



pcpfm: A Python-centric pipeline for high-fidelity and high-performance LC-MS metabolomics data processing.

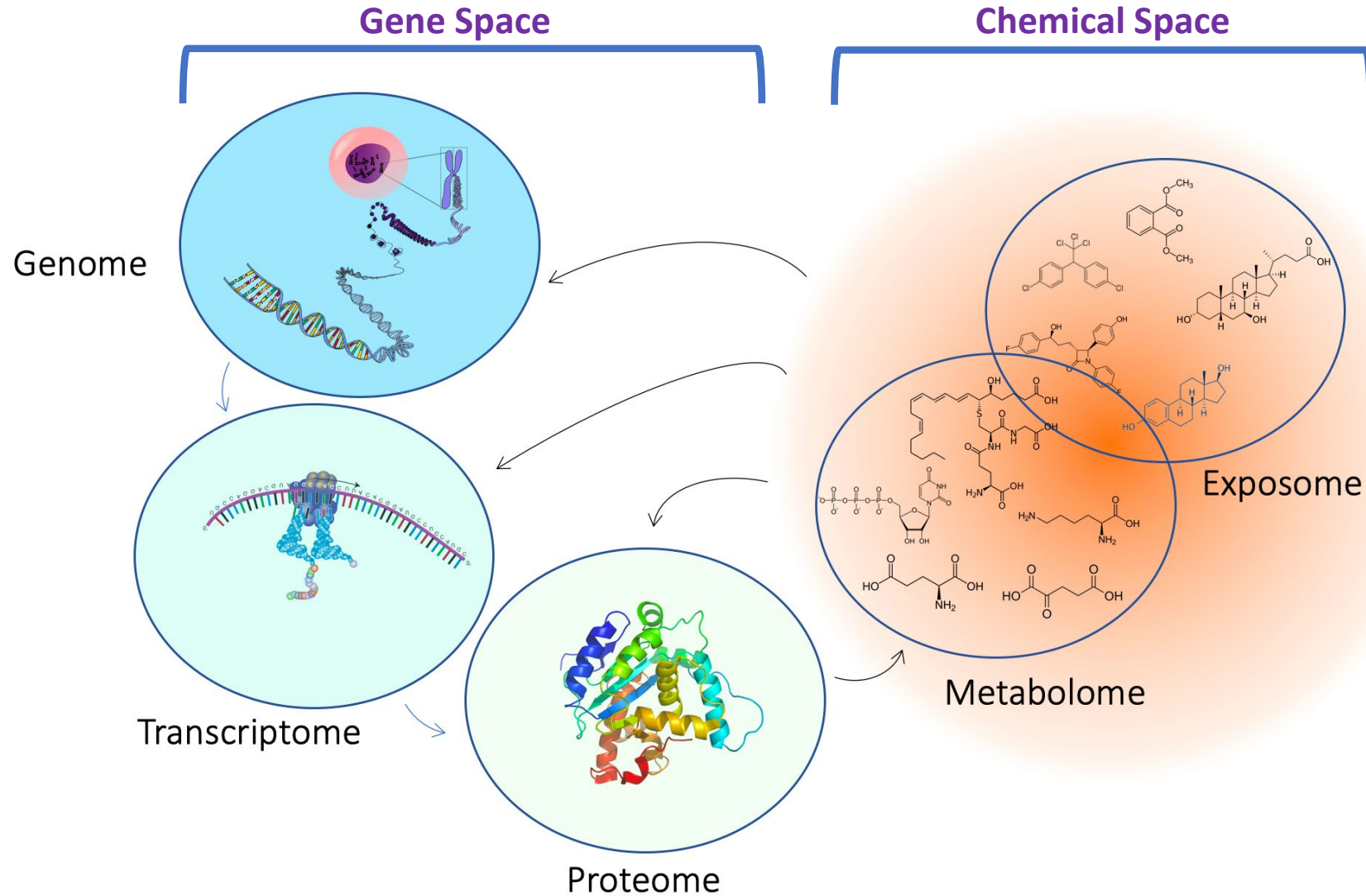
Joshua Mitchell M.D./Ph.D.

Data Scientist

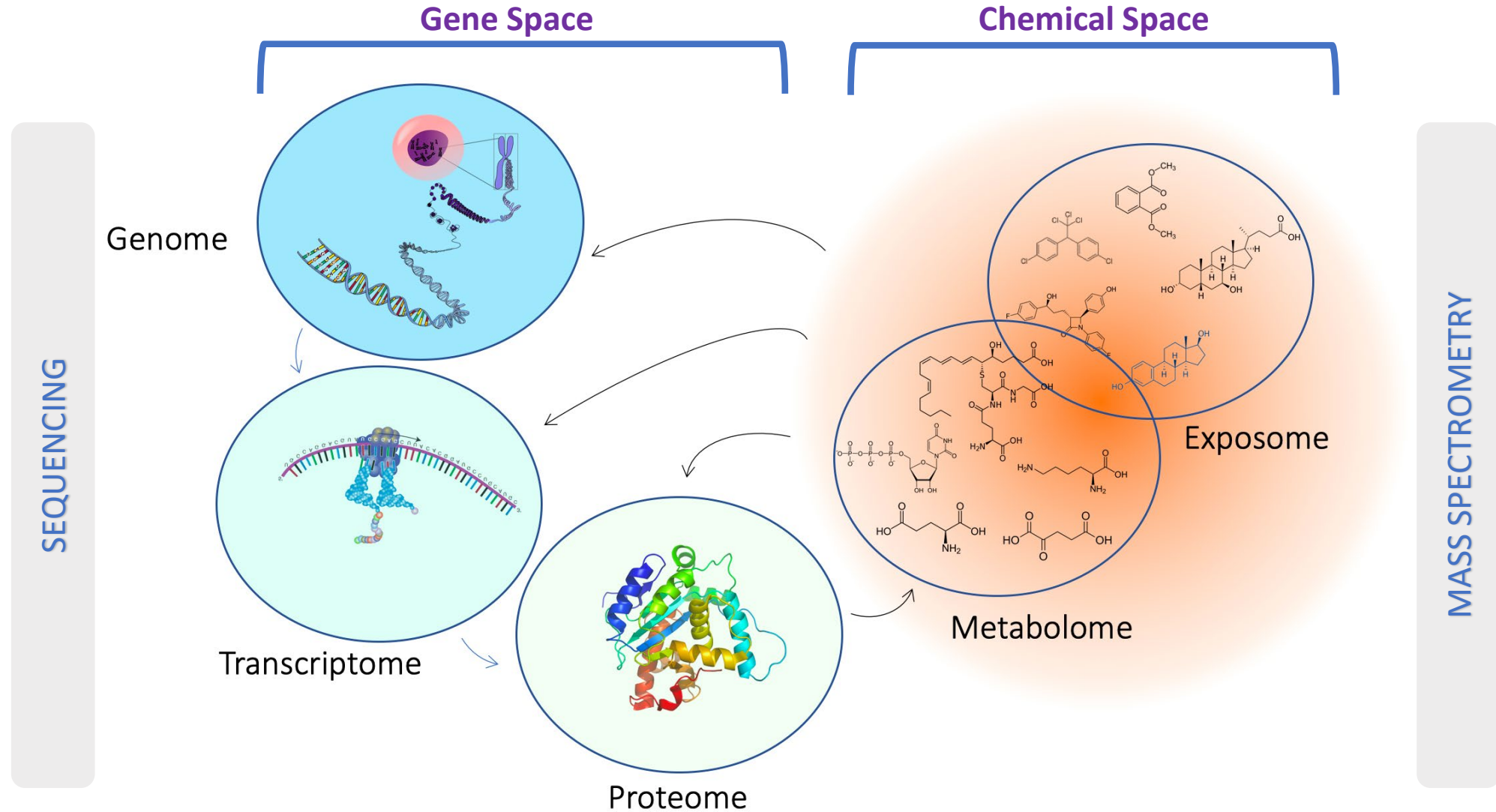
Shuzhao Li Group

The Jackson Laboratory for Genomic Medicine

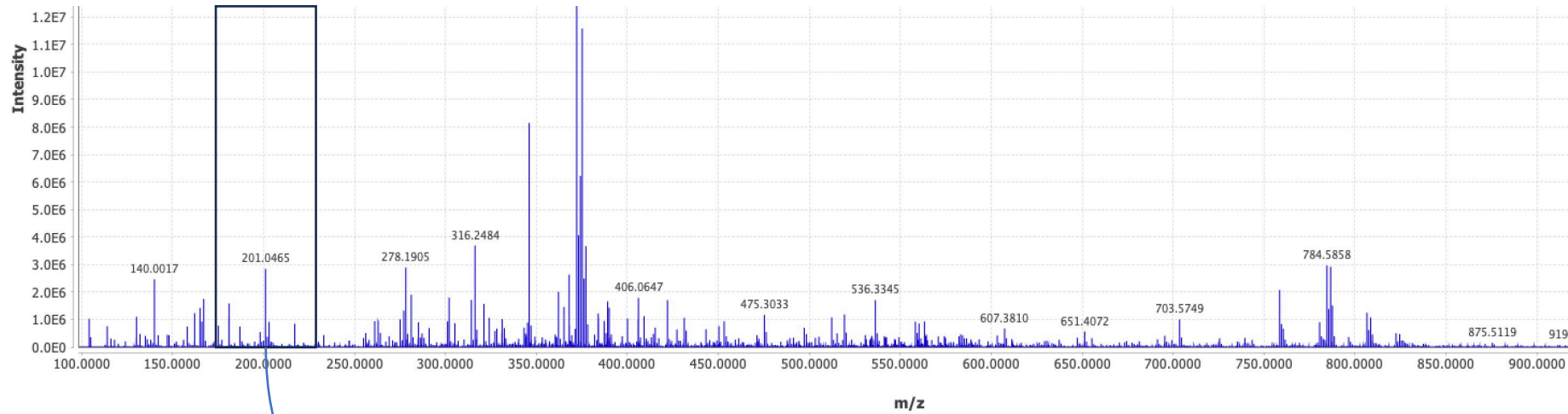
Metabolomics is the comprehensive measurement of biological chemical space



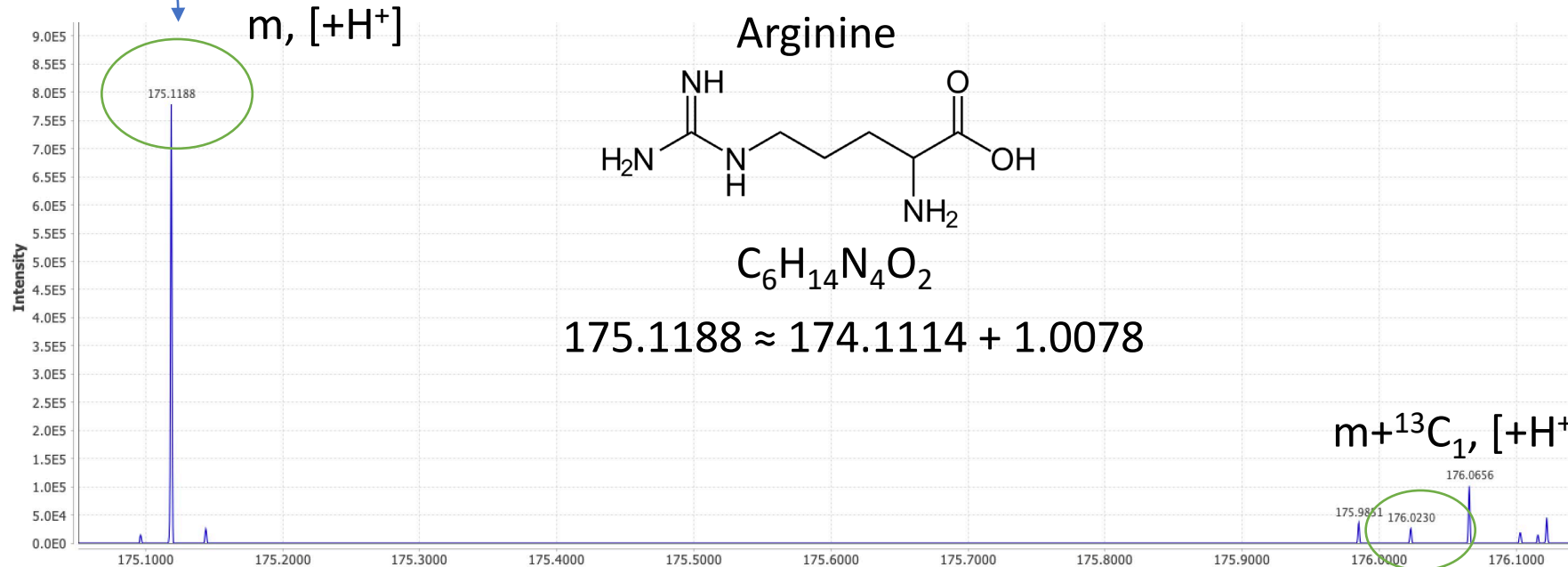
Metabolomics is the comprehensive measurement of biological chemical space



Advantages and challenges of Mass Spectrometry

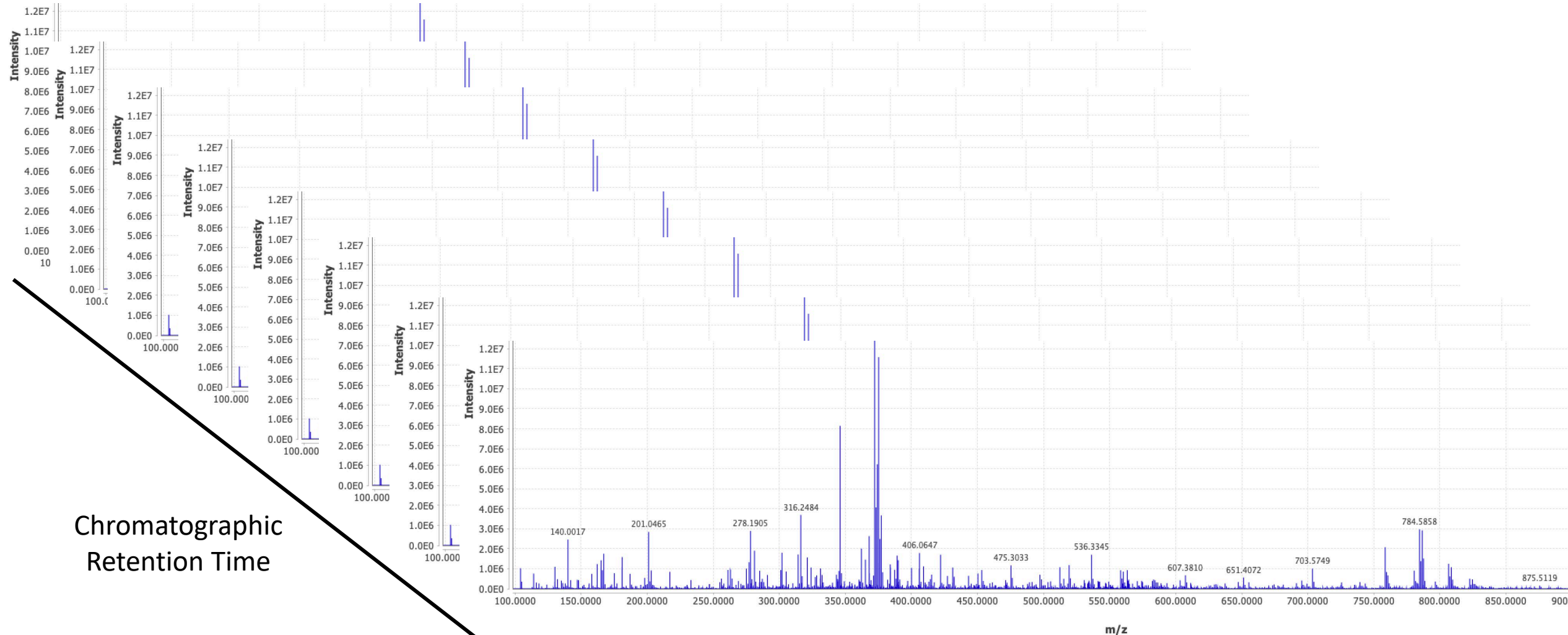


A single mass spectrum (i.e., scan) of a complex mixture is a lot of data.



Each scan is a composition of spectra for many compounds, including different adducted forms, isotopologues, etc.

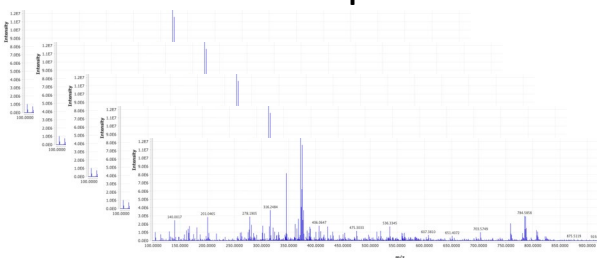
Advantages and challenges of Mass Spectrometry



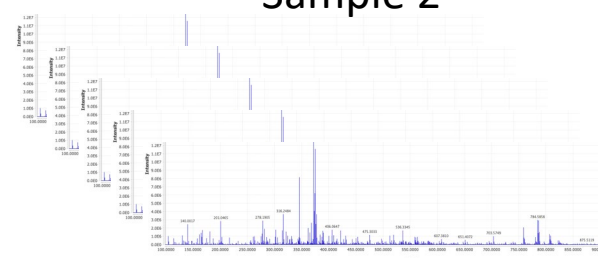
Additionally, most experiments couple chromatography to mass spectrometry meaning multiple spectra for each sample. Data becomes complex quickly.

