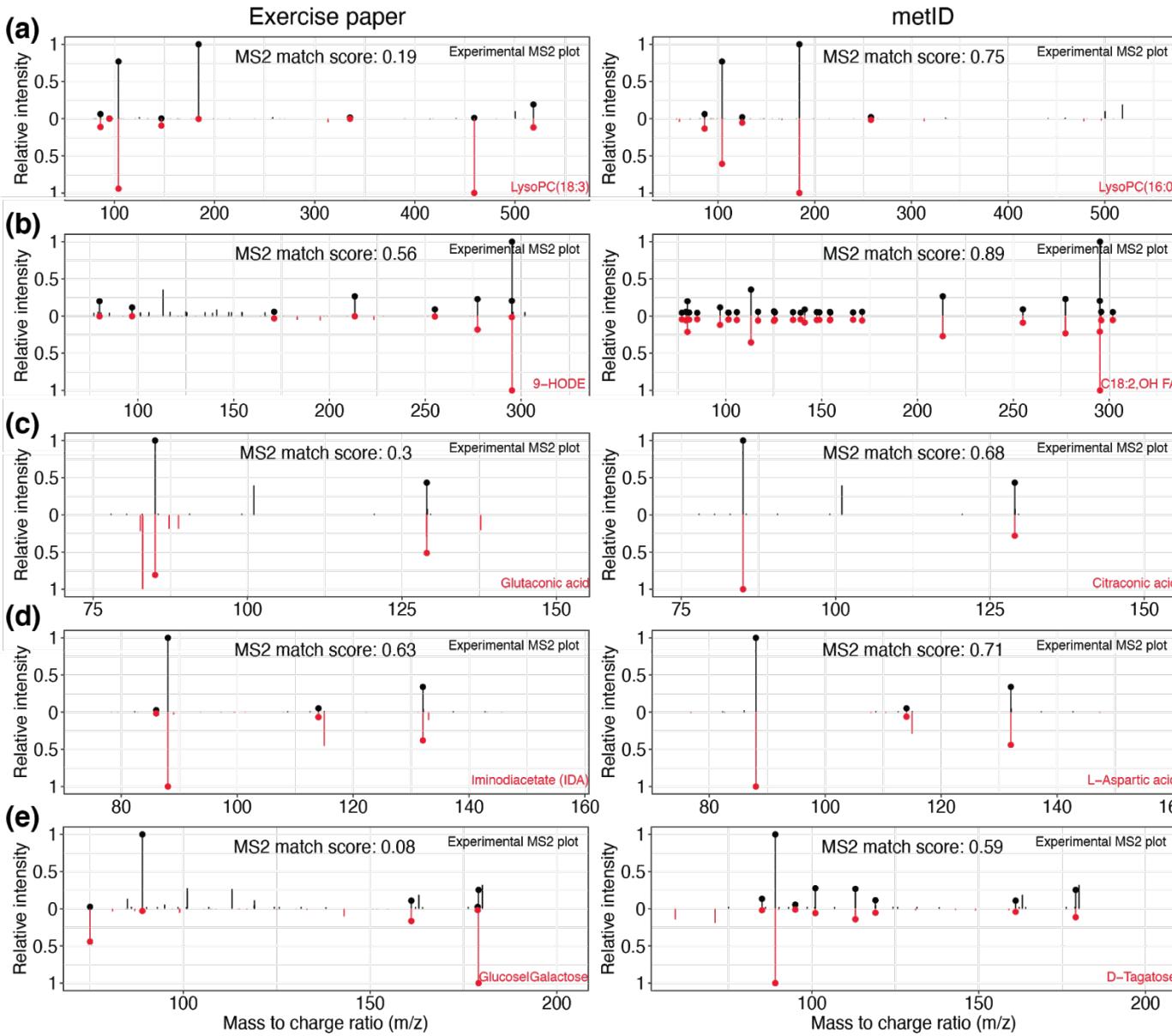
(d)



