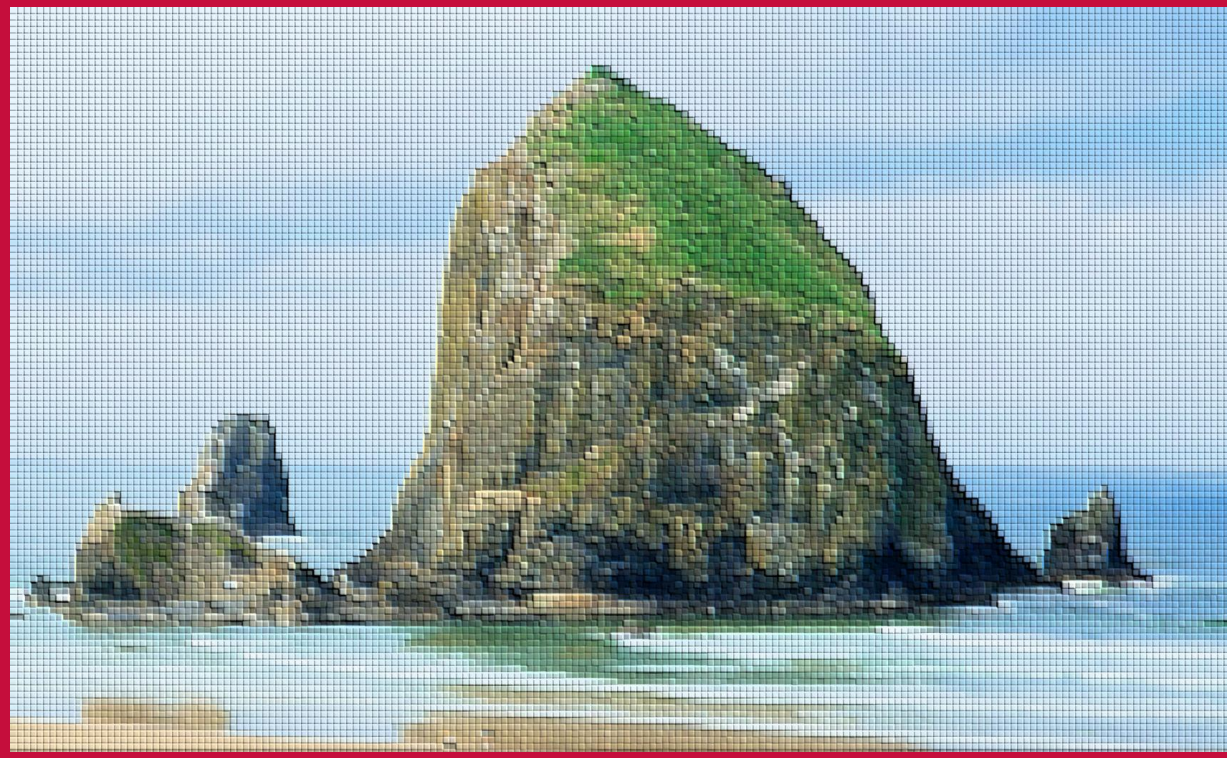
GILEAD
Creating Possible

Fast, Large-Scale Analysis of Mass Spectrometry Data Repositories to Find the Needle in the Haystack



Simon Letarte, Anoop Mishra, Wei-Ting Liu,
Doug Rehder, Franz Gabeau, Yingying Zheng

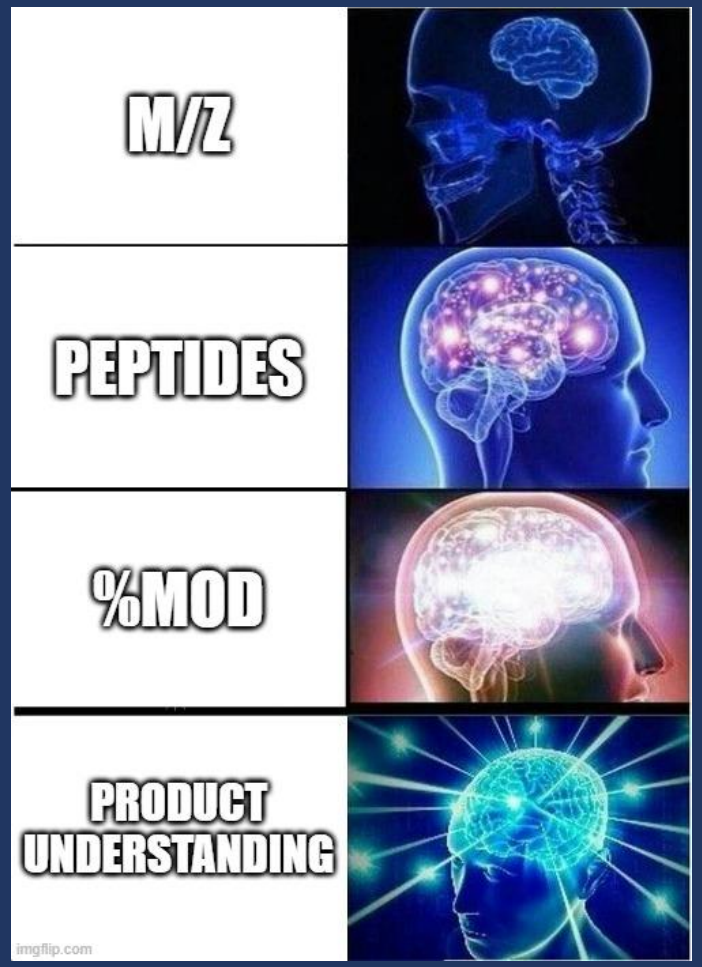
Overview

1. Intro – MS Data
2. Leveraging Big Data for Mass Spectrometry
3. Host Cell Proteins detection by Mass Spectrometry
4. Peptide Mapping PQAs