LC-MS Pre-Processing

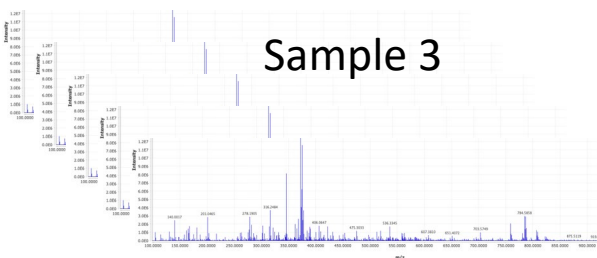
Sample 1



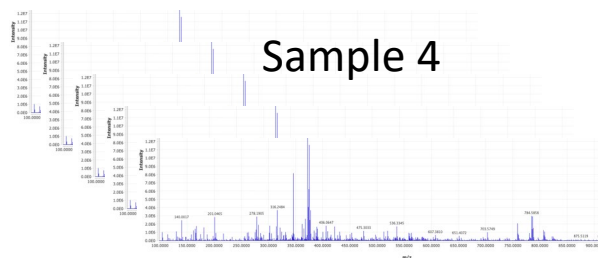
Sample 2



Sample 3



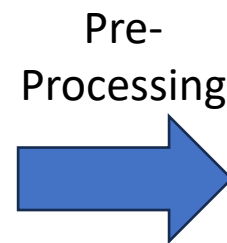
Sample 4



...

Feature Table

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	id_number	rt	rt_min	rt_max	parent_mass	area	collected	goodness	fit	detection	col_MRT	col_ZScore	Blank	col_ZScore
2	F2	124.0151	38.21	36.36	398	409843	1	0.98	467	5	0	522134	1045352	
3	F4	61.0204	206.31	203.53	207.46	8	79540284	1	0.98	3035	63	1193957	0	
4	F5	61.0204	209.31	207.46	210.29	8	15489437	1	0.95	4713	63	2304811	0	
5	F8	61.9882	206.31	203.53	209.08	11	1217045114	1	0.93	9	69	2643976	3604328	
6	F9	124.0402	28.07	26.92	30.56	400	4448312	0.96	0.99	6	10	0	0	
7	F10	124.0402	34.03	31.25	36.82	400	3207732	0.91	0.93	3	10	0	0	
8	F11	124.0402	119.08	116.14	122.03	400	8482945	0.93	0.86	4	10	0	0	
9	F14	124.0422	218.87	216.46	221.65	401	2066174	1	0.99	212	1	0	0	
10	F15	124.0402	24.5	23.12	26.26	402	1514915	1	0.99	205	2	0	0	
11	F16	155.952	189.04	186.27	191.81	709	465638855	1	0.94	8091	60	7094041	0	
12	F18	124.9838	56.48	53.7	59.27	404	183998857	1	0.97	4374	53	3482320	0	
13	F20	124.9912	39.78	38.88	43.57	405	129291374	1	0.97	37	67	5817462	0	
14	F21	124.9912	170.11	168.3	171.96	405	121408353	1	0.98	3	67	2716889	0	
15	F23	125.0248	208.62	207	210.7	406	21908826	1	1	3403	49	305161	0	
16	F24	156.0411	31.48	29.17	34.27	712	3486995	1	0.92	227	1	0	0	
17	F25	125.006	208.39	207	210.24	408	11054293	1	0.99	1680	11	0	0	
18	F27	61.9892	56.48	53.93	59.04	12	2847953	1	0.77	63	63	516391	84573	
19	F31	61.9892	205.61	195.04	210.93	12	64184184	1	0.87	77	67	355627	7582	
20	F32	125.0066	208.62	207	210.47	409	46530908	1	1	7118	45	783456	0	
21	F33	125.0106	208.62	206.77	210.7	410	21137162	1	1	31393	62	263046	0	
22	F34	156.0573	53	50.22	55.79	714	2588500	1	0.98	756	2	0	0	
23	F35	156.0573	63.91	61.12	66.69	714	538750	0.92	0.95	99	2	0	0	
24	F36	156.8657	165.08	162.3	166.46	715	47623515	1	0.98	43387	62	6902097	0	
25	F37	156.8657	168.53	166.46	171.26	715	835602376	1	0.95	58455	62	1644864	0	
26	F44	174.9373	306.61	303.89	309.32	1013	109493484	1	0.99	260	44	251341283	489439001	
27	F45	174.9373	318.92	314.89	319.69	1013	154628462	0.75	0.95	1077	60	73862	912745	
28	F46	125.0241	24.72	22.19	27.37	411	4888171	1	0.98	22	38	867415	0	
29	F51	125.0241	209.31	206.77	212.09	411	5036982	0.88	0.87	39	38	1133949	0	
30	F59	156.9267	306.83	304.12	309.12	717	8049979	1	0.99	3467	34	3519972	6074216	
31	F60	156.9267	316.92	314.16	319	717	7514063	0.79	0.88	771	6	0	0	
32	F61	156.9267	322	320.84	323.98	717	72709	0.71	0.59	127	1	0	0	
33	F65	125.0431	215.98	213.18	218.16	412	47515139	1	0.96	27	69	6518415	6929543	
34	F66	125.0431	219.34	218.16	221.11	412	24869889	1	0.97	19	69	3284530	2473004	
35	F70	125.0607	26.48	24.5	29.17	413	6001614	1	0.98	47	4	900609	2021984	
36	F76	125.0971	26.48	25.38	28.95	414	4000002	1	0.98	46	6	0	0	
37	F77	174.9869	26.7	24.5	28.4	1017	62920885	1	0.99	71607	62	12941486	0	
38	F78	68.959	40.25	38.88	43.04	30	626299	1	0.82	733	3	0	0	
39	F79	174.9897	30.1	28.73	32.18	1018	227031544	1	0.99	36363	35	0	0	
40	F80	174.9897	33.11	32.18	35.89	1018	20337941	1	0.94	1434	34	0	0	
41	F81	156.9513	188.81	186.04	191.58	720	1981386	1	0.83	486	11	0	0	
42	F83	68.9561	40.48	39.11	43.27	24	87821	1	0.93	102	1	0	0	
43	F84	125.9945	39.78	38.21	42.57	416	3050872	1	0.99	447	3	0	0	
44	F85	126.001	208.39	206.31	210.93	418	60261661	1	1	59310	61	5323458	0	
45	F86	174.9911	30.1	28.95	32.41	1019	2338306	1	0.98	94	6	4187096	0	
46	F93	126.0124	208.62	207	210.47	419	5152956	1	1	7694	56	707113	0	
47	F95	126.0307	24.07	21.6	30.38	420	1583846	1	0.92	488	7	366916	0	
48	F97	156.9909	200.88	198.05	203.76	722	88670676	1	1	212	58	37725493	995728	
49	F100	126.0664	218.87	215.98	221.65	421	1484815	1	0.98	120	1	0	0	
50	F101	174.9939	28.26	25.6	28.95	1021	7944469	1	0.91	166	2	0	0	
51	F102	174.9939	29.86	28.95	31.95	1021	2503796	1	0.99	62	1	0	0	
52	F103	126.9041	31.48	29.17	34.27	422	60254379	1	0.99	10	69	1133420	230222	
53	F104	174.9952	26.48	25.38	28.95	1022	9540391	1	0.97	17109	50	1901567	0	
54	F105	126.987	39.78	38.66	42.57	423	9510196	1	0.96	1644	8	0	0	
55	F106	127.0011	212.52	209.77	215.24	424	118170205	1	1	112	69	98217991	4390804	



Goals:

1. Identify regions of interest in acquisitions (feature)
2. Provide an estimate of that feature's abundance per sample

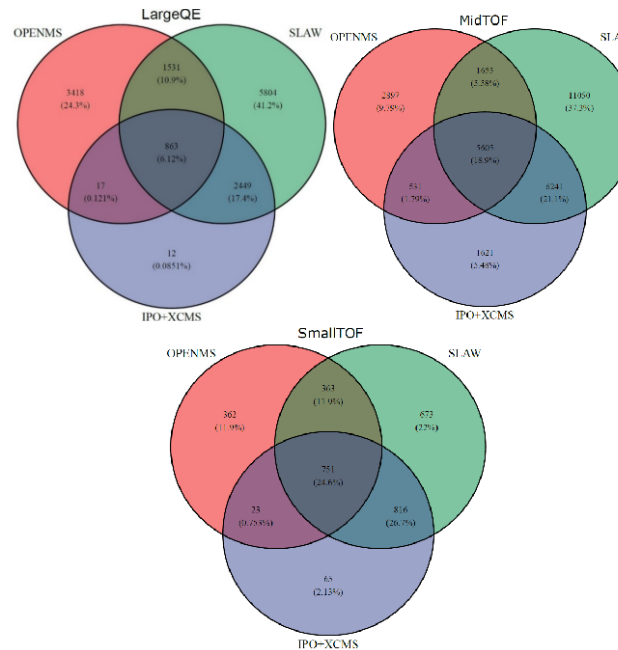
Existing tools are:

1. Computationally expensive
2. Produce inconsistent features
3. Poorly align retention time and mass across samples

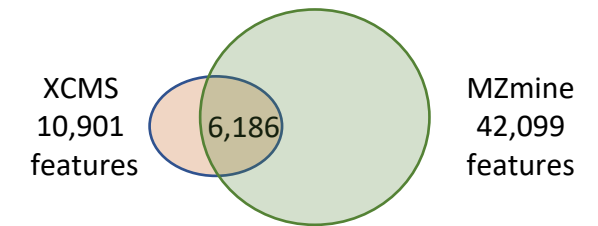
Reproducibility issues in LC-MS metabolomics data processing

andom sample of 400 peaks, with replacement,
d visually inspected. We scrutinized each peak
t the random sample of peaks into three categ
ssibly to real compounds, the second category i
ounds, and the third is peaks that cannot easily
e on we refer to these three categories as good,
a peak was considered good if it met the follow
the peak appear to encapsulate the majority of t

Myers et al. (2017).
Analytical Chemistry,
89(17): 8689



Delabriere et al., (2021).
Analytical Chemistry,
93(45):15024



Li et al., (2023)
Nature Communications,
14(1), 4113.



Top Four Reasons for Poor Reproducibility

- High rate of correspondence errors in large data
- Spurious number of low-quality peaks,
confusion of sensitivity
- Peak detection not transparent enough
- Too many parameters, too dependent on local expertise

Article | [Open access](#) | [Published: 11 July 2023](#)

Trackable and scalable LC-MS metabolomics data processing using asari

[Shuzhao Li](#) , [Amnah Siddiqi](#), [Maheshwor Thapa](#), [Yuanye Chi](#) & [Shujian Zheng](#)

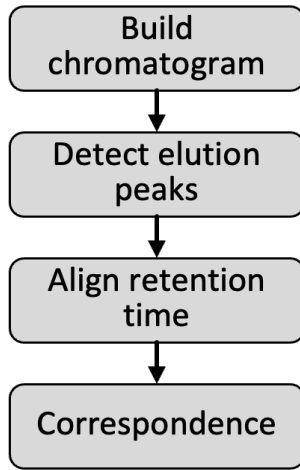
[Nature Communications](#) **14**, Article number: 4113 (2023) | [Cite this article](#)

- Detailed account of technical issues
- Trackable data structure at every level in asari
- New algorithms, new build
- A useful set of quality control metrics
- Performant, easy to deploy, easy to scale

Asari vs. Conventional

A

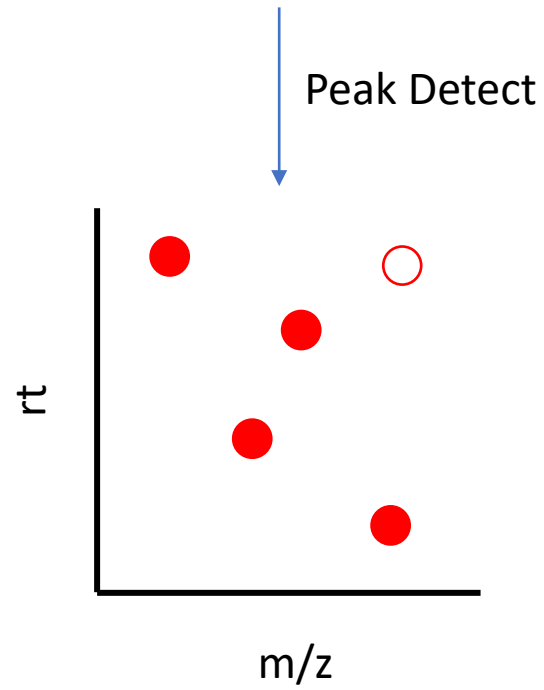
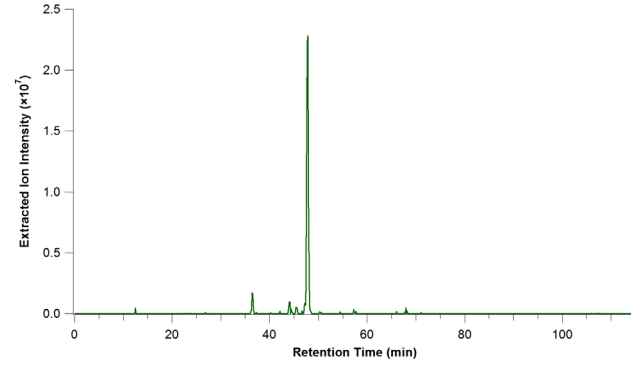
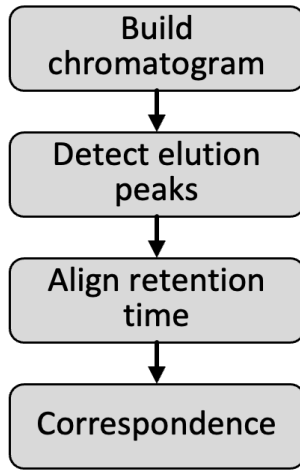
Conventional



Asari vs. Conventional

A

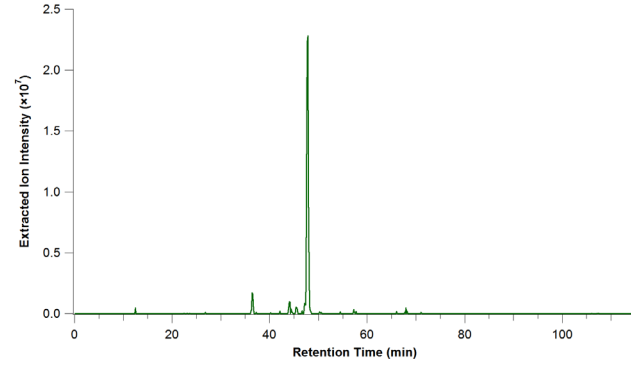
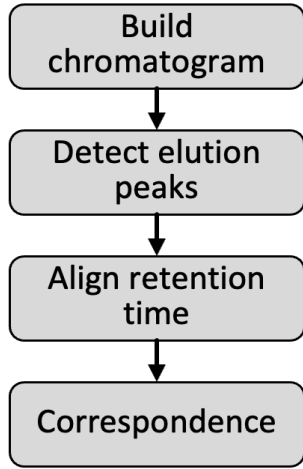
Conventional



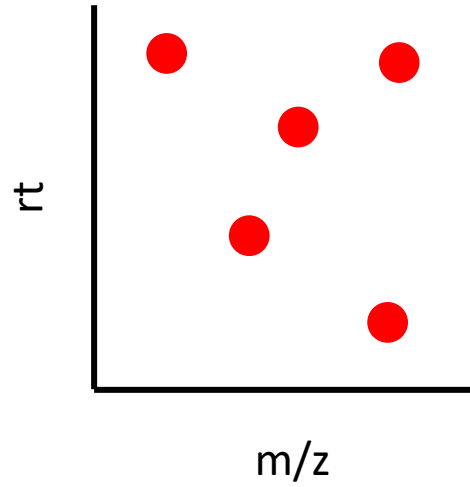
Asari vs. Conventional

A

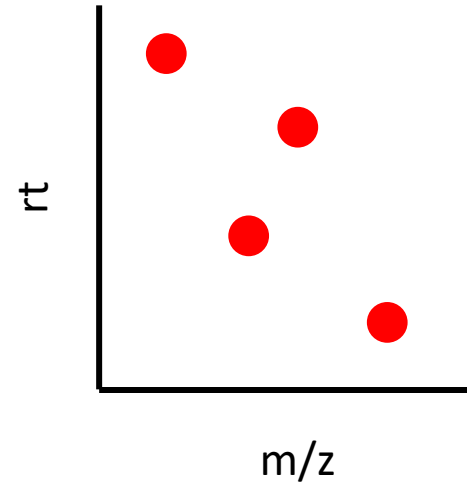
Conventional



Peak Detect



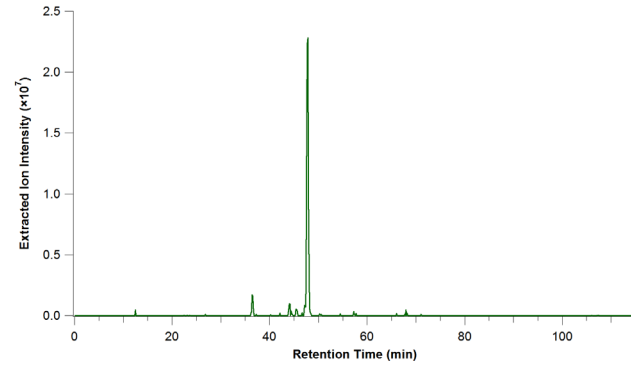
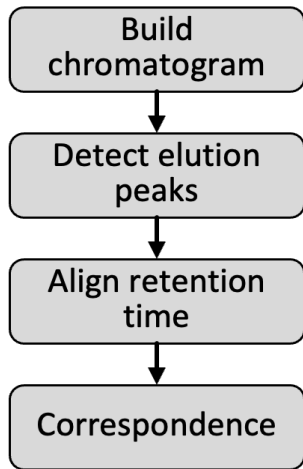
Alignment