peak 1.00_137.0243m/z

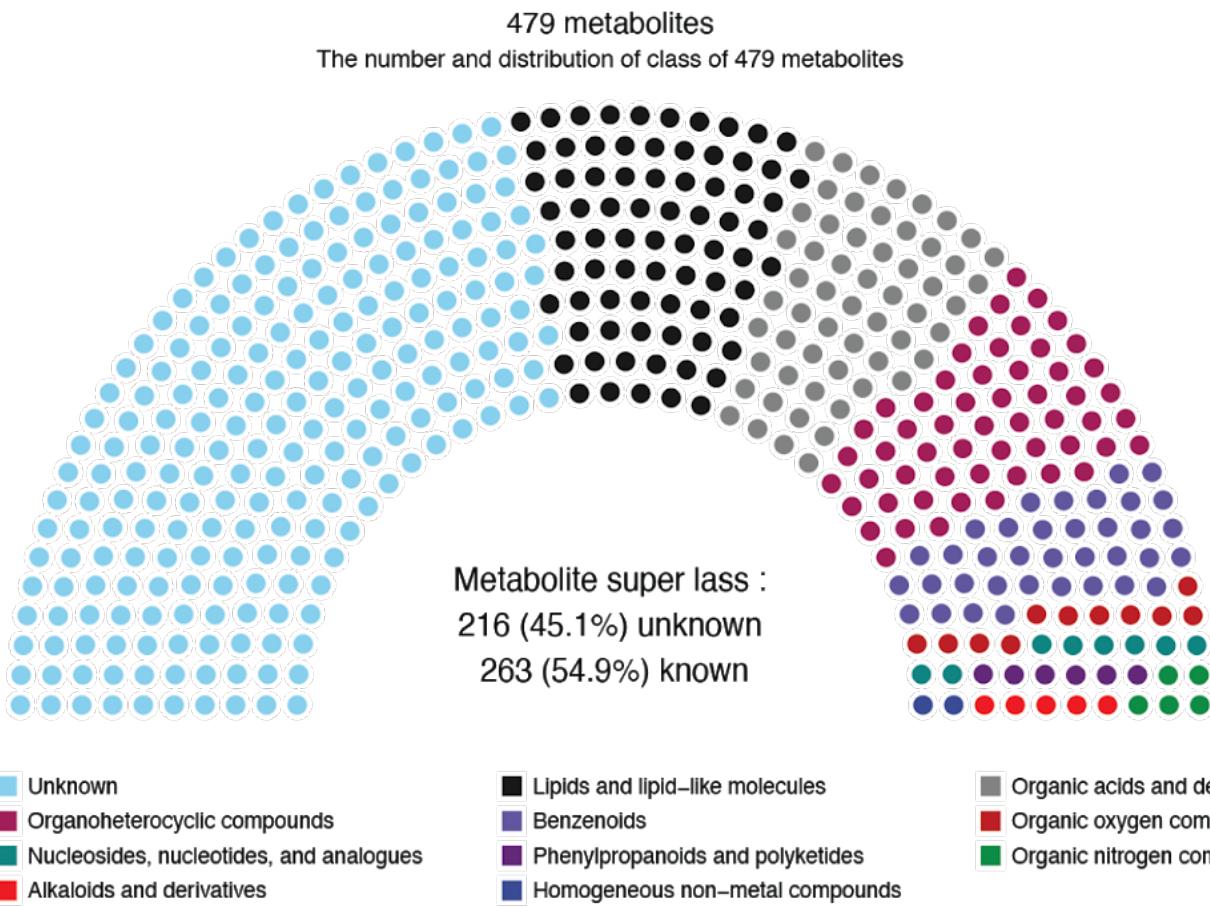
For metID, rank 2 annotation is consistent with the original paper annotation.