Project Initiation

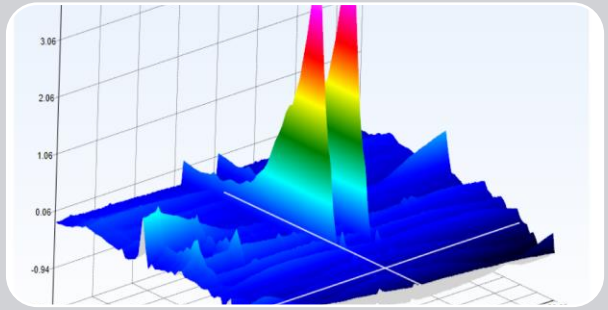
- Collaboration between Structure Characterization Team and Data Science Team
- Find problems we can collaborate together
  - Large amount of data with minimal annotation
  - Wanted a tool to search older data for specific ions
- Added functionality as we go
- Working on other types of data that do not have MS

G.MSIS Gilead Mass Spectrometry Ion Search

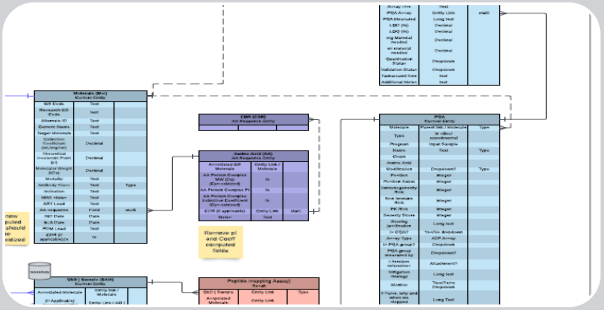


Mass Spectrometry Data

The 3 Stages of Mass Spectrometry Data



Deam					
HC-N392/N387+Succ	GFYPSDIAVEWESNGQPENNYK	2.06	2.27	2.73	2.60
HC N318+Succ	VVSVLTVLHQDWLNGK VVSVLTVLHQDWLNGKEYK	2.05	2.61	2.66	2.25
HC D283+Succ	FNWYVDGVEVHNAK TPEVTCVVVDVSHEDPEVKFNW YVDGVEVHNAK	2.18	3.40	1.56	3.56
HC M255+Oxid	DTLMISR	1.26	1.29	1.05	0.92
HC K450 Lys loss	SLSLSPGK WQQGNVFSCSVMHEALHNHYT QKSLSLSPGK	87.01	88.49	90.73	88.79
HC Q1+Gln->PyroGlu	QVTLR	99.29	99.31	99.75	99.32
HC N300+M5	EEQYNSTYR TKPREEQYNSTYR	1.71	1.29	1.10	1.22
HC N300+A1G0F	EEQYNSTYR TKPREEQYNSTYR	6.05	5.04	11.26	10.95
HC N300+A2G0F	EEQYNSTYR TKPREEQYNSTYR	41.13	40.14	37.36	37.05
N300+A1G1F	EEQYNSTYR	4.06	3.68	5.92	6.13