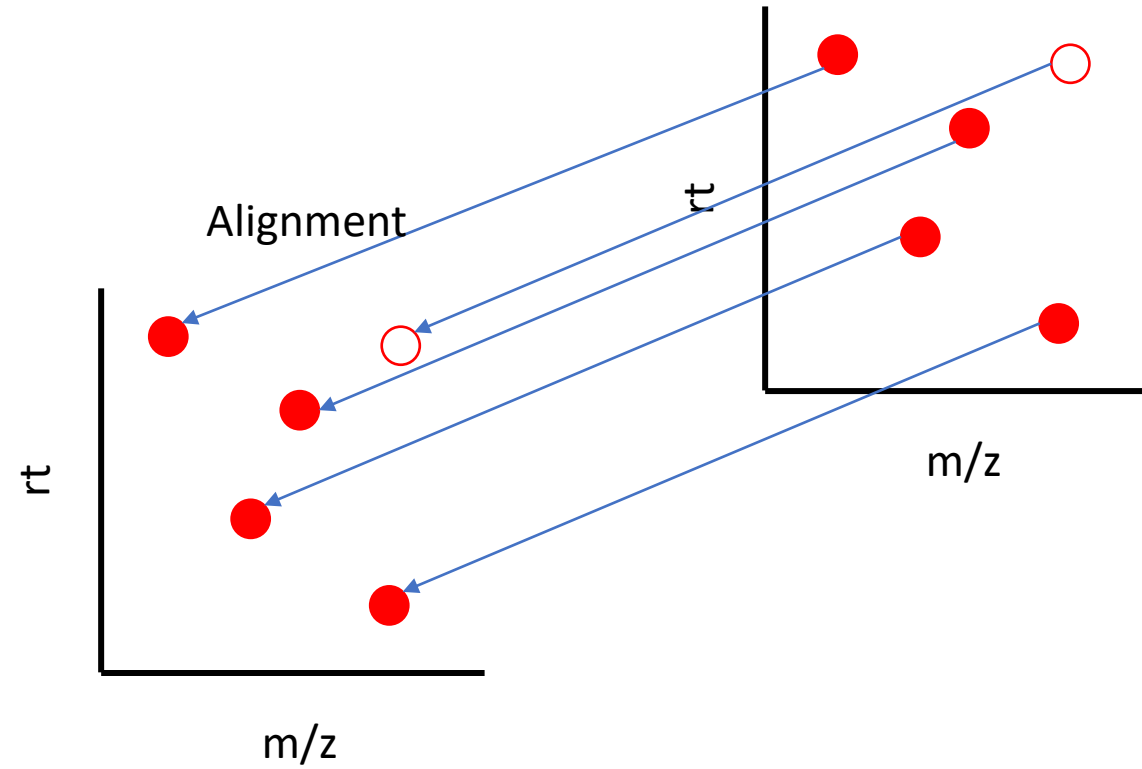
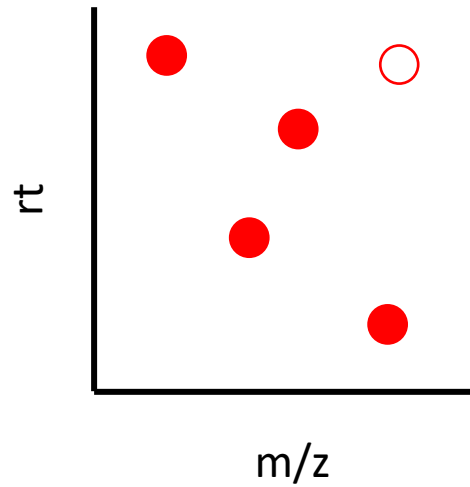
Asari vs. Conventional

A

Conventional

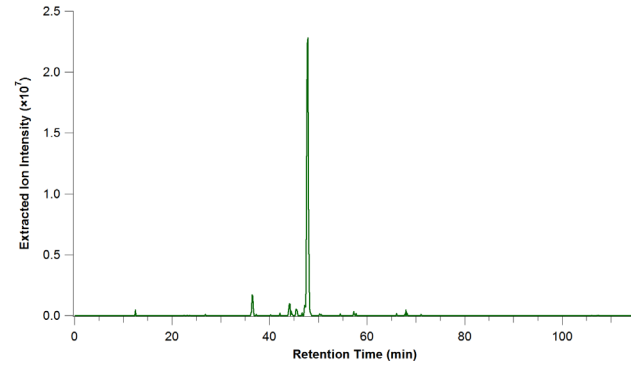


Peak Detect

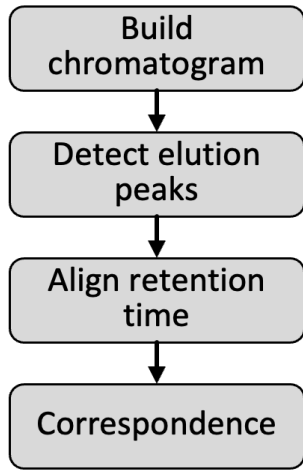


Asari vs. Conventional

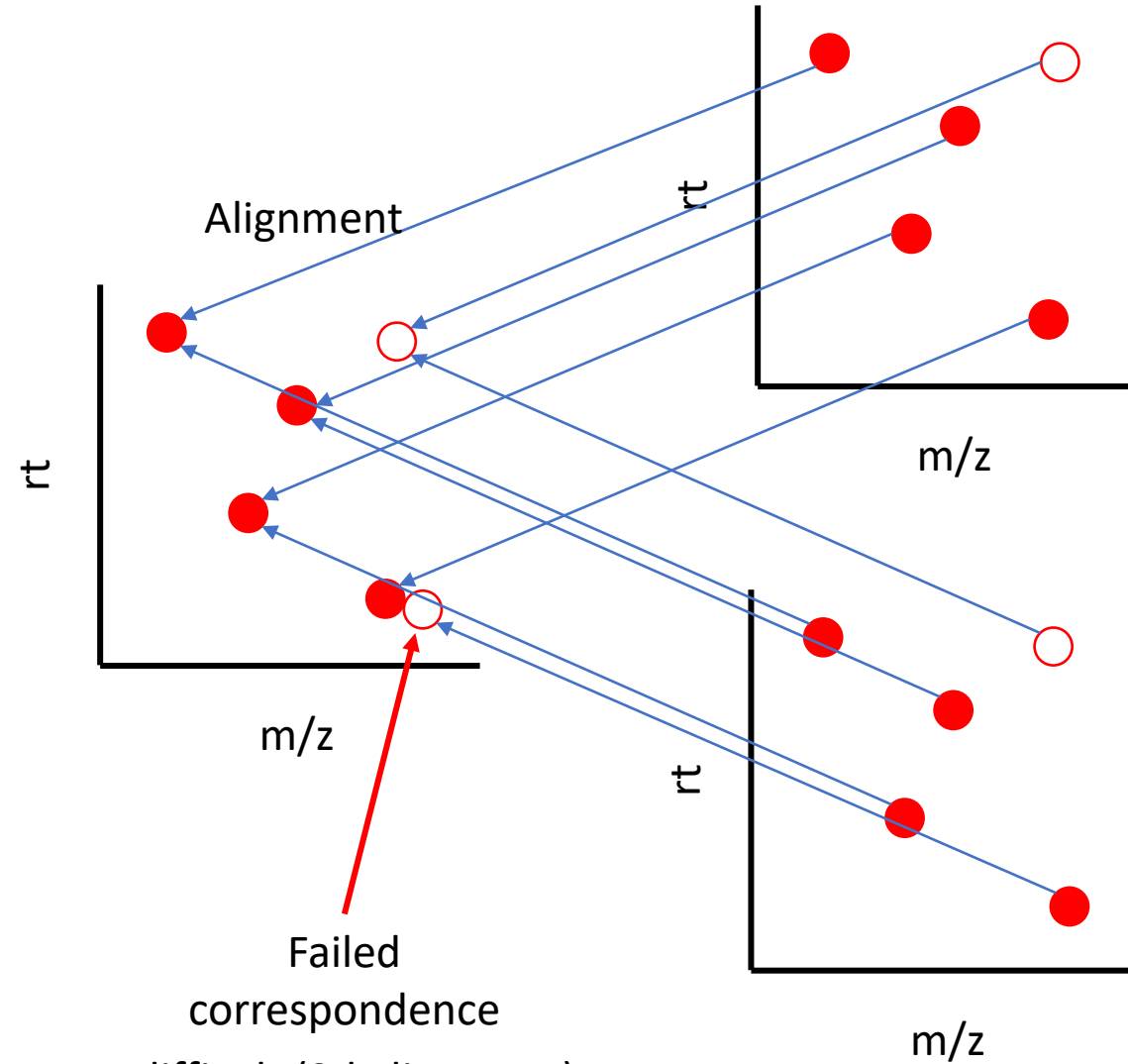
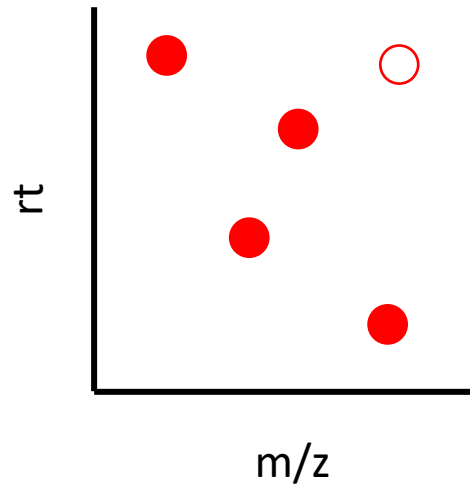
A



Conventional

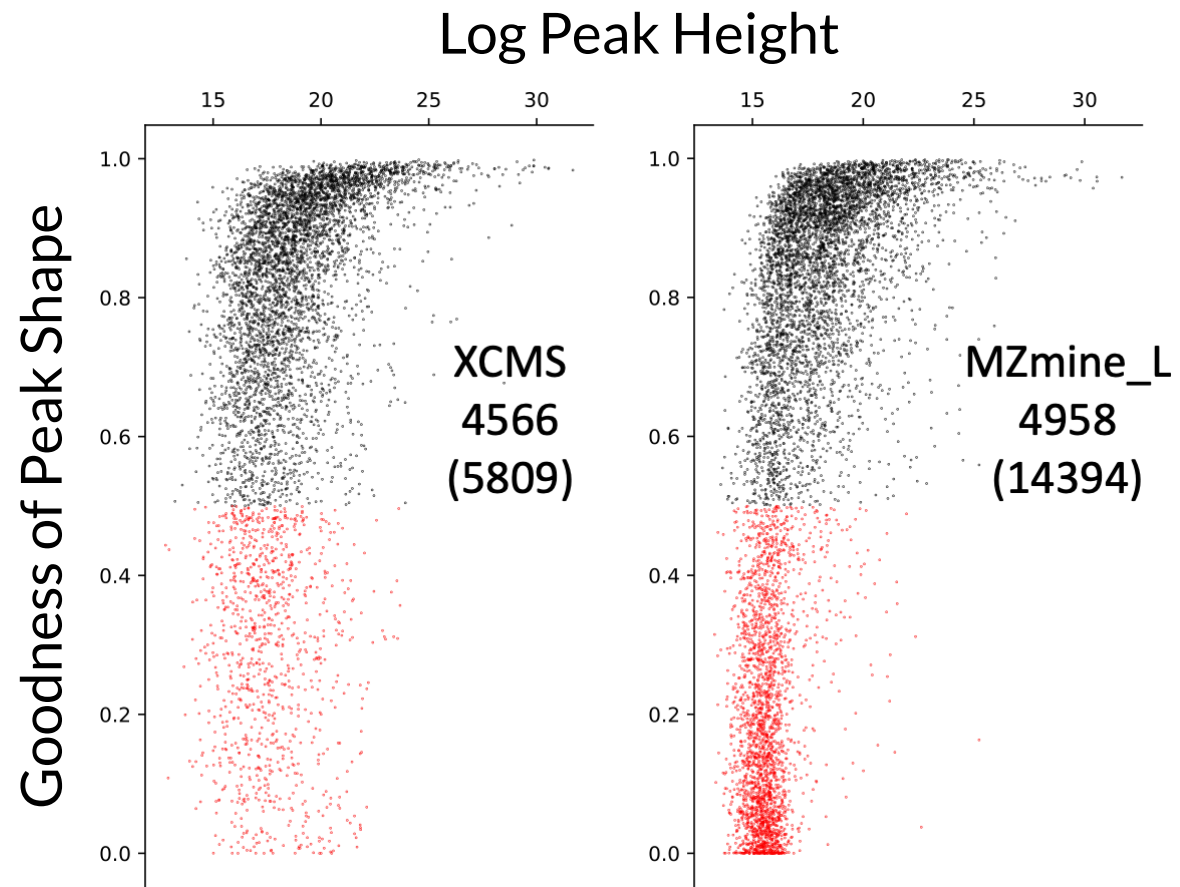


Peak Detect



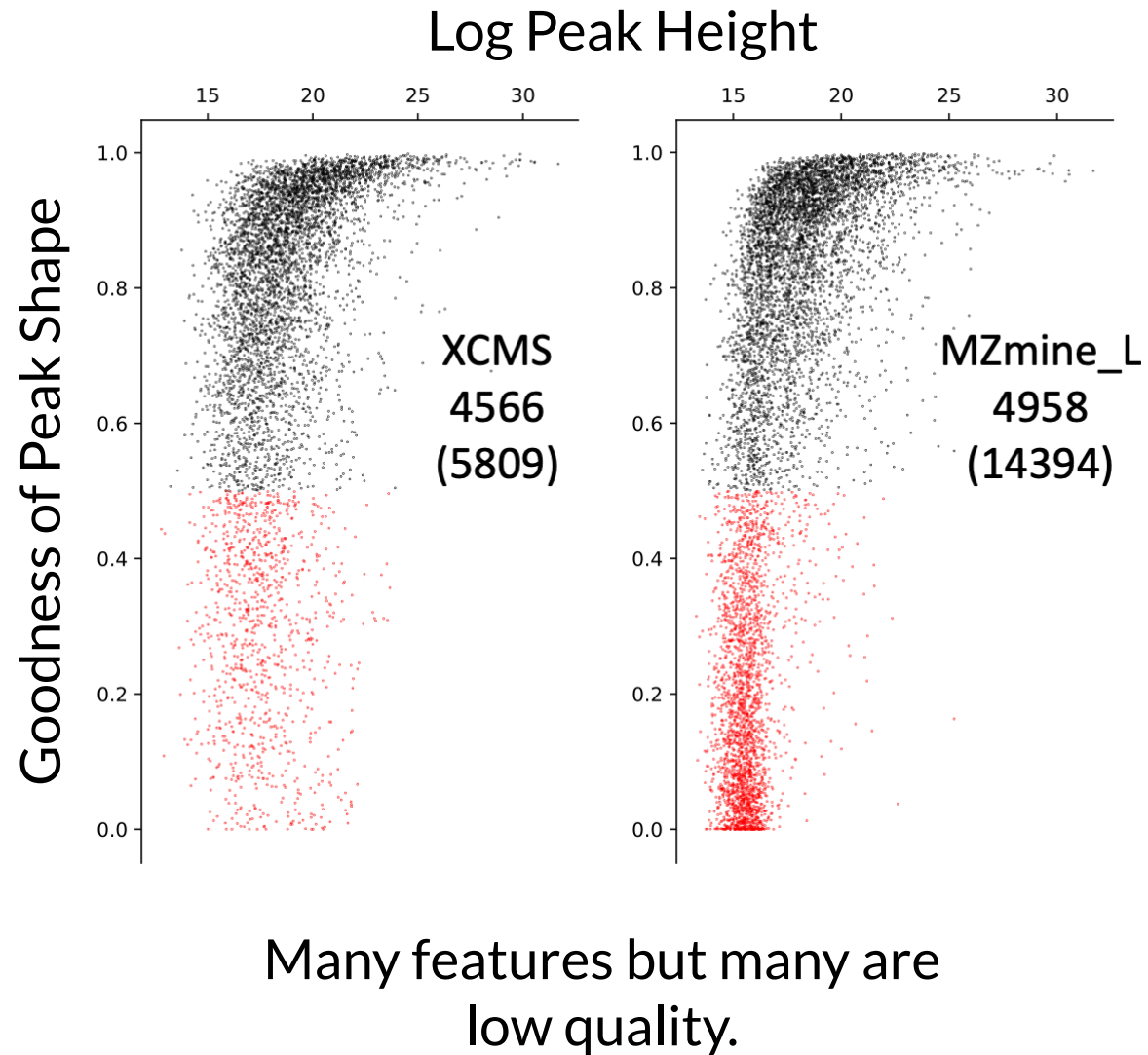
Peak picking then alignment makes alignment more difficult (2d alignment).
Fails to fully leverage the high m/z resolution of modern LC-MS instruments

Many Low-Quality Peaks Due to Low Mass Selectivity



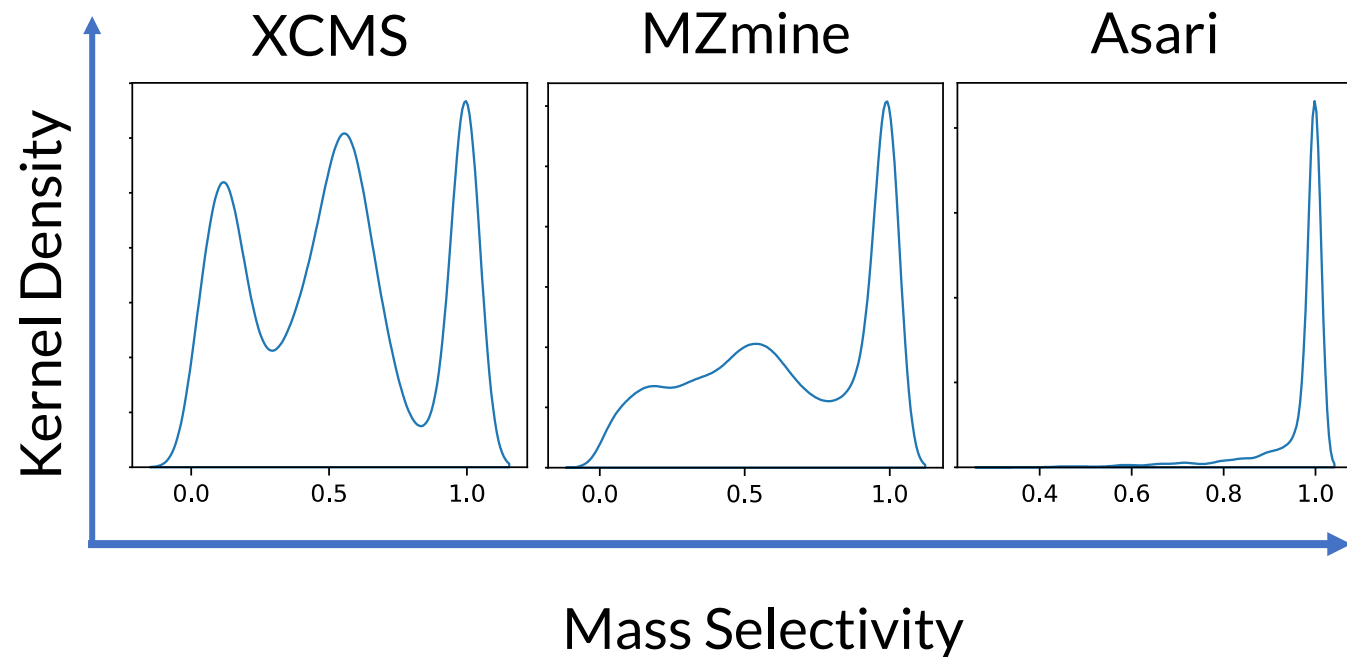
Many features but many are
low quality.