Non-consistent Annotations

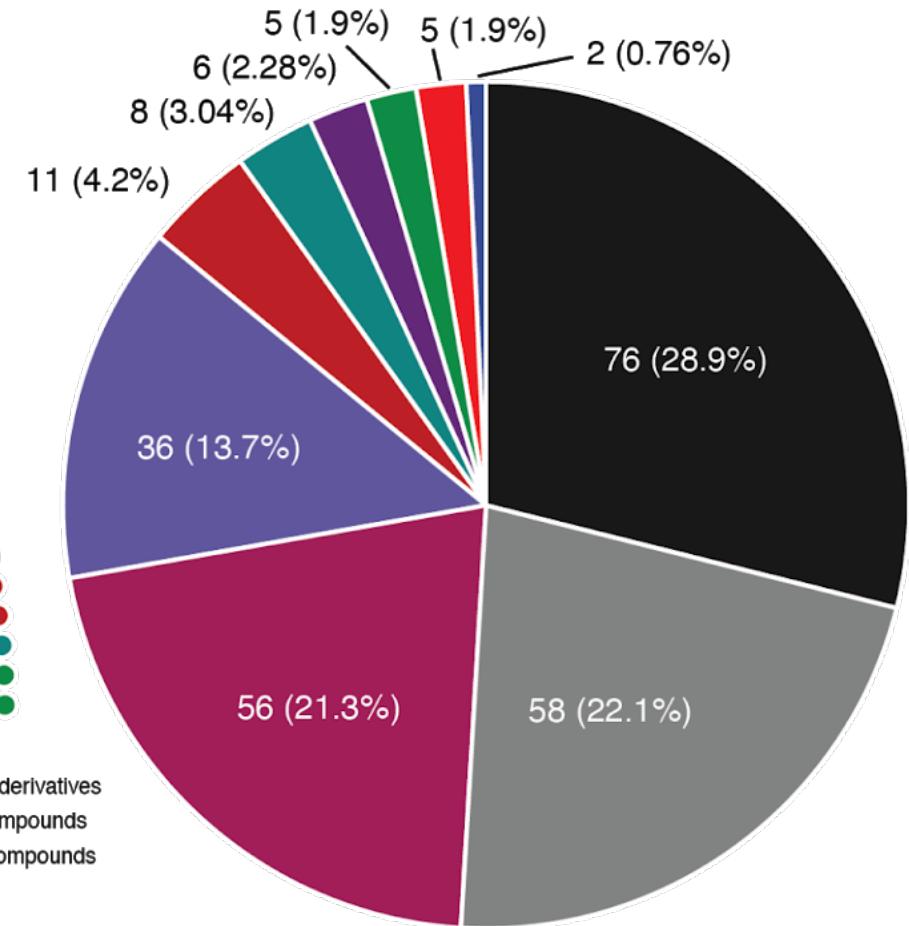


➤ 479 More Metabolites

(a)



(b)



Systems biology

metID: an R package for automated compound annotation for LC–MS-based data

Xiaotao Shen  ^{1,†}, Si Wu ^{1,†}, Liang Liang ¹, Songjie Chen ¹, Kévin Contrepois ¹, Zheng-Jiang Zhu ^{2,*} and Michael Snyder ^{1,*}

¹Department of Genetics, Stanford University School of Medicine, Stanford, CA 94304, USA and ²Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