Raw Data

- RT, m/z
- 1 Gb/file

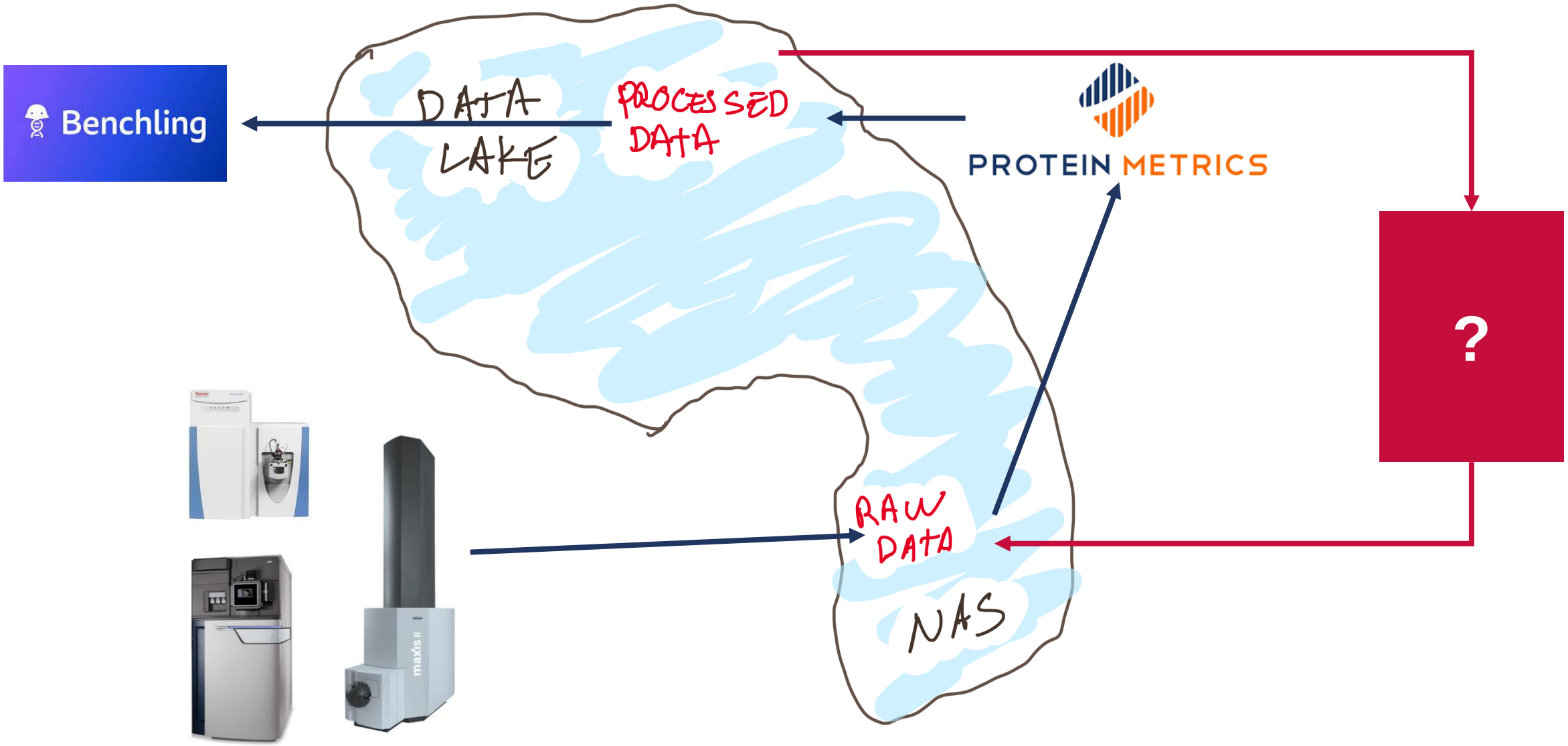
Processed Data

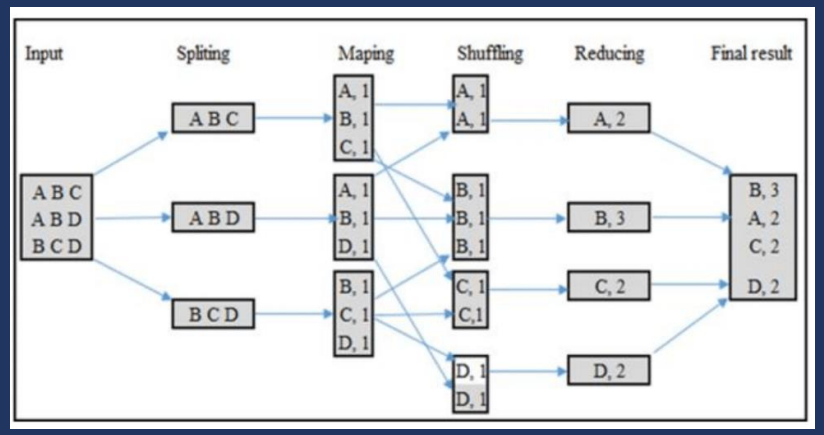
- Peptides
- 100 Kb/file

Connected Data

- Metadata
- Databases

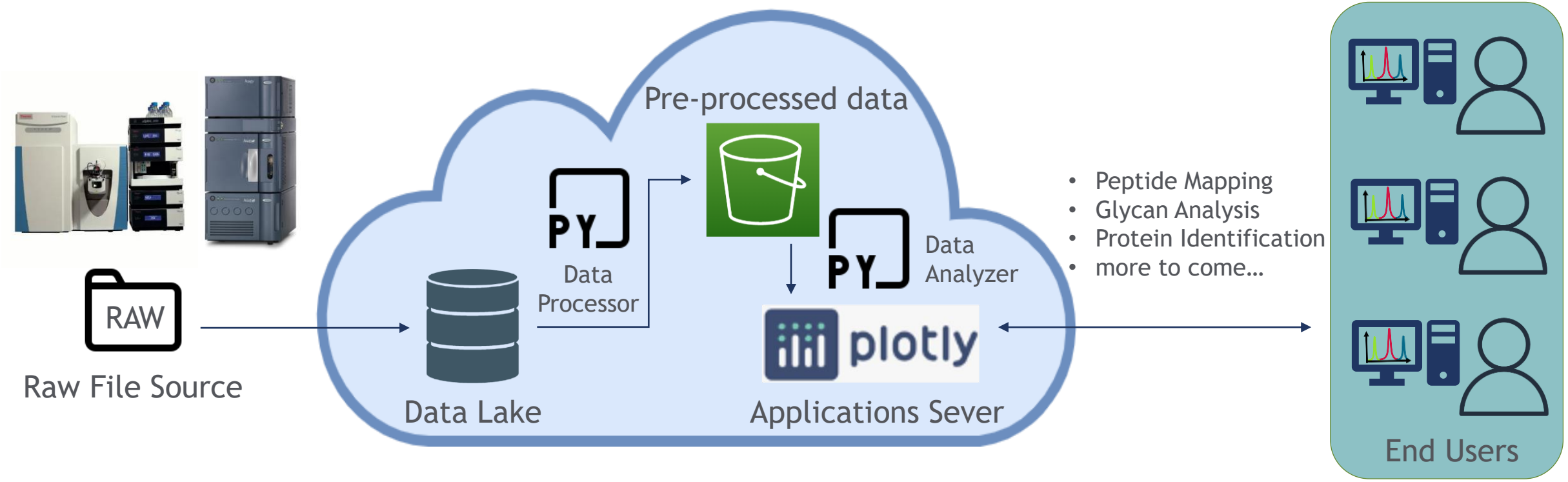
Data Workflow





Leveraging Big Data for Mass Spectrometry

Automated Data Pipeline in the Cloud

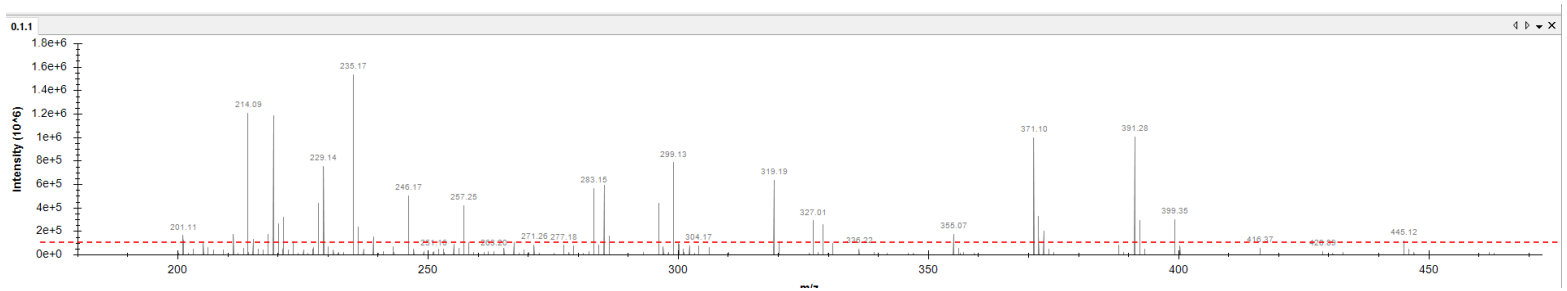


- Developed customized web applications to streamline resource intensive analysis
  - Reduce file size by 90%, speed up data processing time from hours to secs, reduce manual procedures
- End-to-end data pipeline to connect web applications directly with data source
  - Fully automated workflows, central storage, facilitate cross-groups collaborations

Key Technology Behind Faster Ion Search

Key Concept 1: Search Time ∝ Data Size

Step 1 - Low intensity noise is removed from the raw file to significantly reduce the file size by 90%



Example: 1E4 cut-off helps reduce the file size significantly

Note: The cut-off value is customizable

Key Concept 2: Metadata Removes Unwanted Search Candidates

Step 2 – Binning and binary vector representation is applied to generate meta data

1	0	1	1	1	1	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	1	0	

Segment the x-axis m/z values into multiple bins. Where there is a peak within the bin, mark it as “1”. Where there is no peak within the bin, mark it as “0”.