Many Low-Quality Peaks Due to Low Mass Selectivity



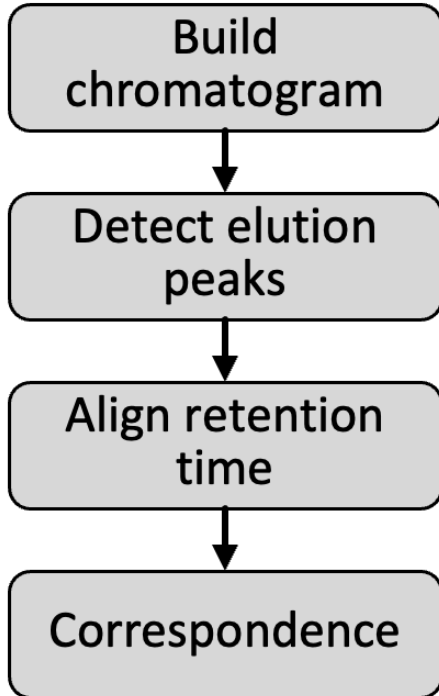
XCMS		MZmine	
m/z	rtime	m/z	rtime
760.5857	73	760.5807	68
760.5833	73	760.5817	72
760.5815	67	760.5826	72
		760.5837	72
		760.5847	73

Falsely resolved
isobaric features due
to poor mass
alignment

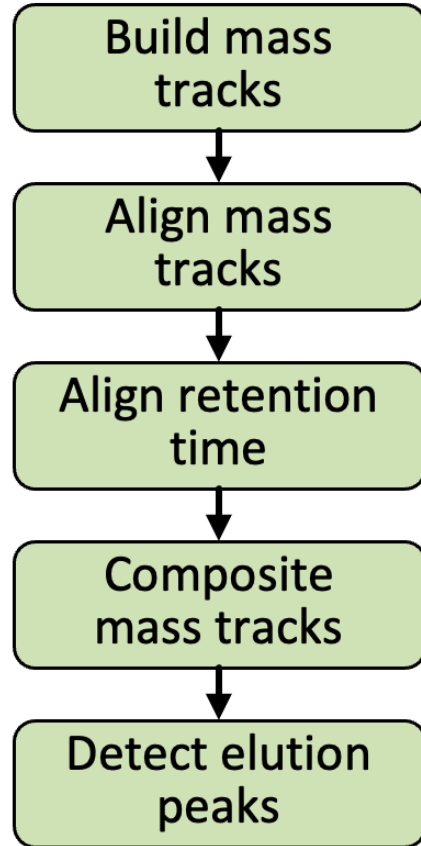


Asari Leverages High Resolution Mass to Improve Data Quality

Conventional

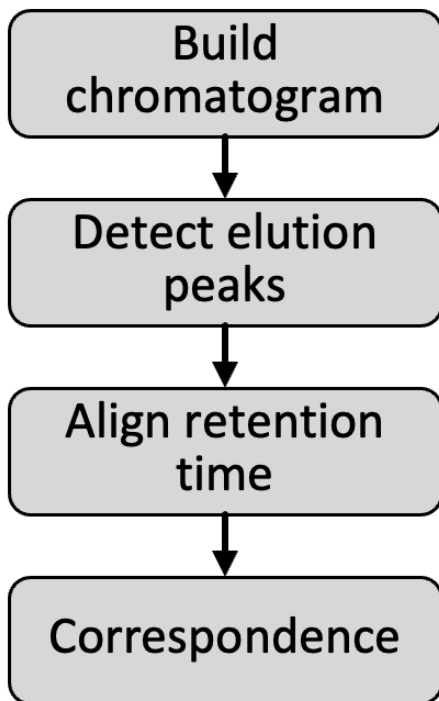


Asari

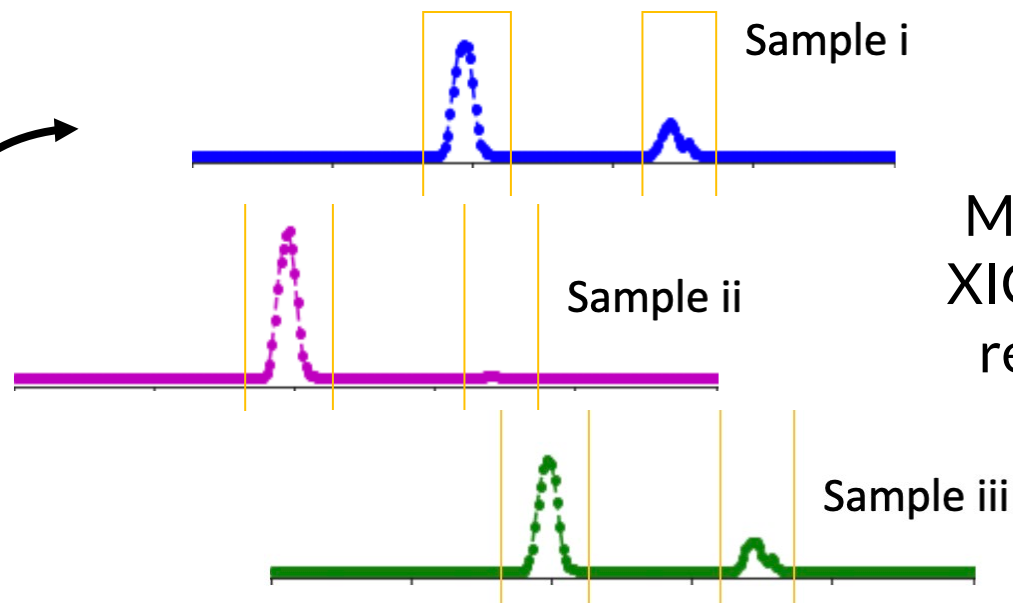
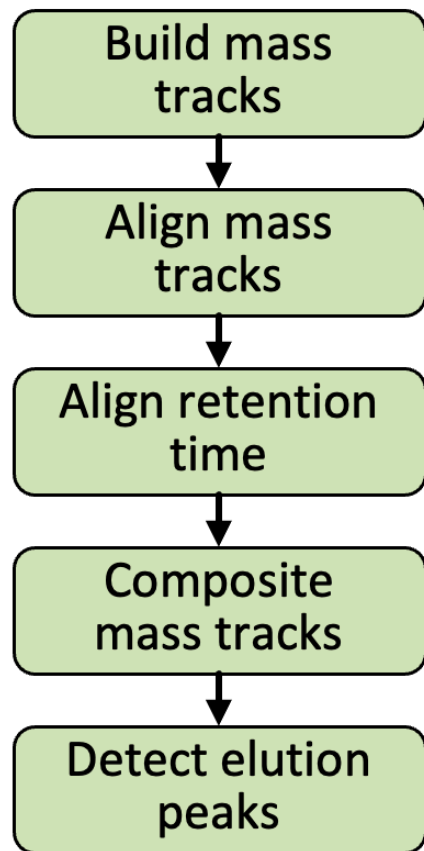


Asari Leverages High Resolution Mass to Improve Data Quality

Conventional



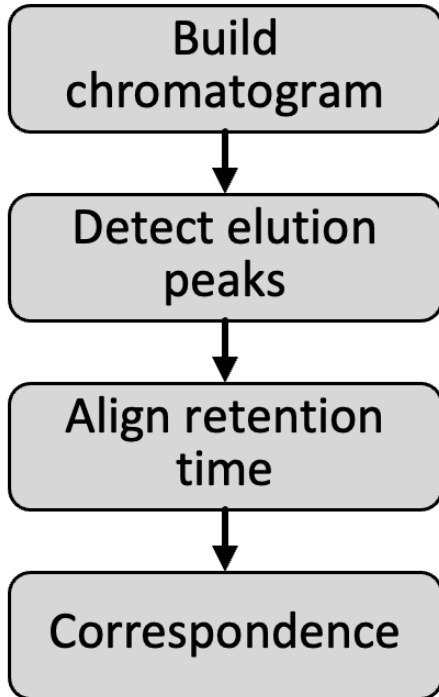
Asari



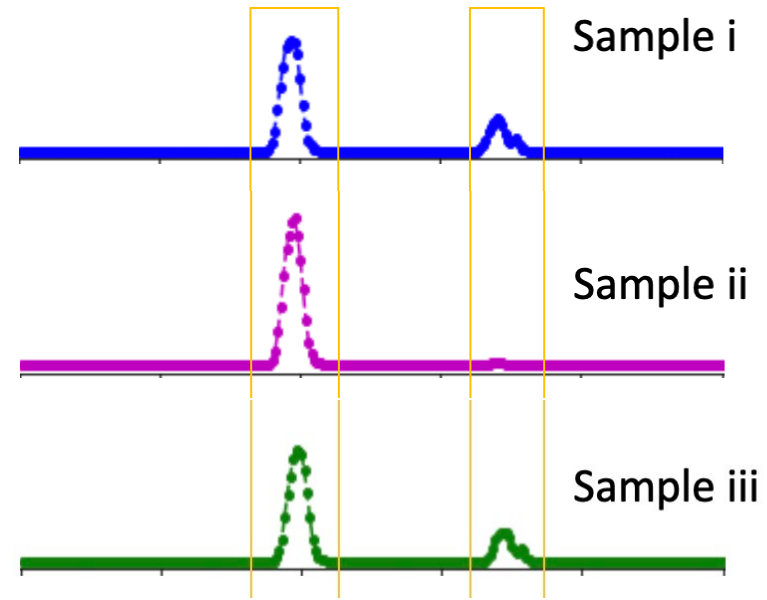
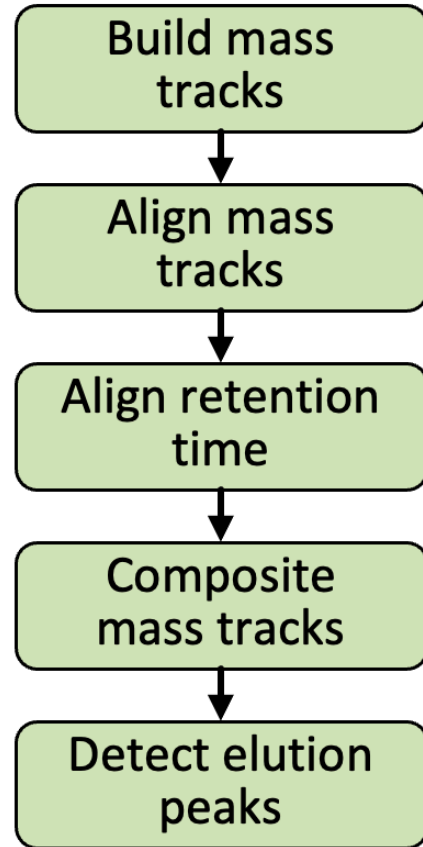
Mass tracks are XICs spanning all retention time

Asari Leverages High Resolution Mass to Improve Data Quality

Conventional

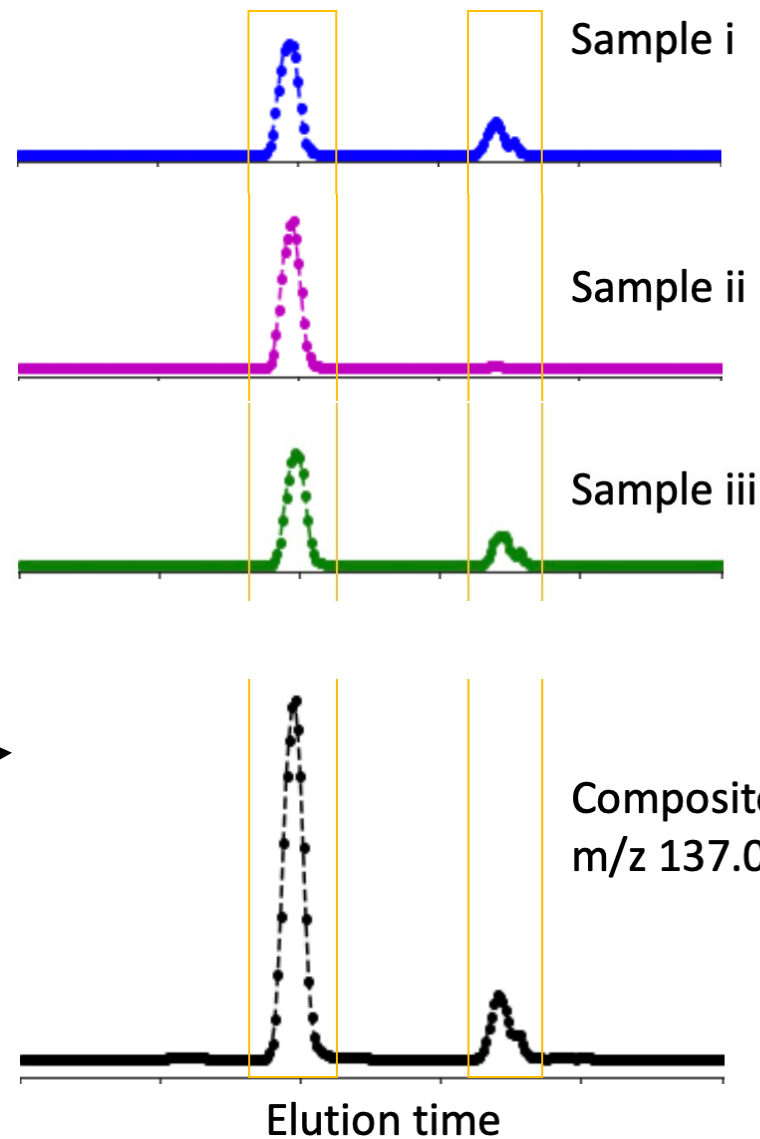
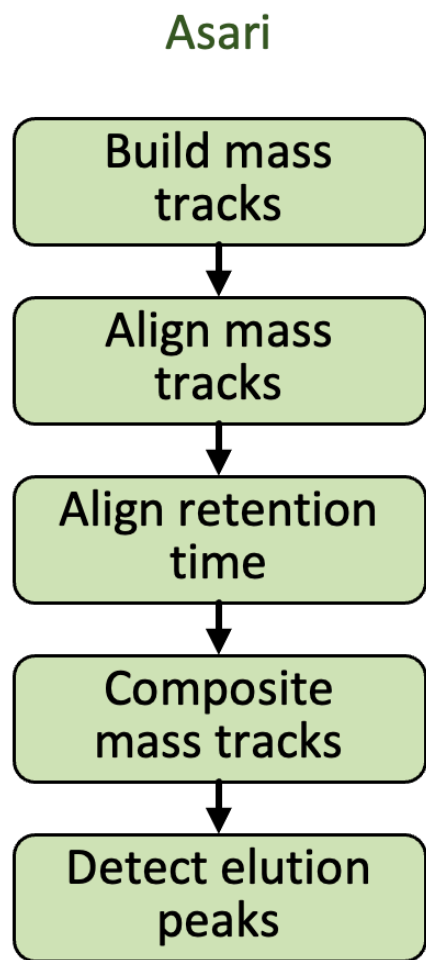
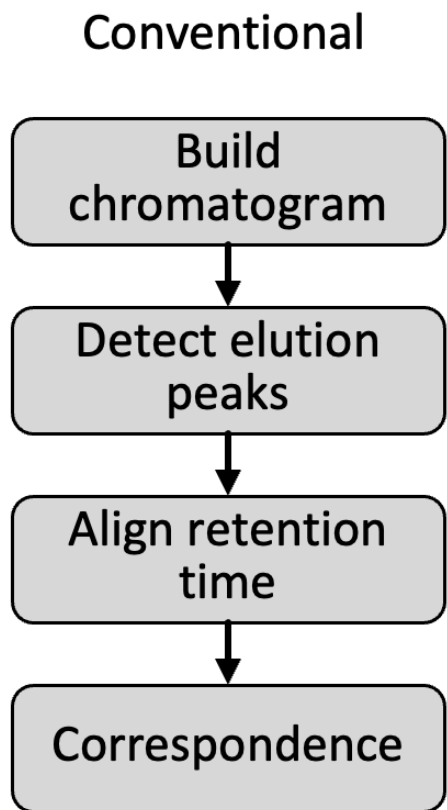