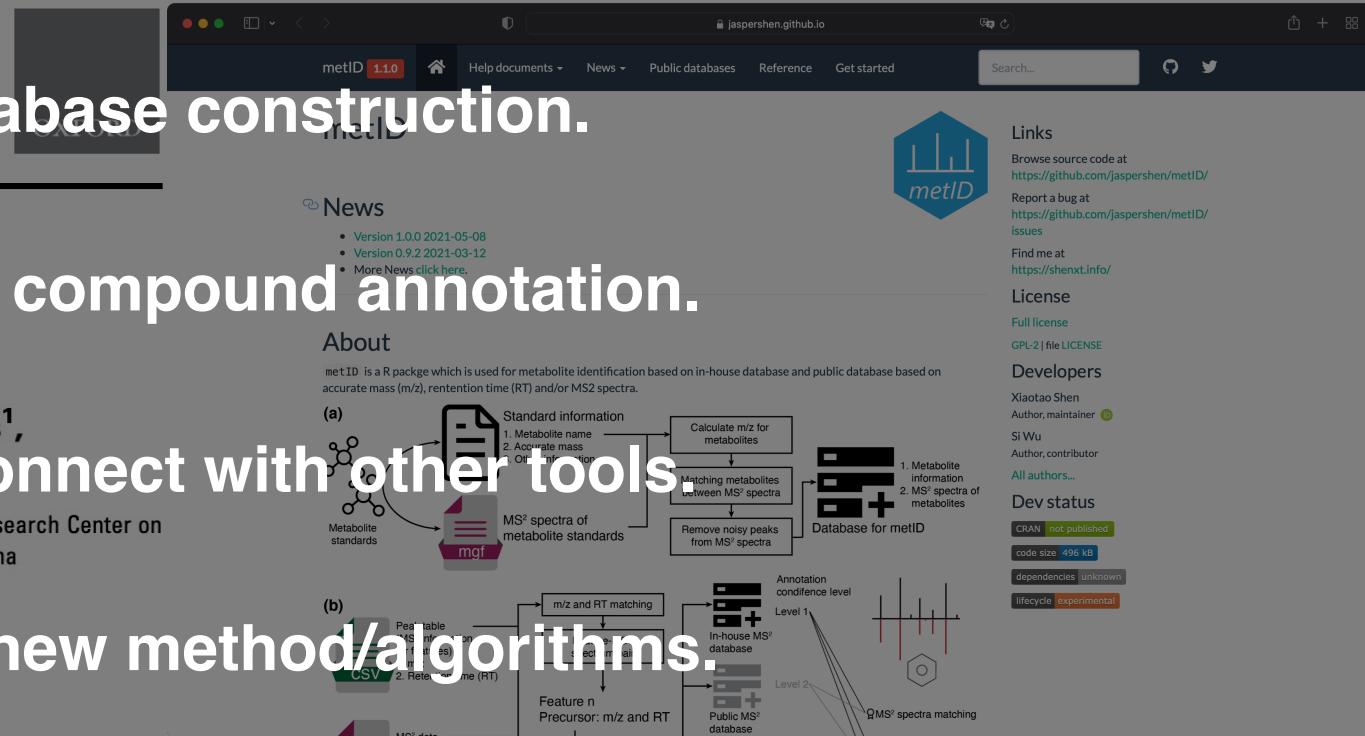
Associate Editor: Olga Vitek

Received on May 17, 2021; revised on July 13, 2021; editorial decision on July 31, 2021

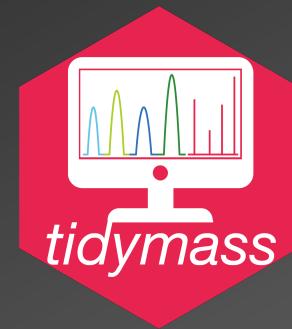


In-house database construction. Automatable compound annotation.

</> Flexible to connect with other tools.
Easy to add new method/algorithms.