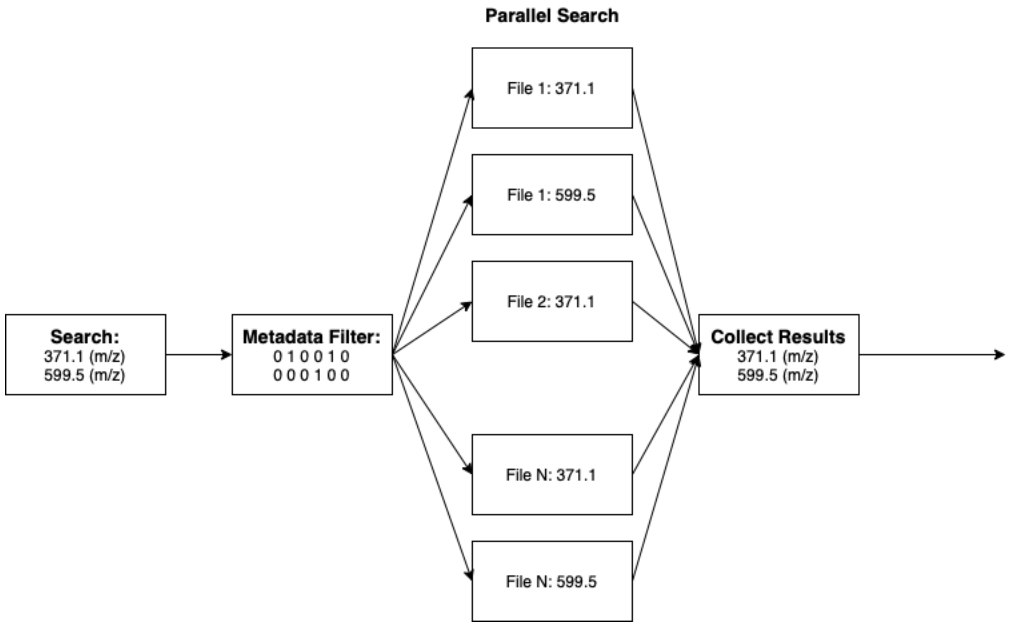
0 1 0 0 1 0 1 0 0 0 0 ✓
0 0 0 1 0 0 0 0 0 0 0 ✗
0 0 0 1 0 0 1 1 0 0 0 ✓

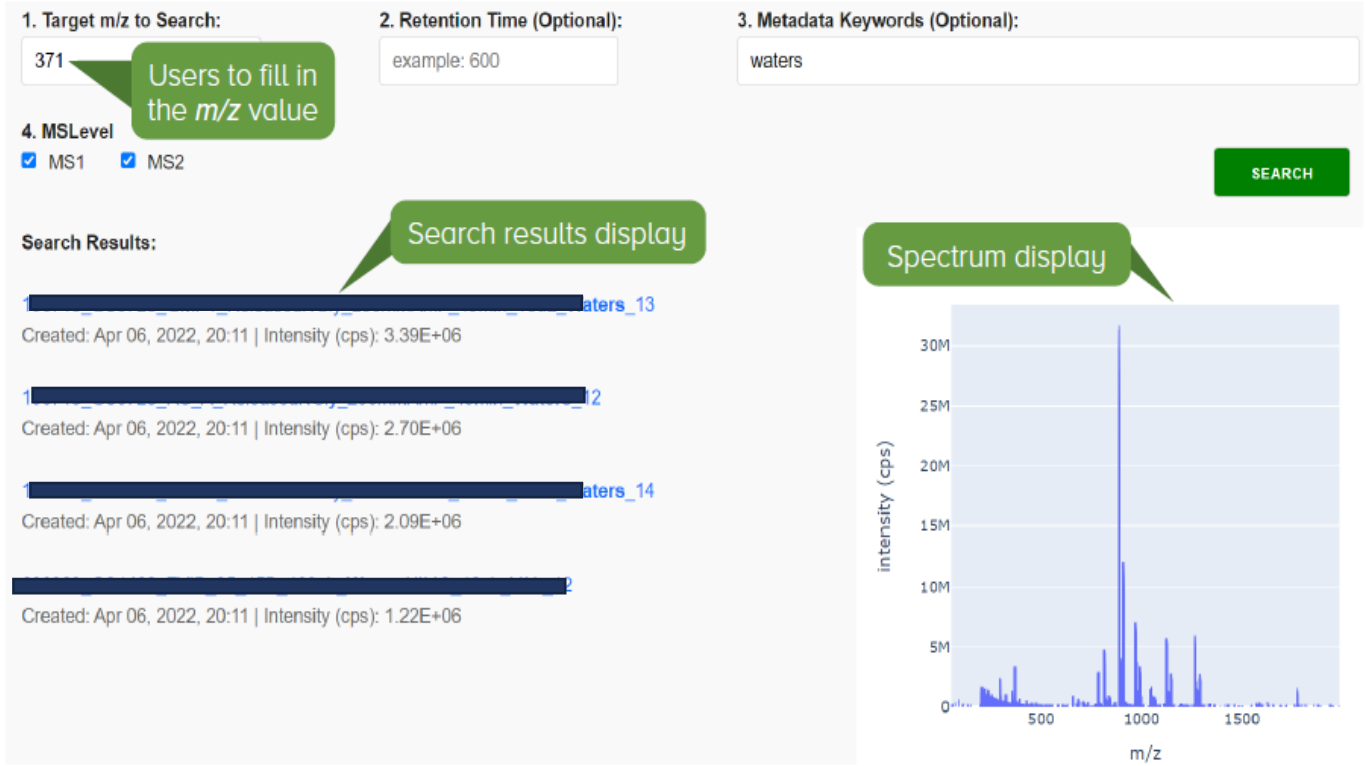
Find Vector Matches (logical AND > 0)