Asari



Mass tracks are XICs spanning all retention time

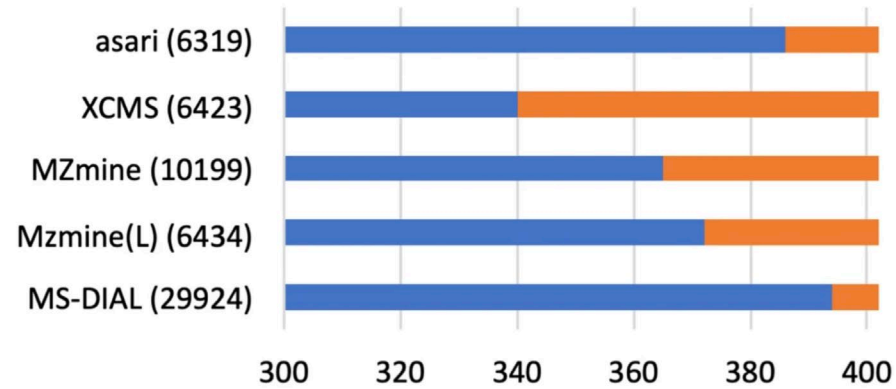
Asari Leverages High Resolution Mass to Improve Data Quality



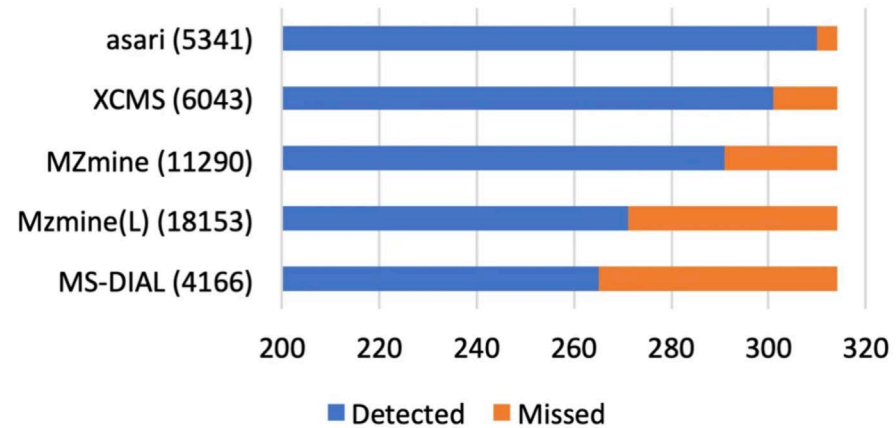
Mass tracks are XICs spanning all retention time

Why Asari?

Detection in HZV029

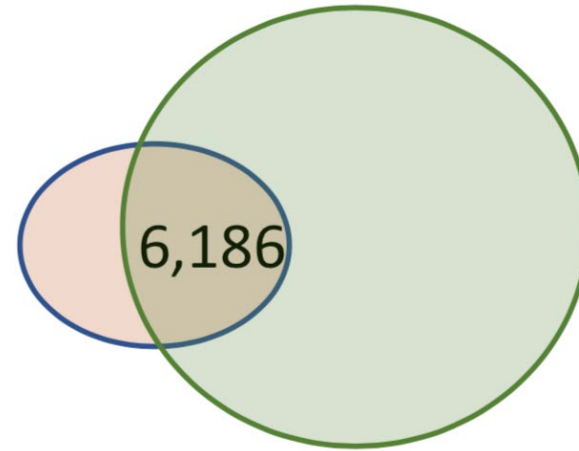


Detection in Yeast2021



Poor Consistency with Multiple Datasets

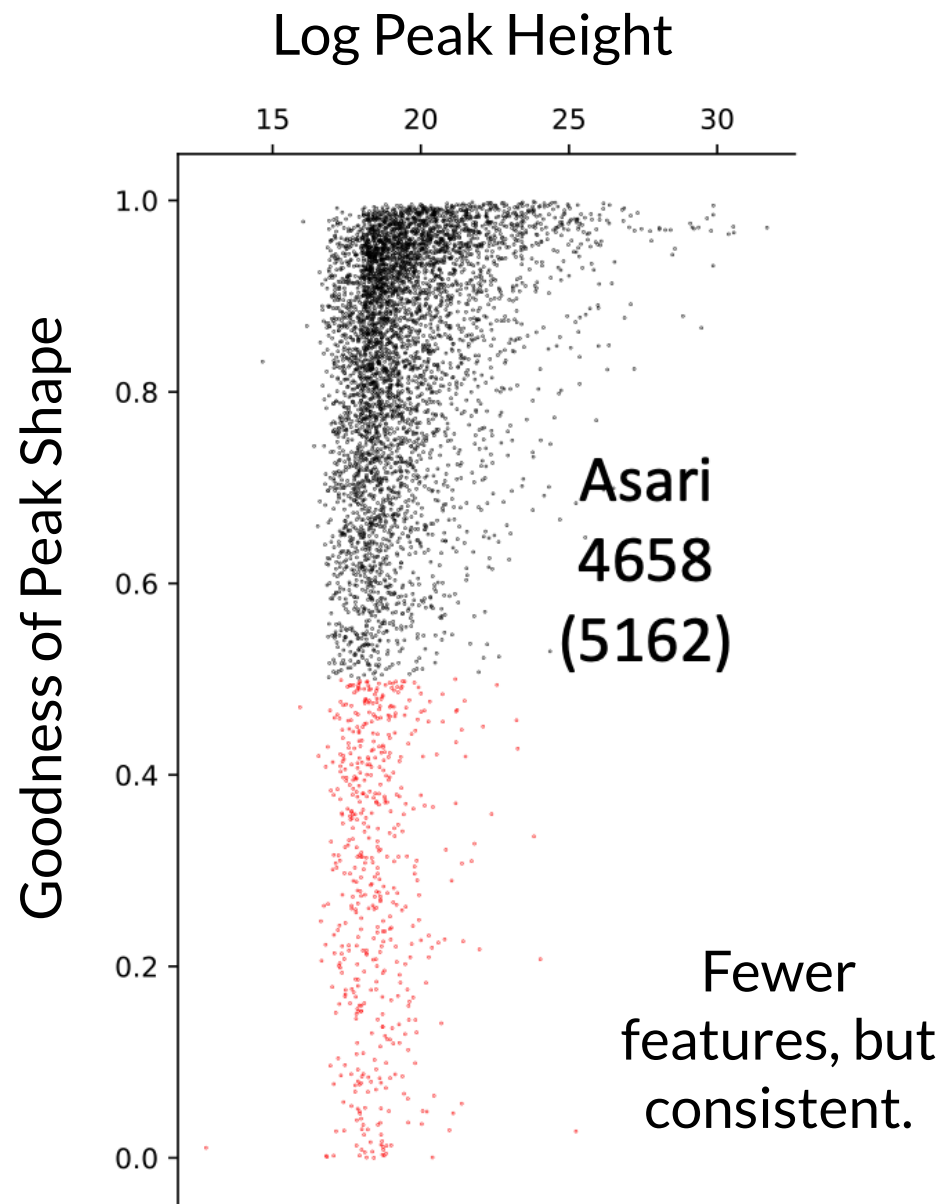
XCMS
10,901
features



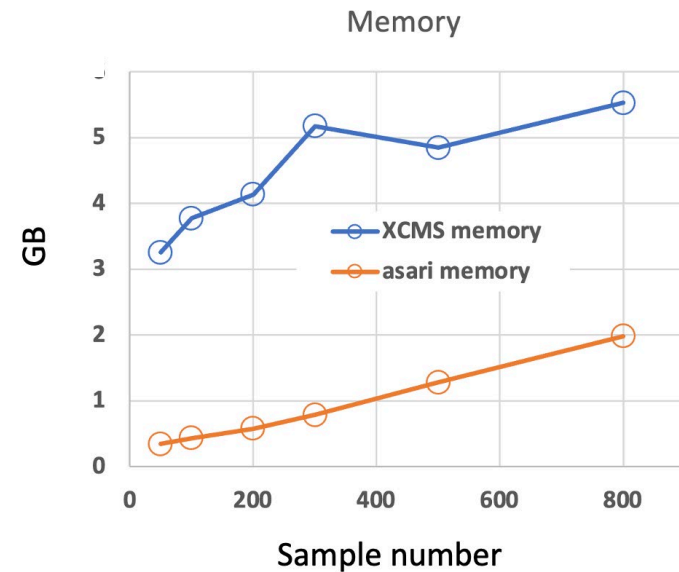
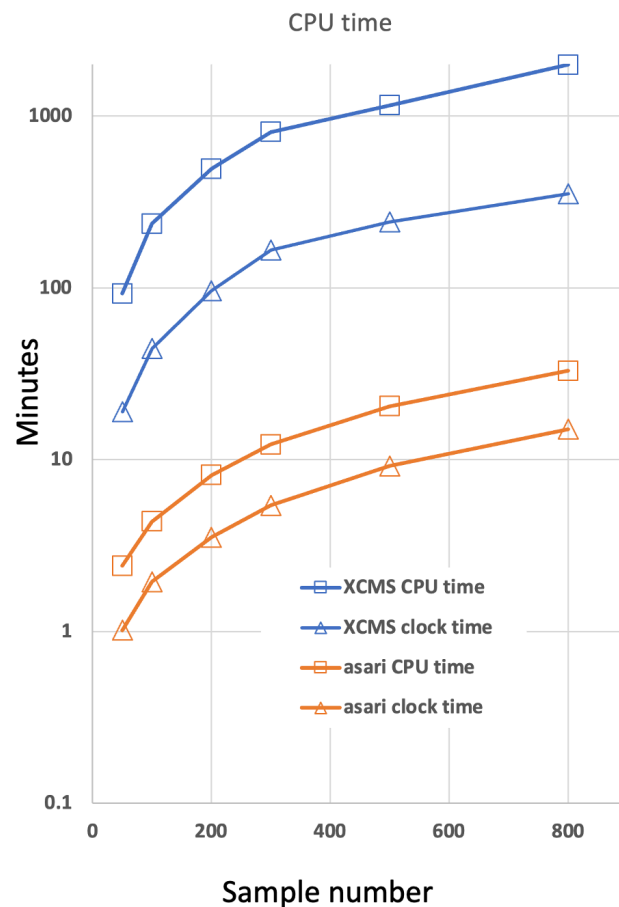
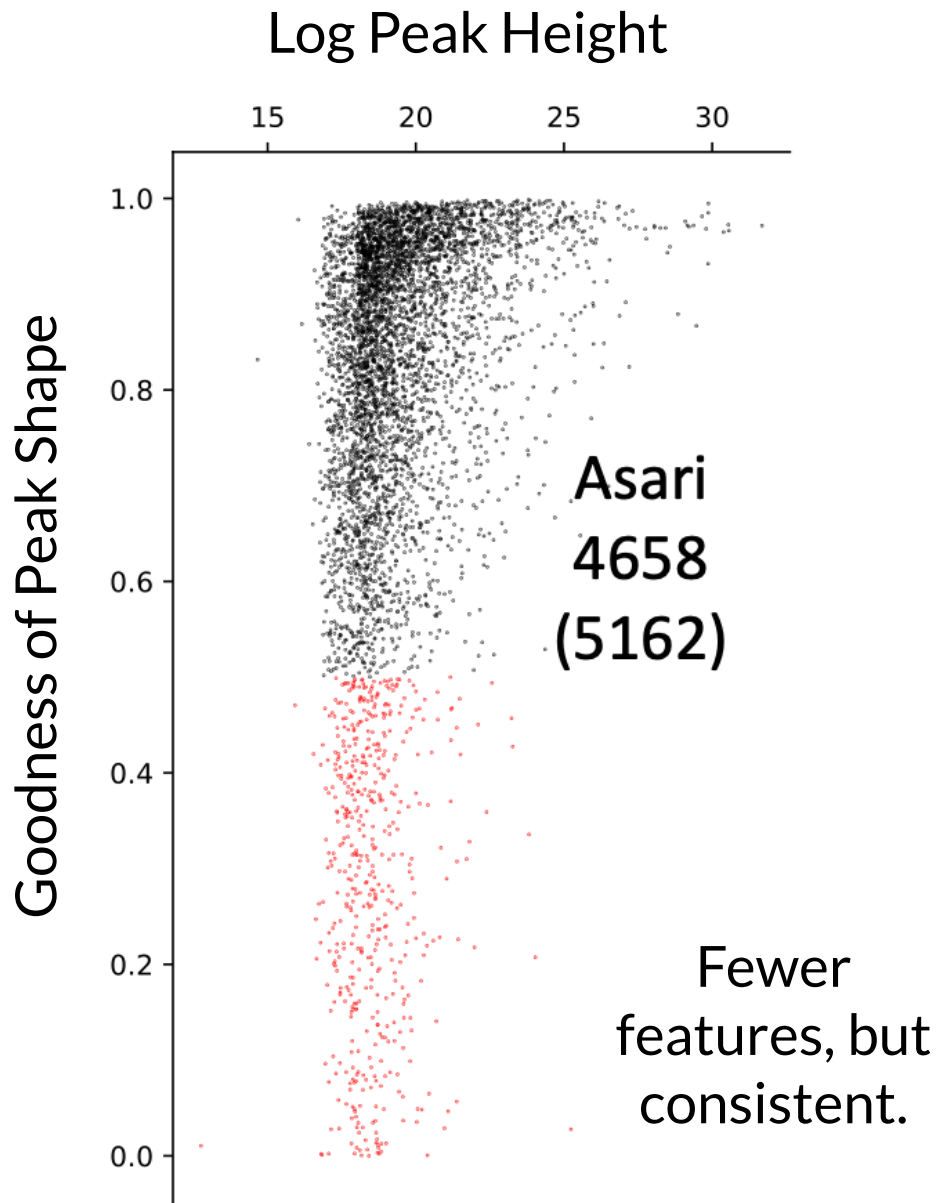
MZmine
42,099
features

Features from 184 Pooled
QC Samples

Asari Improves Feature Quality and Computational Performance



Asari Improves Feature Quality and Computational Performance



Order of magnitude improvement in runtime

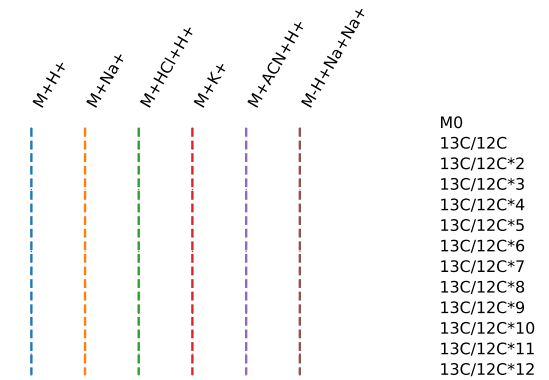


Beyond Feature Tables

Quality Control
Annotation
Reporting

Annotation

- khipu pre-annotation for both regular and isotope tracing data. Yields empirical compounds.
- JSON centric, chaining MS, MS/MS and authentic libraries

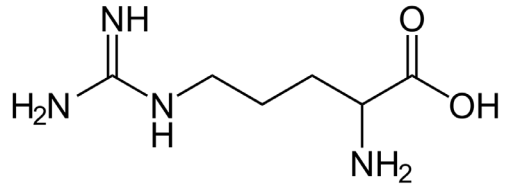


Li and Zheng, 2023.
Analytical Chemistry

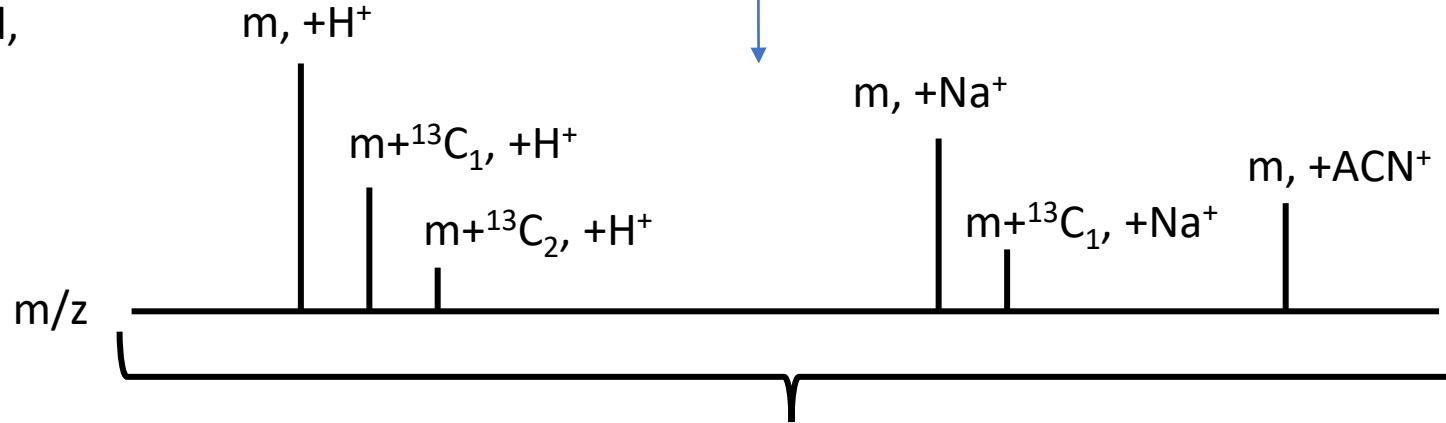


MS/MS search by MatchMS,
accelerated with IntervalTrees

What is an empirical compound?