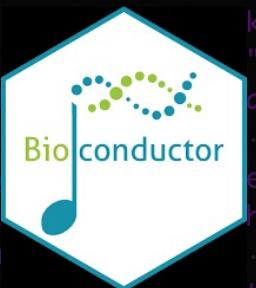
<https://jaspershen.github.io/tidymass/>



TidyMass: A Computational Framework for LC-MS Data Processing and Analysis

R is (or maybe) the most important language for bioinformatics analysis

```
6 #' @param from From.github or gitee, if you are in China, try to set this as "gitee".  
7 #' @param force Force installation even if the remote state has not changed since the previous install  
8 #' @param update CRAN package check for updates  
9 #' "default" respects the value of the R_REMOTES_UPGRADE environment variable if set,  
10 #' and falls back to "ask" if unset. "ask" prompts the user for which out of date  
11 #' packages to upgrade. For non-interactive sessions "ask" is equivalent to "always".  
12 #' TRUE and FALSE are also accepted and correspond to "always" and "never" respectively.  
13 #' @param dependencies Which dependencies do you want to check? Can be a character vector  
14 #' (selecting from "Depends",  
15 #' or a logical vector. TRUE is  
16 #' and "Suggests". NA is short  
17 #' for "NA". Default: FALSE means  
18 #' just check this package.  
19 #' The argument "NA" means the  
20 #' @param which_install What packages you want to install? Default is all. You can set it as a character vector.  
21 #' @param ... parameters from devtools::install_github  
22 #' or tools::install_github  
23 #' or tools::install_git  
24 #' @param ...  
25 #' @exp...  
26  
27 install_tidymass <->  
28 .function(from = c("github", "git",  
29 .force = FALSE,  
30 .upgrade = "never",  
31 .dependencies = NA,  
32 .demo_data = TRUE,  
33 .which_package = c("all", "metID", "metflow2", "lipidflow", "demoData", "metPath"),  
34 .)  
35 .from = match.arg(from)  
36 .which_package = match.arg(which_package)  
37 .which_package = stringr::str_to_lower(which_package)  
38 .  
39 .##detach.packages  
40 .if("metID" %in% search()){  
41 .  detach("metID")  
42 .}
```



Bioconductor



Comprehensive R Archive Network (CRAN)

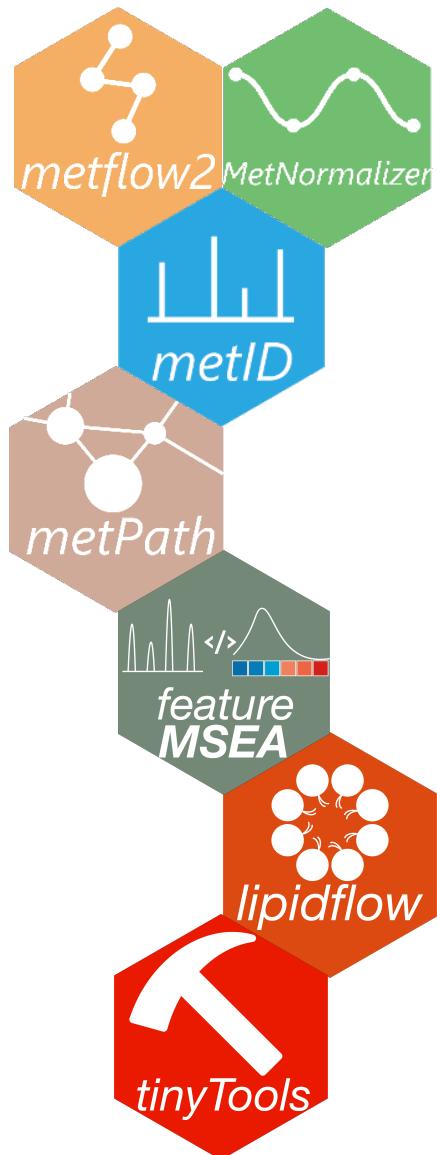


GitHub



R-Forge

› TidyMass: Collection of R Packages for Mass Spectrometry Data Analysis



lifecycle maturing

metflow2: Raw MS data processing and data cleaning

<https://jaspershen.github.io/metflow2/>

lifecycle maturing

metNormalizer: Data normalization and integration

<https://jaspershen.github.io/MetNormalizer/>

lifecycle maturing

metID: Compound identification

<https://jaspershen.github.io/metID/>

lifecycle experimental

metPath: Biological information mining

<https://jaspershen.github.io/metPath/>

lifecycle experimental

lipidflow: Lipidomics data analysis. (Chuchu Wang)