Key Technology Behind Faster Ion Search

Key Concept 3: Parallel Search (Map Reduce) To Deliver Instant Results

Step 3 – Custom python app to search multiple files and multiple data points **all at once**





The Application Interface

M / Z

345.234 & 873.234 & 324.34
or
371.343
or
657.4554 & 45.435

TOLERANCE

0.005

[x] **MS1** [ ] **MS2**

KEYWORDS Glycan Waters

SEARCH

RESULTS

1. HPC/ /abc_file1.raw

345.234: intensity: 324+e5 retention time: 62.34
873.234: intensity: 53+e5 retention time: 78.31
324.34: intensity: 53+e5 retention time: 78.31

2. HPC/ /abc_file3.raw

371.343: intensity: 43+e5 retention time: 56.12

HCP Analysis by Proteomics Mass Spectrometry

PLBL-2 HCP

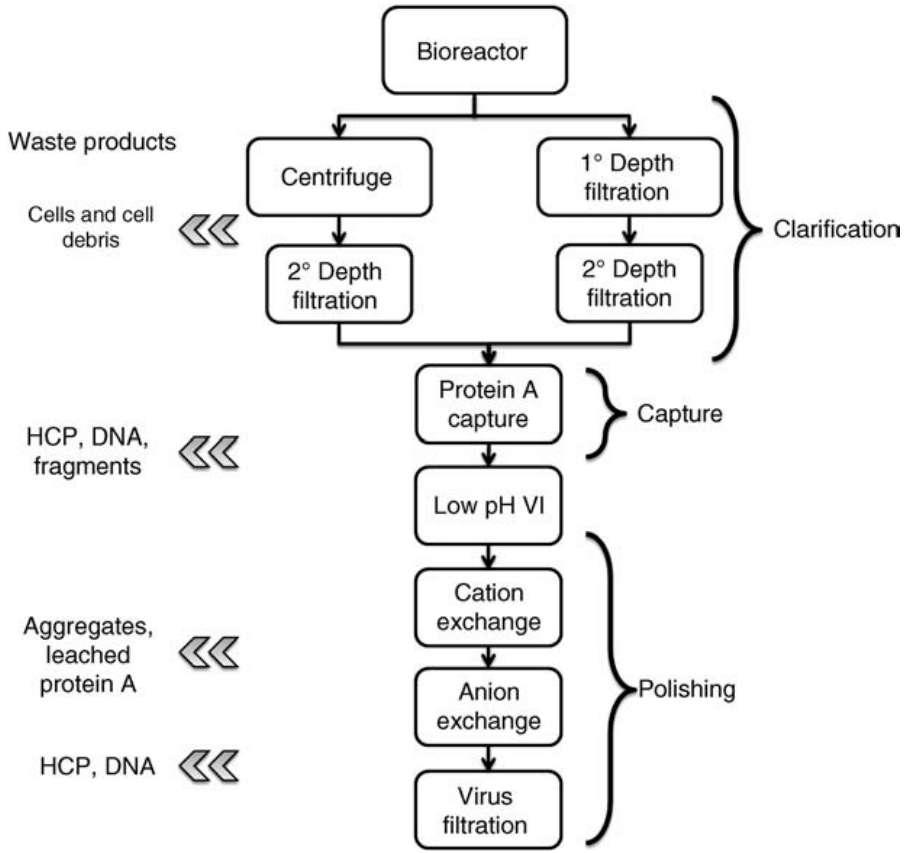
>CHO PLBL-2 Sequence

QNLDPPVSRVRSVLLDAASGQLRLVDGIHPYAVAWANLTNAI
RETGWAYLDLGTNGSYNDSLQAYAAGVVEASVSEELIYMHWM
NTMVNYCGPFEYEVGYCEKLKSFLEINLEWMQREMELSQDSP
YWHQVRLTLLQLKGLEDSYEGRLTFPTGRFTIKPLGFLLLQI
AGDLEDLEQALNKTSTKLSLGSGSCSAIIKLLPGARDLLVAH
NTWNSYQNMLRIIKKYQLQFRQGPQEAYPLIAGNNLVFSSYP
GTIFSGDDFYILGSGLVTLETTIGNKNPALWKYVQPQGCVLE
WIRNIVANRLALDGATWADIFKQFNSGTYNNQWMIVDYKAFI
PNGPSPGSRVLTILEQIPGMVVVADKTEDLYKTTYWASYNIP
FFEIVFNASGLQDLVAQYGDWFSYTKNPRAQIFQRDQSLVED
MNSMVRLIRYNNFLHDPLSLCEACIPKPNAENAISARSDLNP
ANGSYPFQALYQRPHGGIDVKVTSFSLAKRMSMLAASGPTWD
QLPPFQWSLSPFRSMLHMGQPDLWTFSPISVPWD

PLBL-2 is a common Host Cell Protein that pose an immunogenicity risk and may degrade polysorbate

M/Z	Charge	Mass Theo	Peptide
513.23	2	1024.4462	GLEDSYEGR
762.35	2	1522.6756	DQSLVEDMNSMVR
615.35	2	1228.6776	SVLLDAASGQLR
824.42	2	1646.8239	YVQPQGCVLEWIR
396.23	2	790.4337	LTFPTGR
600.32	2	1198.6095	AFIPNGPSPGSR
513.27	2	1024.5302	QNLDPPVSR
427.73	2	853.4446	YQLQFR

What are Host Cell Proteins (HCP)?