One Compound,
Many Features



Traditional techniques,
each feature is queried for
annotation = 6 searches!
More false positives

Empirical Compound



Inferred Neutral Mass

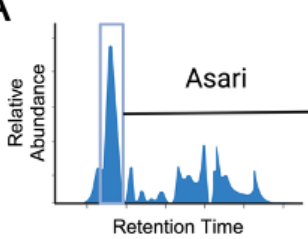


Empirical compound
means one query and
fewer false positives.

mzunit, CAMERA, binner perform a similar 'pre-annotation' but khipu performs the regression that allows for the inference of mass and yields a computable data structure.

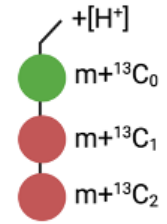
Annotation Levels

A

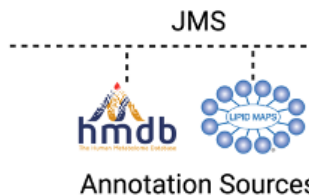


Feature Table		
Id, m/z,	rt, ...	
F3,	205.095, 52.32	
F343,	206.100, 52.34	
F344,	207.104, 52.33	

Khipu



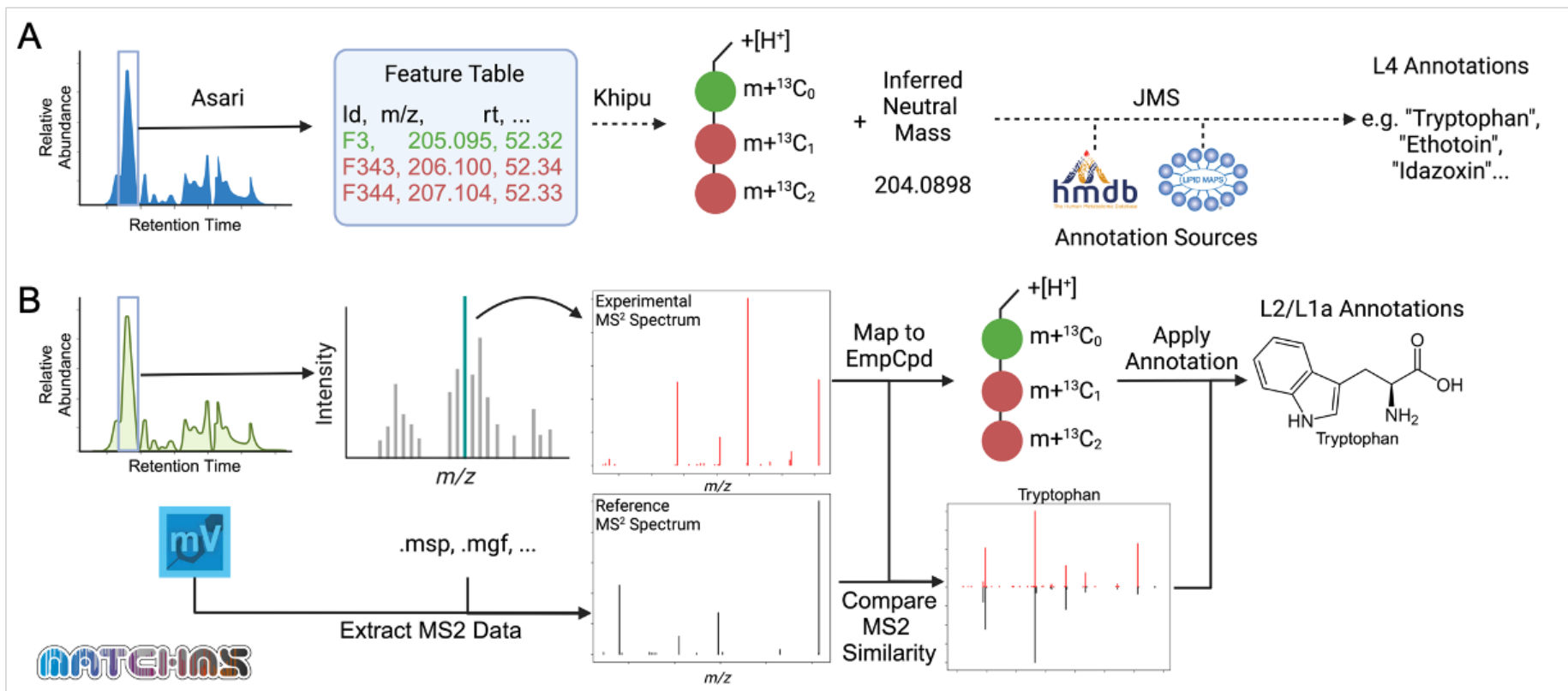
Inferred Neutral Mass
+
204.0898



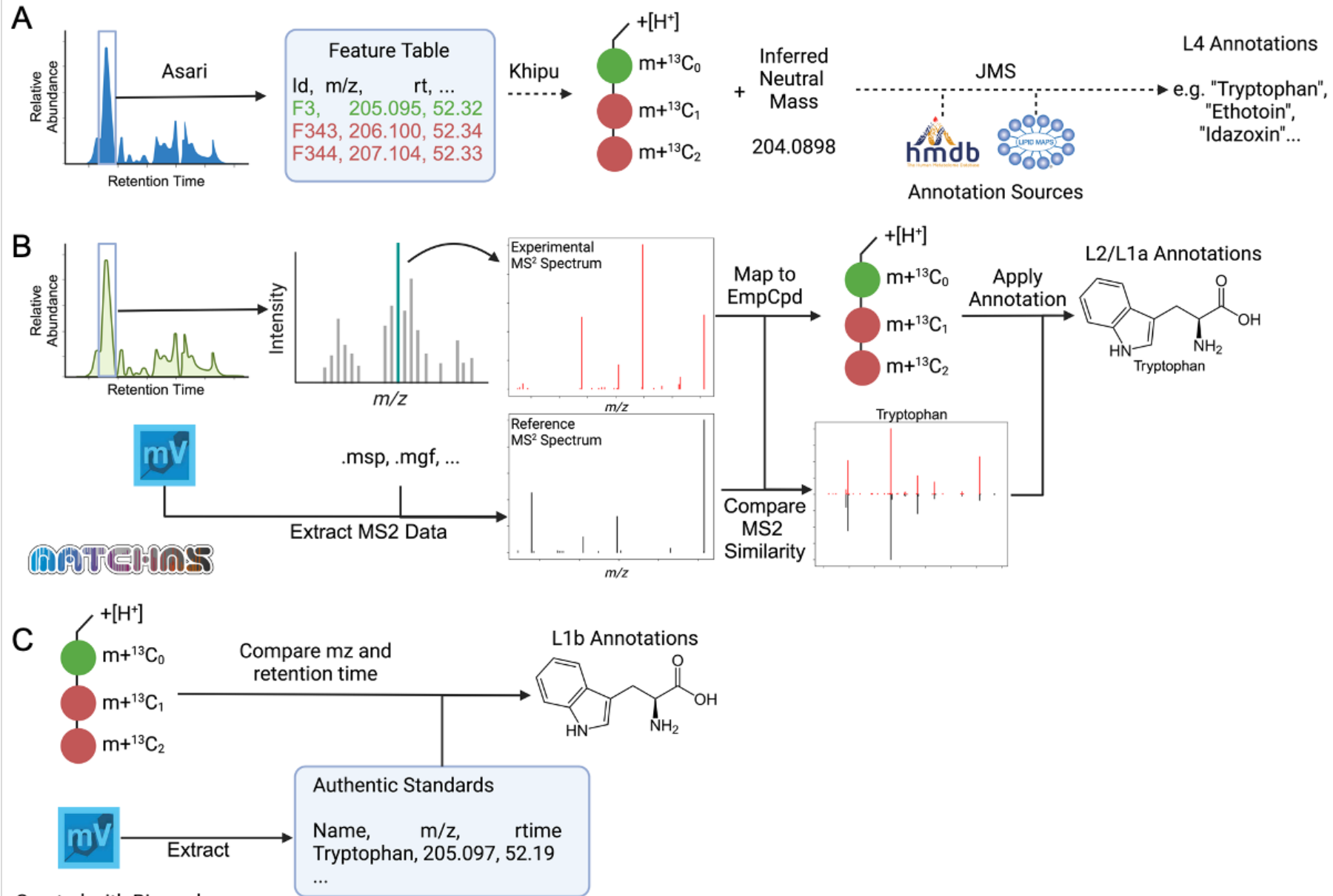
L4 Annotations

e.g. "Tryptophan",
"Ethotoin",
"Idazoxin"...

Annotation Levels



Annotation Levels

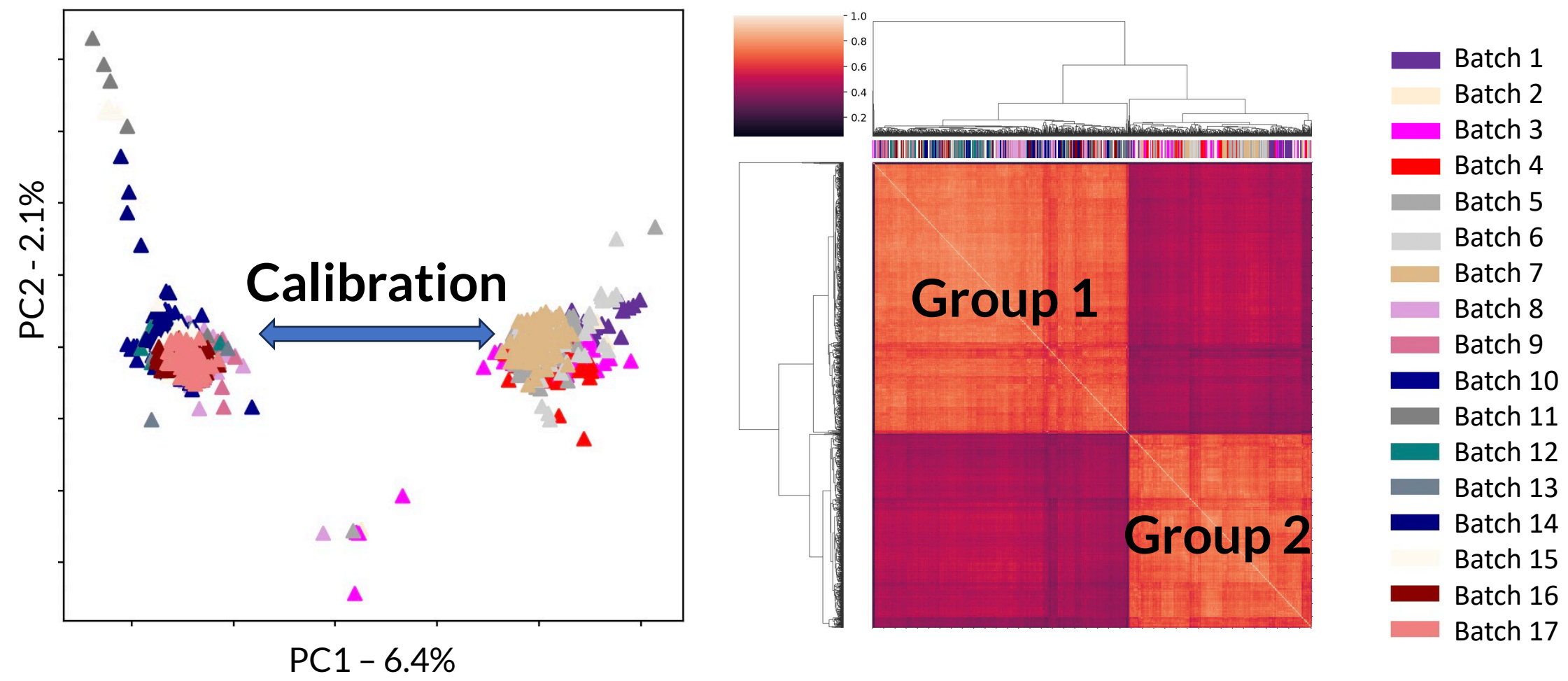


JSON Example of Chained Annotation

```
"kp4_166.0488": {
  "interim_id": "kp4_166.0488",
  "neutral_formula_mass": 166.04884896677,
  "Database_referred": [
    "MoNA-export-LC-MS-MS_Negative_Mode.msp", "HMDBv5"],
  "identity": [{"3-Methylxanthine", 0.98}],
  "MS1_pseudo_Spectra": [
    {
      "id_number": "F21",
      "mz": 165.0418,
      "rtime": 73.72,
      "isotope": "M0",
      "modification": "M-H-",
      "ion_relation": "M0,M-H-",
    },...],
  "MS2_Spectra": [
    {"ms_level": 2,
      "precursor_ion_mz": 165.041976928711,
      "list_mz": [ 55.02987289428711, ...],
      "list_intensity": [0.010449324526009295,...],
      "rtime": 69.007309953,
      "precursor_ion_id": "165.041976928711_69.007309953_plasma_ID_01.mzML",
      ...],
    "list_matches": [
      ["C6H6N4O2_166.049075",...],...
    ],
  "Level_4": [{
    "accession": "HMDB0001886",
    "name": "3-Methylxanthine",
    "chemical_formula": "C6H6N4O2",
    "primary_db": "HMDBv5"
  },...],
}
```

Made possible using empirical compounds

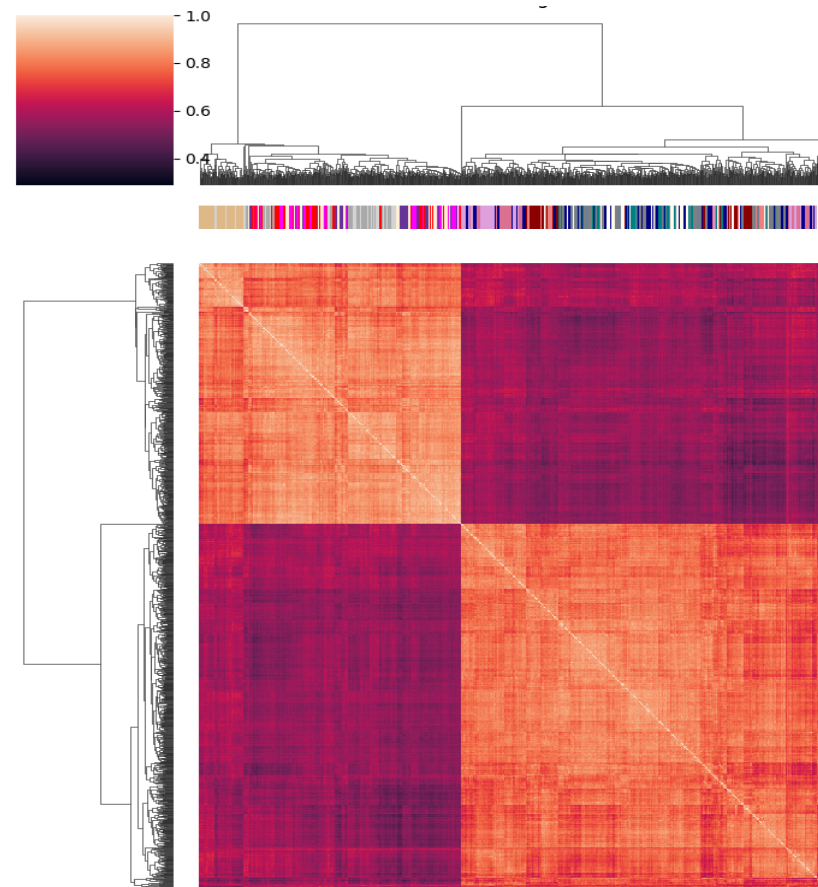
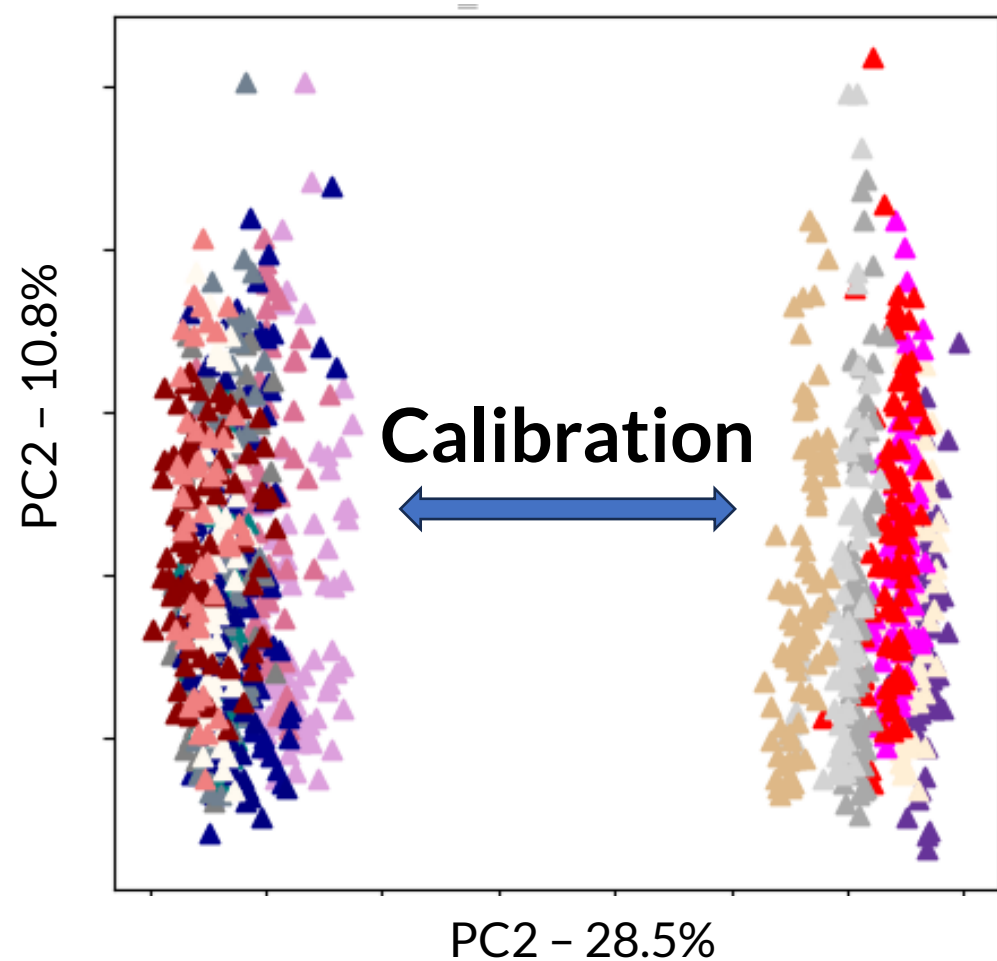
Large Experiment QA/QC Example