<https://jaspershen.github.io/lipidflow/>

lifecycle experimental

fMSEA: feature based metabolite set enrichment analysis

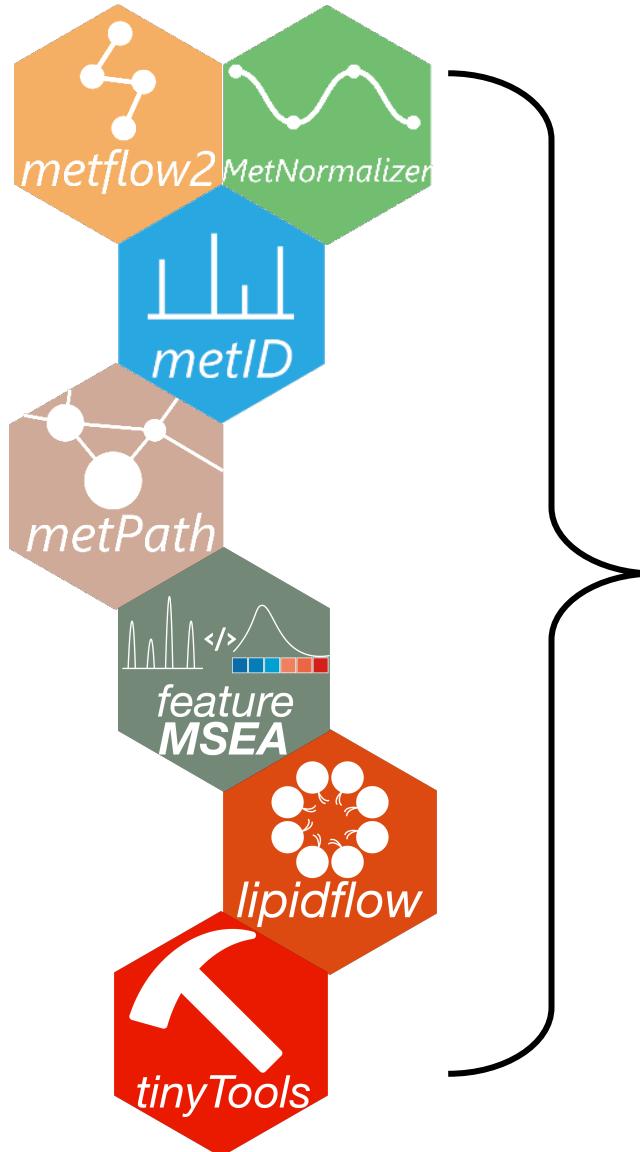
<https://jaspershen.github.io/fMSEA/>

lifecycle maturing

tinyTools: Toolkit for MS data processing.

<https://jaspershen.github.io/tinyTools/>

› TidyMass: Collection of R Packages for Mass Spectrometry Data Analysis



Screenshot of the `tidymass` GitHub project page (<https://jaspershen.github.io/tidymass/>):

The page displays the following content:

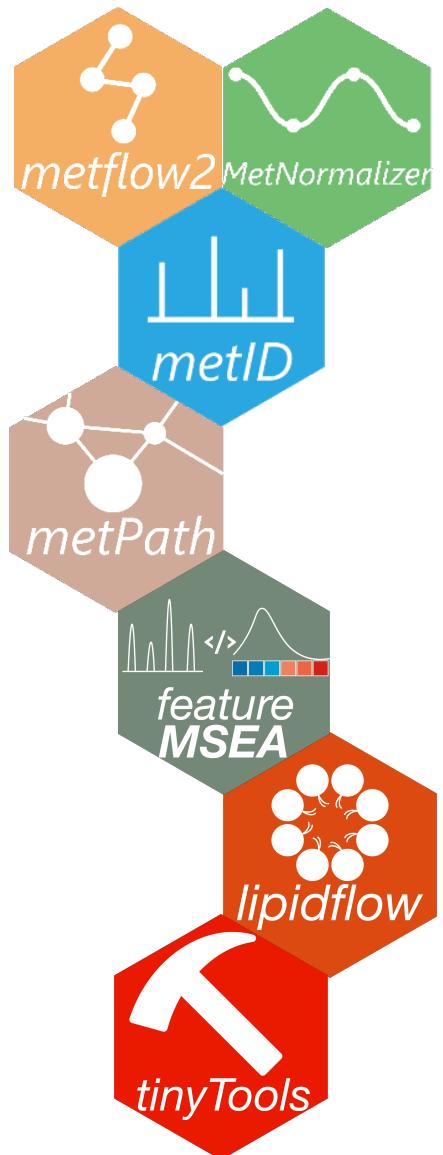
- tidymass 0.0.1** (version)
- About**: `tidymass` is a collection of R packages for mass spectrometry data processing, analysis.
- Installation**:
 - You can install `tidymass` from [Github](#).
 - Code snippet:

```
if(!require(devtools)){
  install.packages("devtools")
}
devtools::install_github("jaspershen/tidymass")
```
 - Then you can use `install_tidymass()` to install all the packages in `tidymass`.
 - Code snippet:

```
library(tidymass)
tidymass::install_tidymass(from = "github", force = FALSE)
```
- Usage**: `tidymass` contains several packages:
 - metflow2** (represented by an orange hexagon icon)
- Links**:
 - Browse source code at <https://github.com/jaspershen/tidymass/>
 - Report a bug at <https://github.com/jaspershen/tidymass/issues>
 - Find me at <https://shenxt.info/>
- License**: MIT + file [LICENSE](#)
- Developers**: Xiaotao Shen (Author, maintainer)
- All authors...**
- Dev status**:
 - CRAN: not published
 - code size: 13.4 kB
 - dependencies: unknown
 - lifecycle: experimental

<https://jaspershen.github.io/tidymass/>

› TidyMass: Collection of R Packages for Mass Spectrometry Data Analysis



metflow2 0.9.2 metflow2
MetNormalizer 1.3.02 MetNormalizer
metID 1.0.0 metID
lipidflow 0.0.1 lipidflow
metPath 0.0.9 metPath
tinyTools 0.9.0 tinyTools