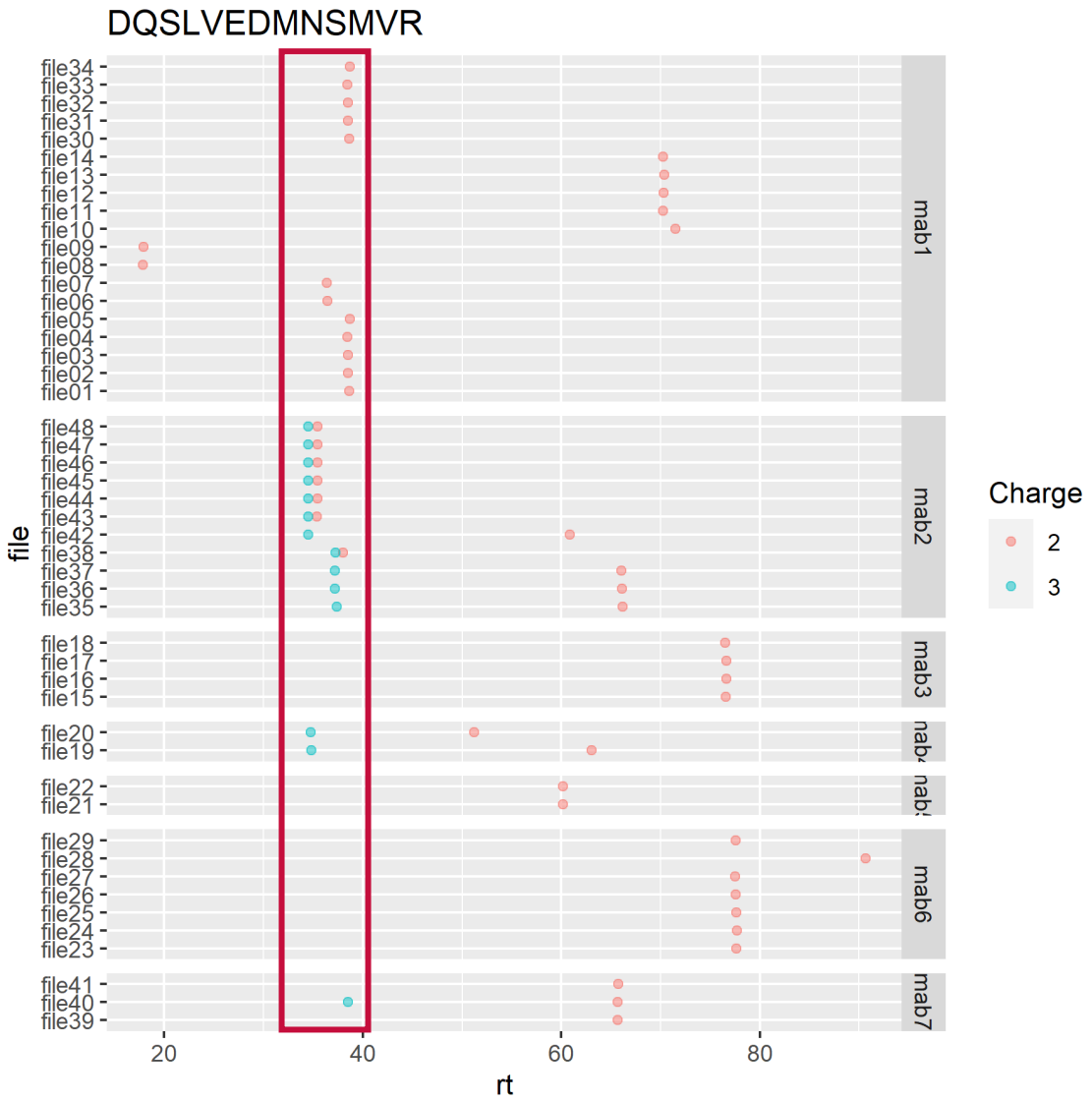
Gillespie *et al*, 2015

- Proteins from host cell organism that are present in the drug substance
- Incomplete purification by the downstream process:
  - Same physico-chemical properties as the drug
  - Affinity with the drug and hitchhiker effect.