~1700 samples over 17 batches

Raw Data

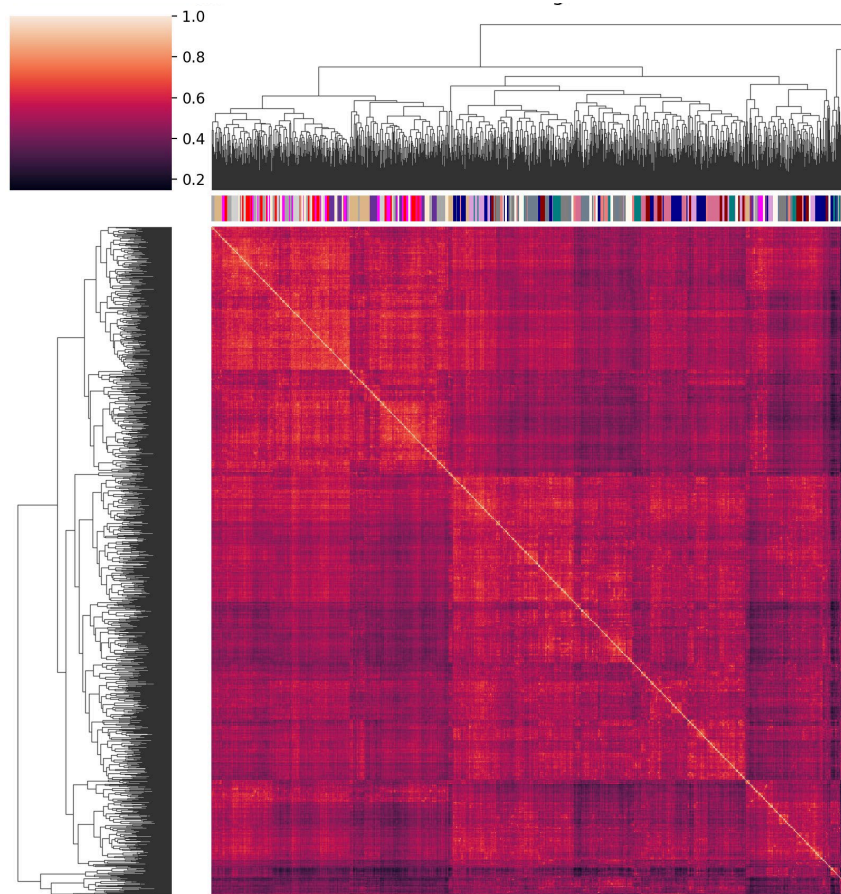
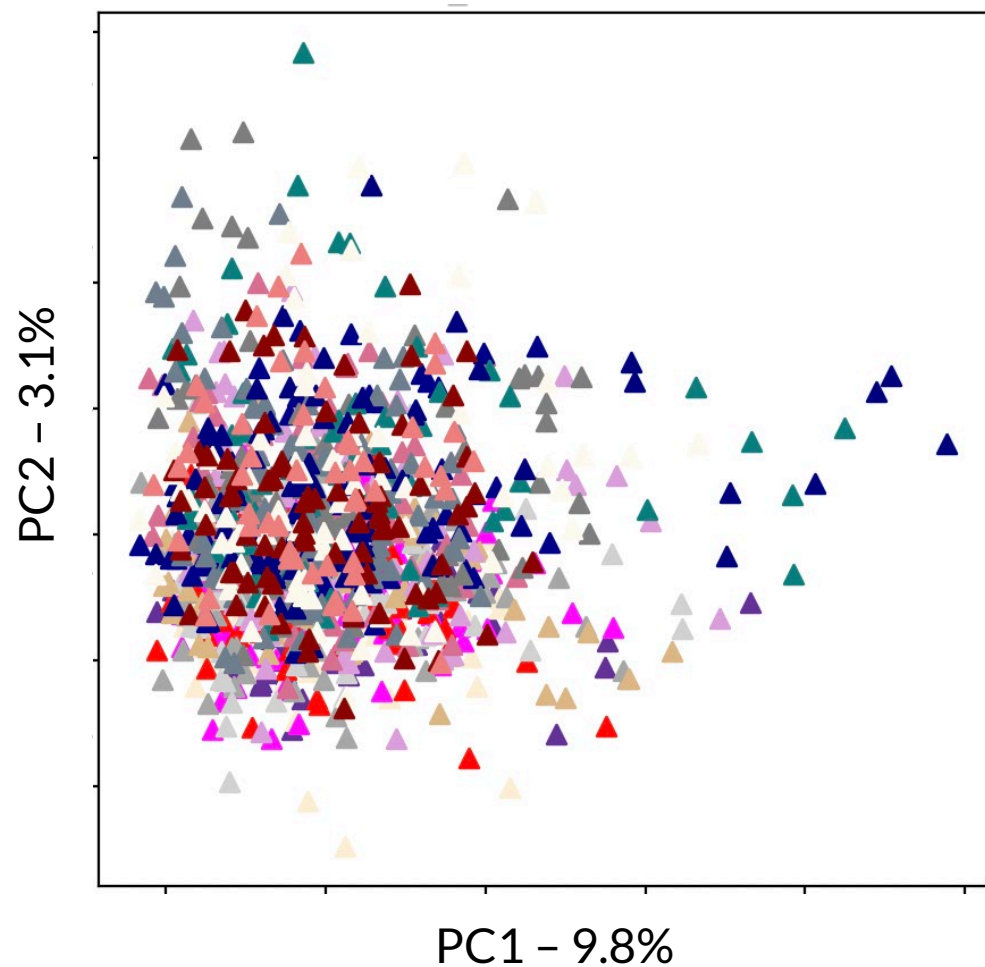
Large Experiment QA/QC Example



- Batch 1
- Batch 2
- Batch 3
- Batch 4
- Batch 5
- Batch 6
- Batch 7
- Batch 8
- Batch 9
- Batch 10
- Batch 11
- Batch 12
- Batch 13
- Batch 14
- Batch 15
- Batch 16
- Batch 17

Normalized, Interpolated

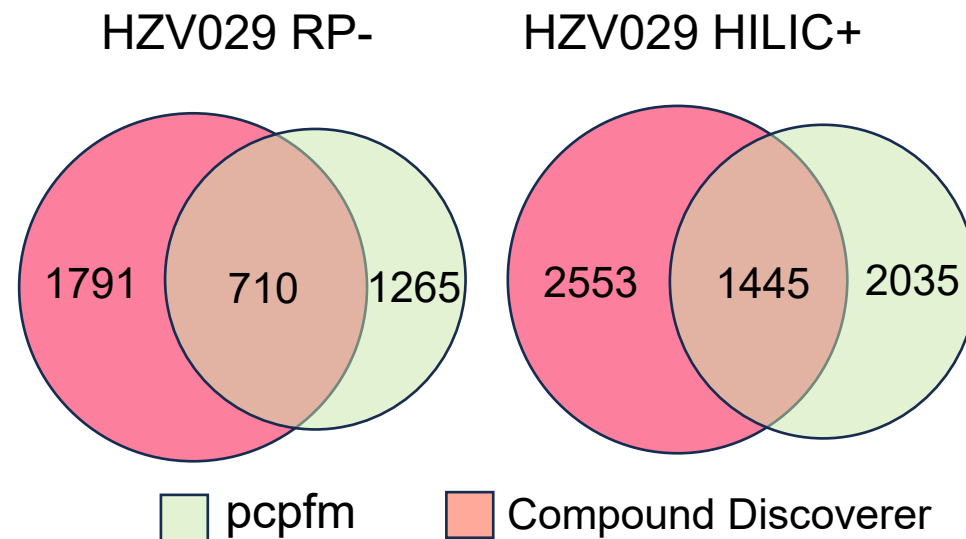
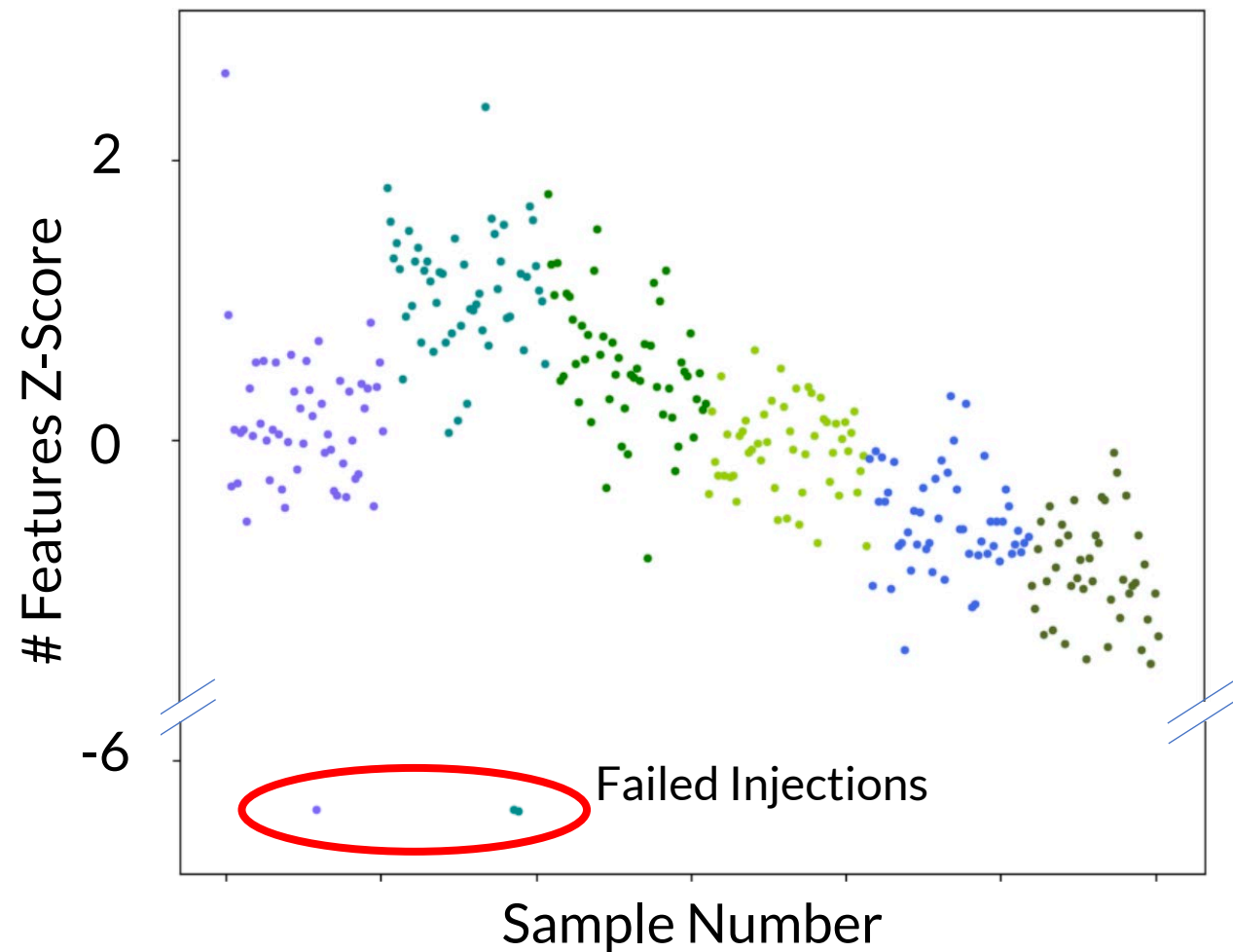
Large Experiment QA/QC Example



- Batch 1
- Batch 2
- Batch 3
- Batch 4
- Batch 5
- Batch 6
- Batch 7
- Batch 8
- Batch 9
- Batch 10
- Batch 11
- Batch 12
- Batch 13
- Batch 14
- Batch 15
- Batch 16
- Batch 17

Batch Corrected

Outlier and Annotation Example Results



High MS² annotation similarity
between CD and pcpfm

Example Reports

PCPFM Report - dmpa_pcpfm

Timestamp
Report generated on 2023-10-04 13:30:48.567659

Feature Table Summary

A feature denotes a region of a spectrum believed to represent an ion of a compound with a retention time and a mass-to-charge ratio. Multiple features often represent the same metabolite due to isotopologues, adduct, multiple charges etc. and thus, the number of features is only a rough proxy for the number of detected metabolites. Due to noise, artifacts, rare metabolites, etc. the number of features often increases with the number of samples.

Table Name, Num Samples, Num Features

full, 47, 238069
preferred, 47, 96228
preferred_blank_masked, 47, 96228
masked_preferred_unknows, 24, 81703
pref_qaqc_filtered_unknows, 22, 81336
pref_normalized, 22, 81336
pref_missing_dropped, 22, 67237
pref_interpolated, 22, 67237
log_transformed_for_analysis, 22, 67237

empCpd Table Summary

Empirical compounds are computational intermediates representing sets of features suspected to correspond to the same compound. Each empirical compound is a khipu, thus, the number of khipus is an estimate of the number of detected metabolites. Each khipu represents multiple features; however, unless singletons were added to the khipu during construction, the number of features grouped captured by empCpds is less than the number of features.

EmpCpd Name, Num Khipus, Num Features

asari, 155596, 238070
preferred, 21509, 58296
HMDB_LMSD_annotated_preferred, 21509, 58296

Annotation Summary

Annotations are mappings of features / empCpds to suspected chemical entities. Annotations can be higher or lower confidence depending on the origin in which they are generated. In general, MS1 annotated features are lower confidence than MS2 annotated features.

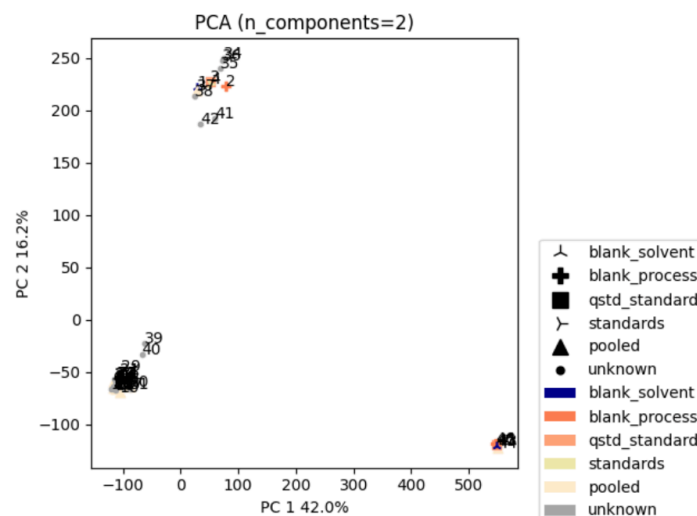
Feature Tables

Table Name, # Features, # MS1 Annotated Features, # MS2 Annotated Features

full, 238069, 0, 0
preferred, 96228, 0, 0
preferred_blank_masked, 96228, 0, 0
masked_preferred_unknows, 81703, 0, 0
pref_qaqc_filtered_unknows, 81336, 0, 0
pref_normalized, 81336, 0, 0
pref_missing_dropped, 67237, 0, 0
pref_interpolated, 67237, 0, 0
log_transformed_for_analysis, 67237, 0, 0