Links
Browse source code at <https://github.com/jaspershen/metflow2/>

Links
Browse source code at <https://github.com/jaspershen/MetNormalizer/>

Links
Browse source code at <https://github.com/jaspershen/metID/>

Links
Browse source code at <https://github.com/jaspershen/lipidflow>

Links
Browse source code at <https://github.com/jaspershen/metPath/>

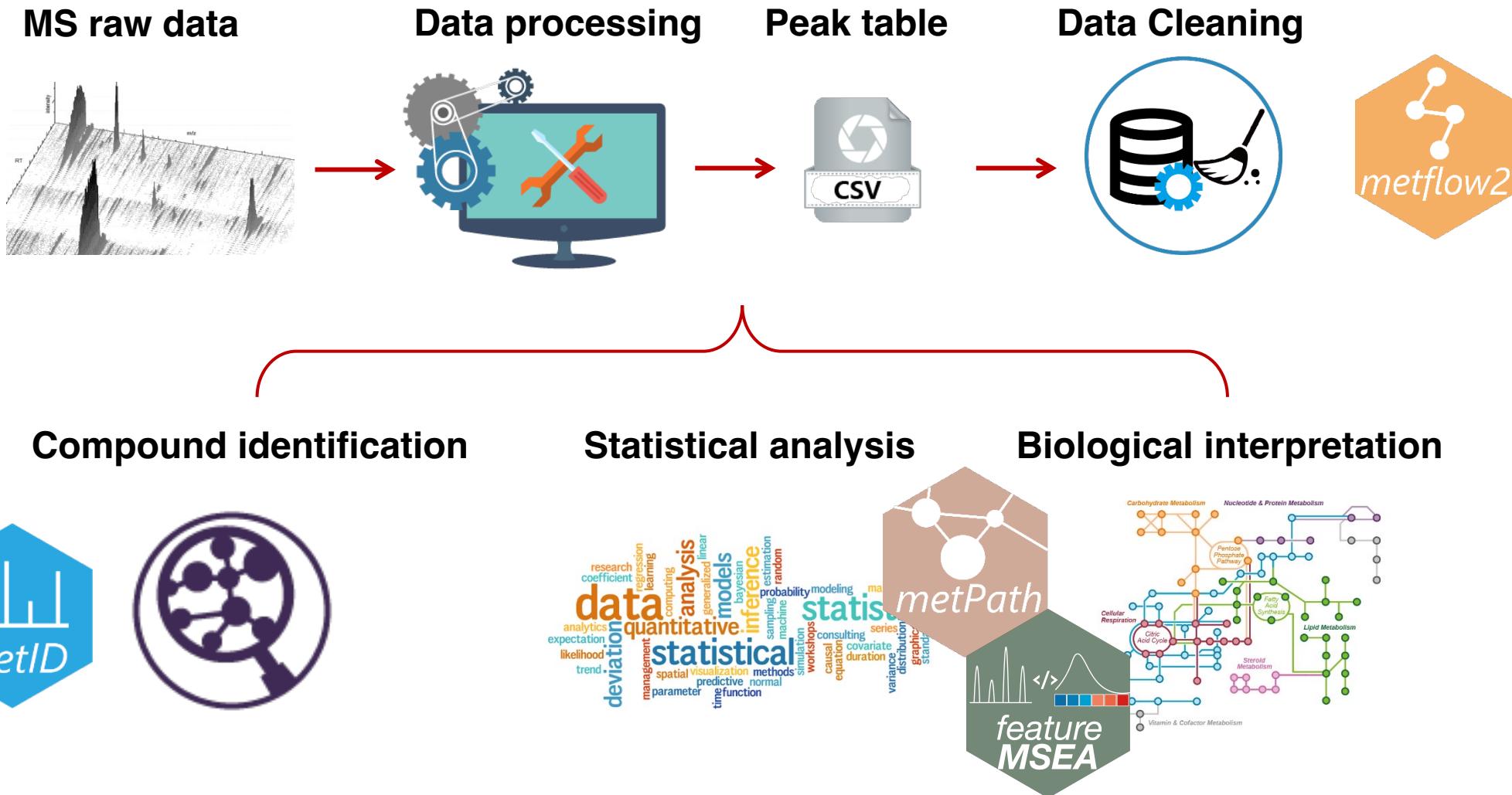
Links
Browse source code at <https://github.com/jaspershen/tinyTools>

<https://jaspershen.github.io/tidymass/>

32

Workflow of LC-MS Data Processing and Analysis

Same data type across all analysis processing (expression data, variable information and metadata)!



› ACKNOWLEDGEMENTS



Dr. Si Wu
(Stanford University)



Dr. Liang Liang
(Stanford University)



Dr. Songjie Chen
(Stanford University)



Dr. Kévin Contrepois
(Stanford University)



Prof. Zheng-Jiang Zhu
(CAS)



Prof. Michael Snyder
(Stanford University)



Stanford
MEDICINE

department of
GENETICS



中国科学院大学
University of Chinese Academy of Sciences

Thanks for your attention!

Q&A

Xiaotao Shen PhD

Stanford University School of Medicine
Department of Genetics



shenxt@stanford.edu



shenxt.info



@xiaotaoshen1990



github.com/jaspershen