Results Overview

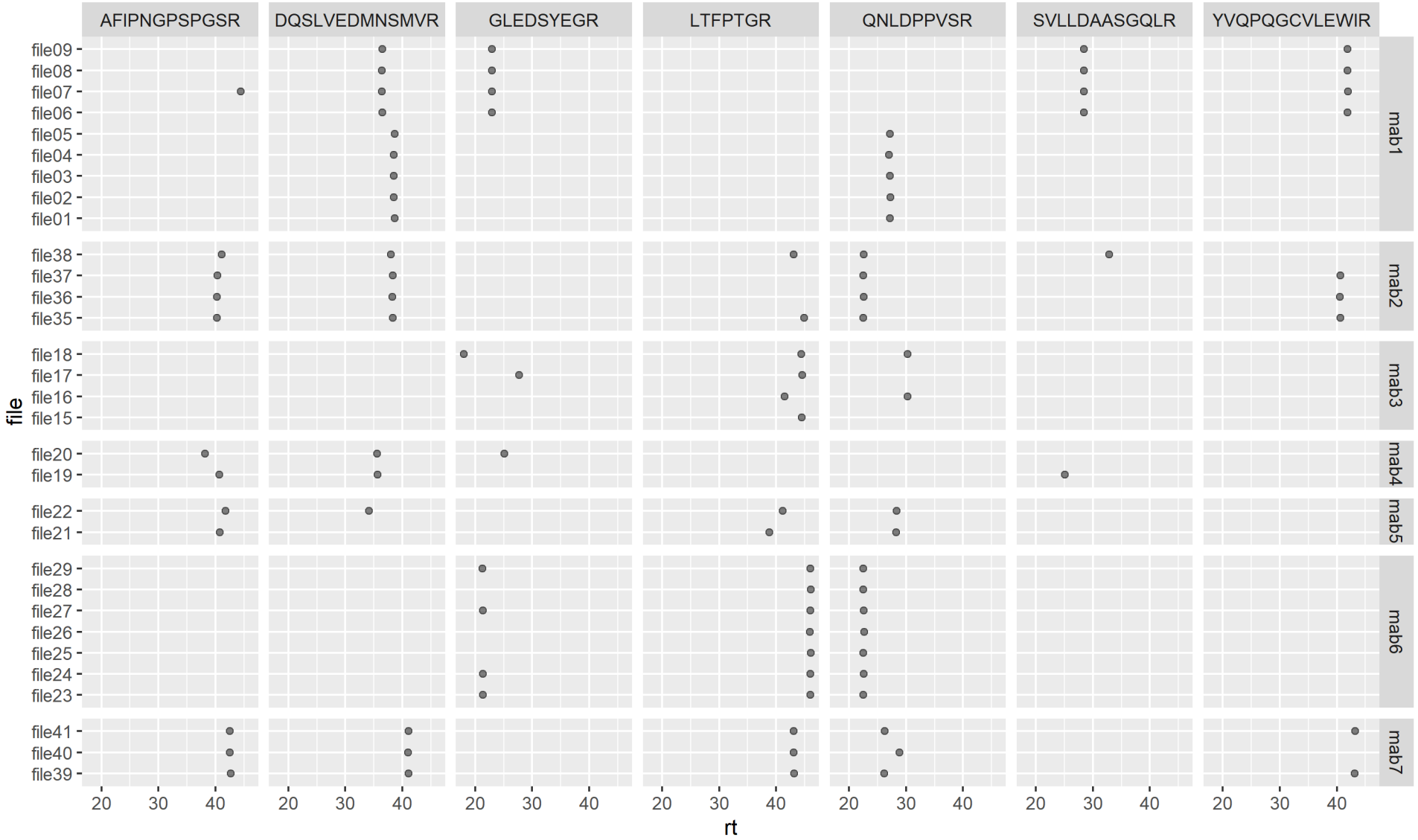
Intensity > 1e4
Accuracy < 0.005 Da

Retention Time Analysis

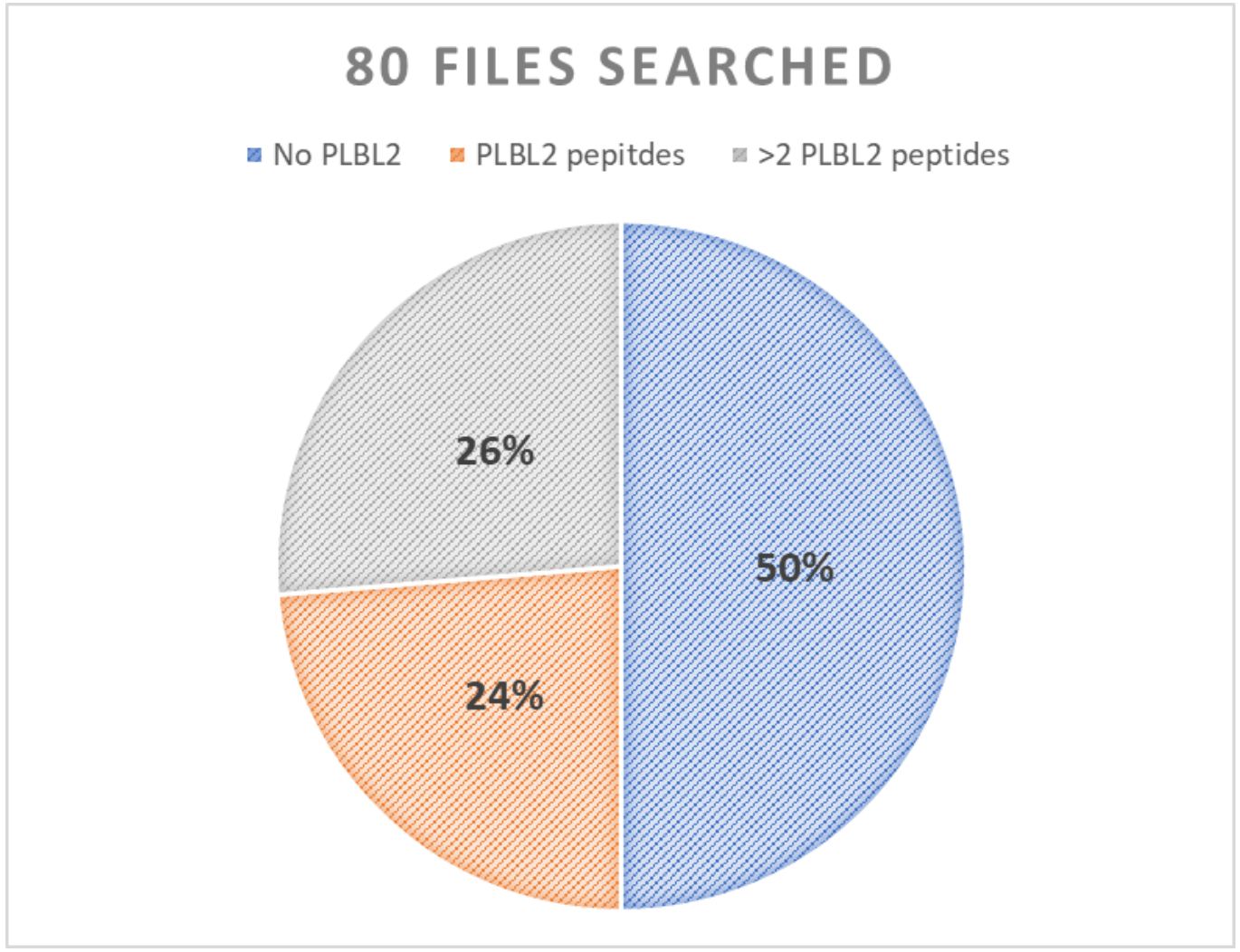


- Different gradients can be used for different experiments
- Algorithm currently selects the most intense peak within the user-specified RT and MS1 tolerance

HCP RT Filtering



file	molecule	nbPep
file06	mab1	4
file07	mab1	5
file08	mab1	4
file09	mab1	4
file35	mab2	5
file36	mab2	4
file37	mab2	4
file38	mab2	5
file18	mab3	3
file19	mab4	3
file20	mab4	3
file21	mab5	3
file22	mab5	4
file23	mab6	3
file24	mab6	3
file27	mab6	3
file29	mab6	3
file39	mab7	5
file40	mab7	4
file41	mab7	5



file	molecule	nbPep
file06	mab1	4
file07	mab1	5
file08	mab1	4
file09	mab1	4
file35	mab2	5
file36	mab2	4
file37	mab2	4
file38	mab2	5
file18	mab3	3
file19	mab4	3
file20	mab4	3
file21	mab5	3
file22	mab5	4
file23	mab6	3
file24	mab6	3
file27	mab6	3
file29	mab6	3
file39	mab7	5
file40	mab7	4
file41	mab7	5