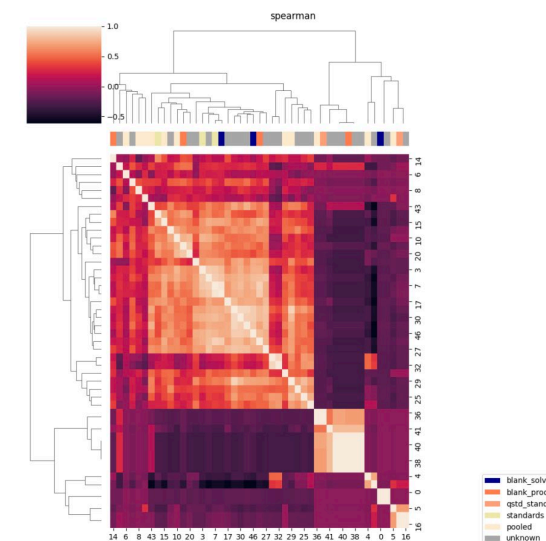
PCPFM Report - dmpa_pcpfm

Table: preferred

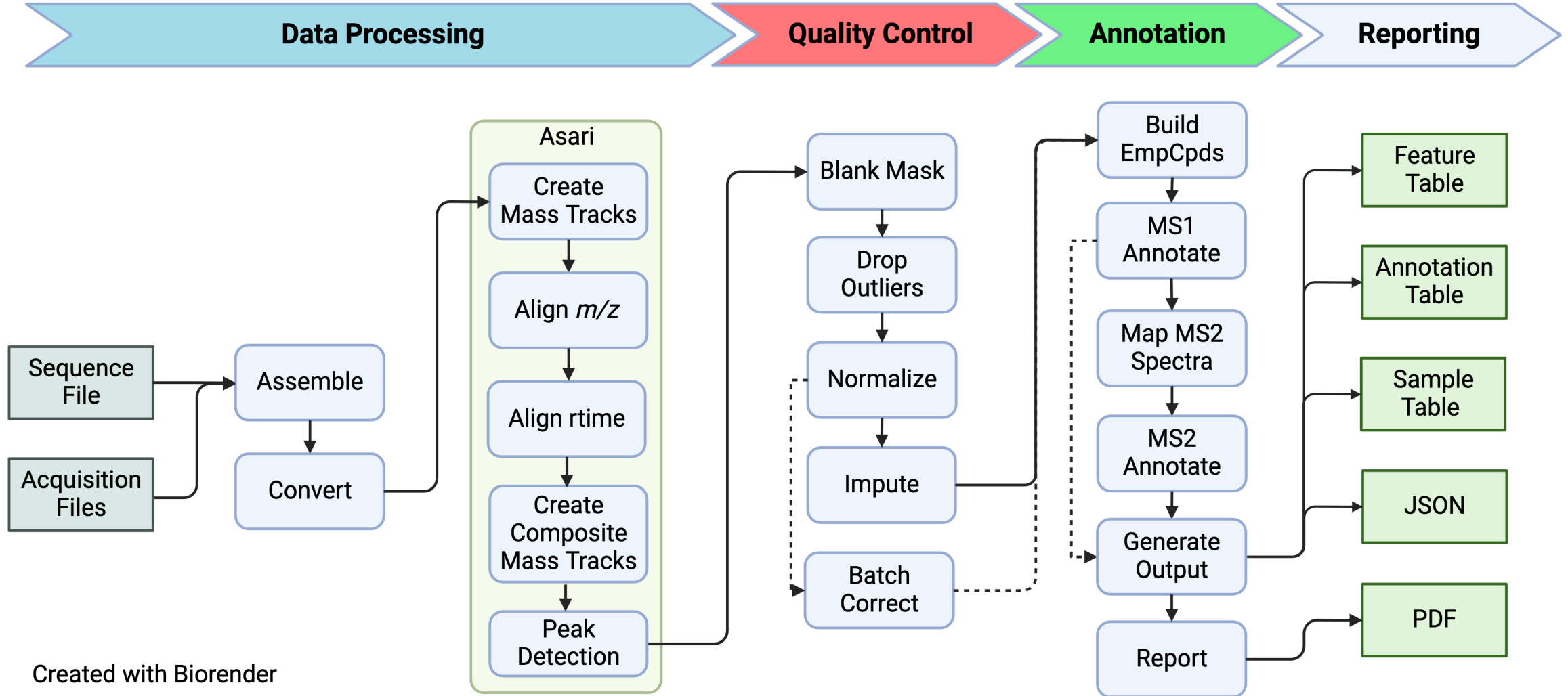


PCPFM Report - dmpa_pcpfm

Table: preferred

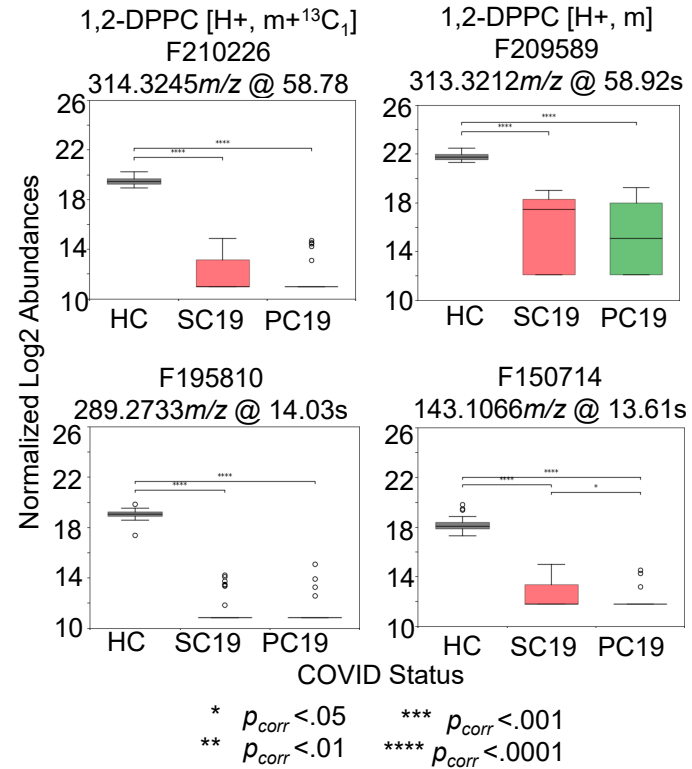
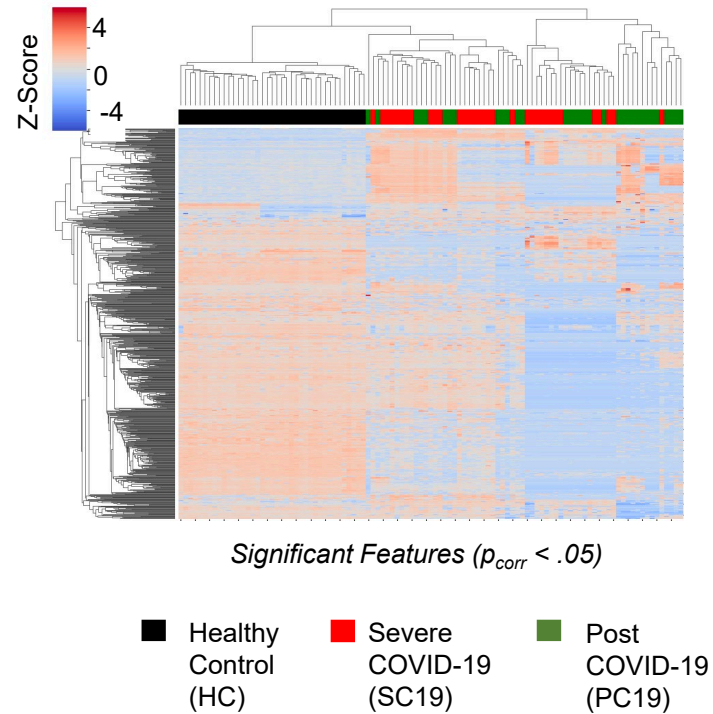


Example Workflow



Example pcpfm Results

Ansone 2021 re-analysis



Pcpfm recapitulates the clustering of patients from the original manuscript

But produces different annotations.

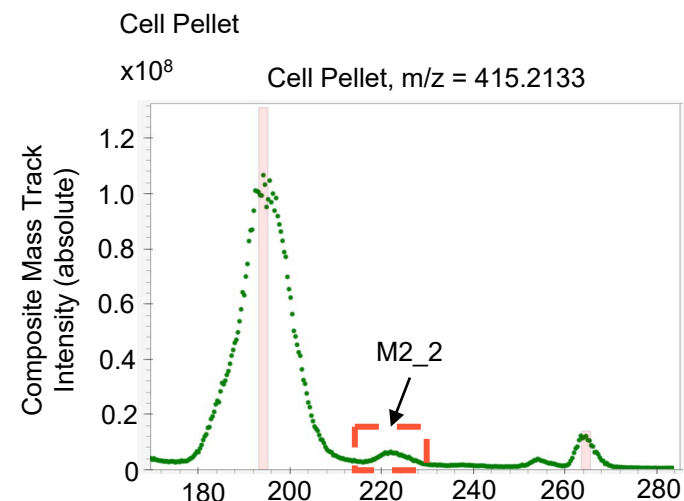
Example pcpfm Results

Bowen 2023 re-analysis

Formula	ID	Bowen	Preferred	Full
C ₂₂ H ₂₇ FN ₄ O ₂	M0_1	Detected	Detected	Detected
C ₂₂ H ₂₇ FN ₄ O ₂	M0_2	Detected	Not Detected	Detected
C ₂₀ H ₂₃ FN ₄ O ₂	M1	Detected	Not Detected	Detected
C ₂₂ H ₂₇ FN ₄ O ₃	M2_1	Detected	Not Detected	Detected
C ₂₂ H ₂₇ FN ₄ O ₃	M2_2	Detected	Not Detected	Not Detected
C ₁₈ H ₁₉ FN ₄ O ₂	M3	Detected	Detected	Detected
C ₂₀ H ₂₃ FN ₄ O ₃	M4_1	Detected	Detected	Detected
C ₂₀ H ₂₃ FN ₄ O ₃	M4_2	Detected	Detected	Detected
C ₂₂ H ₂₅ FN ₄ O ₃	M12	Detected	Detected	Detected
C ₂₀ H ₂₁ FN ₄ O ₃	M14_1	Detected	Not Detected	Detected
C ₂₀ H ₂₁ FN ₄ O ₃	M14_2	Detected	Not Detected	Detected
C ₈ H ₁₈ N ₂ O	M20	Detected	Detected	Detected

Formula	ID	Bowen	Preferred	Full
C ₂₂ H ₂₇ FN ₄ O ₂	M0_1	Detected	Detected	Detected
C ₂₂ H ₂₇ FN ₄ O ₂	M0_2	Detected	Detected	Detected
C ₂₀ H ₂₃ FN ₄ O ₂	M1	Detected	Not Detected	Detected
C ₂₂ H ₂₇ FN ₄ O ₃	M2_1	Detected	Detected	Detected
C ₂₂ H ₂₇ FN ₄ O ₃	M2_2	Detected	Detected	Detected
C ₂₂ H ₂₅ FN ₄ O ₃	M12	Detected	Detected	Detected
C ₂₀ H ₂₁ FN ₄ O ₃	M14_1	Detected	Detected	Detected
C ₈ H ₁₈ N ₂ O	M20	Detected	Detected	Detected

■ Not Detected
■ Detected

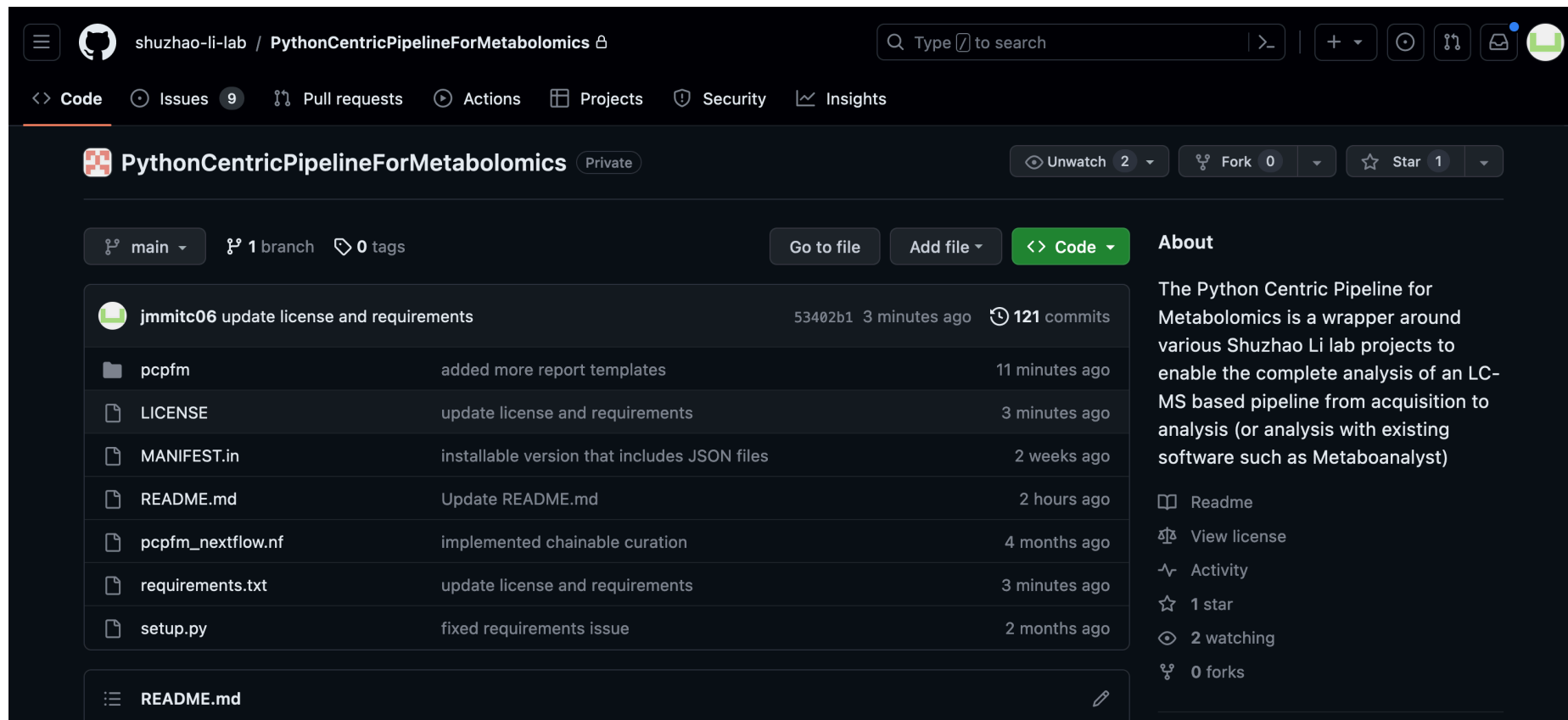


Pcpfm finds most features without a specialized workflow

Trivial inspection of missing features

Open Source Matters

- Both Asari and pipeline are open-source
- Reuse welcome
- Feel free to submit issues and request features



The screenshot shows the GitHub interface for the repository `shuzhao-li-lab / PythonCentricPipelineForMetabolomics`. The repository is private and has 121 commits, 2 watchers, 0 forks, and 1 star. The main branch is selected, and the file list shows recent updates to `pcpfm`, `LICENSE`, `MANIFEST.in`, `README.md`, `pcpfm_nextflow.nf`, `requirements.txt`, and `setup.py`. The `README.md` file is currently selected and open in the editor view.

PythonCentricPipelineForMetabolomics Private

Unwatch 2 Fork 0 Star 1

main 1 branch 0 tags

Go to file Add file Code

About

The Python Centric Pipeline for Metabolomics is a wrapper around various Shuzhao Li lab projects to enable the complete analysis of an LC-MS based pipeline from acquisition to analysis (or analysis with existing software such as Metaboanalyst)

Readme View license Activity 1 star 2 watching 0 forks

File	Commit Message	Time Ago
pcpfm	added more report templates	11 minutes ago
LICENSE	update license and requirements	3 minutes ago
MANIFEST.in	installable version that includes JSON files	2 weeks ago
README.md	Update README.md	2 hours ago
pcpfm_nextflow.nf	implemented chainable curation	4 months ago
requirements.txt	update license and requirements	3 minutes ago
setup.py	fixed requirements issue	2 months ago



Conclusions

- Still looking for feedback, evolving but stable enough.
- [Data analysis notebooks in the repository.](https://github.com/shuzhao-li-lab/PythonCentricPipelineForMetabolomics)
<https://github.com/shuzhao-li-lab/PythonCentricPipelineForMetabolomics>
- [Link to other tutorials and examples.](https://github.com/shuzhao-li-lab/pcpfm_tutorials)
https://github.com/shuzhao-li-lab/pcpfm_tutorials (work in progress)
- [Data repo for datasets](https://github.com/shuzhao-li-lab/data)
[https://github.com/shuzhao-li-lab/data/](https://github.com/shuzhao-li-lab/data)

Contact me at: joshua.mitchell@jax.org

Acknowledgements

Shuzhao Li Group

- Shuzhao Li
- Yuanye Chi
- Shujian Cheng
- Maheshwor Thapa
- Minghao Gong
- Amnah Siddiqa (Former)

Other JAX Collaborators

Robson Lab

- Juliana Alcoforado Diniz
- Zukai Liu
- Arti Taggar
- Dylan Baker
- Anahita Amiri

Non-Jax Collaborators

- Stephen Barnes
- Jianguo (Jeff) Xia
- Lei Xu

The Jackson Laboratory for
Genomic Medicine
Farmington, CT



Funding

- UM1HG012651
- RO1AI149746
- U01CA235493


A comment on Metabolomics from 2009

webofstories.com/play/james.watson/97

WEB of STORIES

STORYTELLERS | THEMES | BLOG | ABOUT | HELP

A STORY LIVES FOREVER



Glycolysis, cancer and metabolomics
James Watson Scientist

Play all

So, you know, the molecular biologists don't think biochemistry. And so the cancer field is now controlled by molecular biologists. You know, whereas, you know, when I was a boy, molecular biologists didn't exist and the big people that everyone respected were the good biochemists, starting with Warburg.

So you know, if I were, you know, doing a PhD, I'd do it metabolomics, you know, if you wanted to be a biochemist because you're likely to get a good job afterward, 'cause everyone will see the need for it.

Related [Transcript](#)

of glycolysis. So
biologists don't think bio
cancer field is now contr
biologists. You know, whe
when I was a boy, wh
exist and the big p
respected were th
with Warburg. And
DNA is made and
and, you know, th
learned it. There a
if it's interfered wi
into a rational exp
working on cancer
metabolism and this
metabolomics, where you
the small molecules pr
quantitate them
and
or you know, how much c