HCP Summary

1. Used molecule known to have PLBL2 HCP present as control
2. MS1 and retention time to locate previously identified PLBL2 peptides
3. Filter with at least 2 peptides to identify other runs that might have HCP
4. Use Protein Metrics Byos software to confirm hits

5. Working on using MS2 data to create “transitions” instead of using RT

Peptide Mapping PQAs

Rationale - more than m/z

- For some applications, PQAs might not be a fixed m/z
- Depends on sequence to calculate m/z based on peptides
- Need new interface:
  - Input protein and digest
  - Select modifications
- Return results
- Filter by selecting MS1, charge states, RT, intensity and MS2

Methionine Oxidation Characterization

Assay workflow

protein

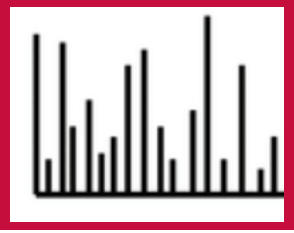
WVTFISLLFLFM*SSAYS**R**SEVAH**R**D
LGEEM*NF**K**ALVLIAFAQYLQQCPF

Trypsin digestion
Reduction, Alkylation

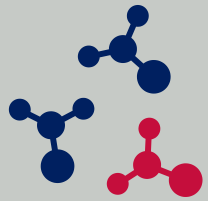
resulting peptides

WVTFISLLFLFM*SSAYSR
SEVAHR
DLGEEM*NFK
ALVLIAFAQYLQQCPF

LC-MS spectra



Data analysis pipeline

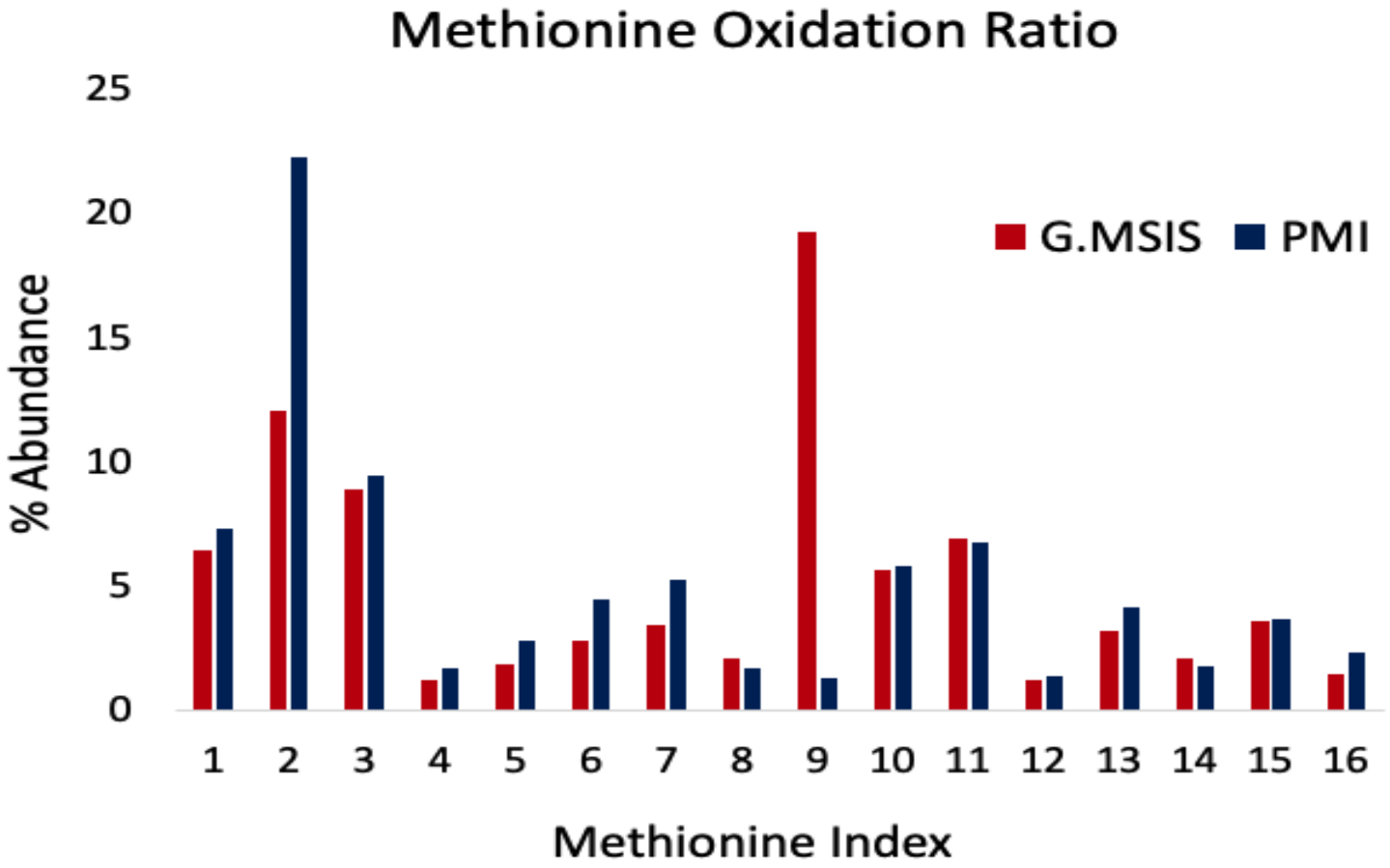


Calculate theoretical mass for both native and modified forms

Batch search &
Report generation

Comparison with Valited Results

- **The data processing time was reduced from a couple hours to a few secs**
- Overall results generally agrees with benchmark analysis
- Our team is working closely with data science team to further optimize this web application



* G.MSIS picked a different peak for Methionine 9 due to the original peak is in very low abundance

Performance Metrics

- 2 seconds to search 40 files with 0.005 Da mass window using serial search
- < 1s for parallel search in large number of files (depending on processing power)
- 70% of sites have methionine oxidation values within 20% of the validated results
- Discrepancies arise from:
  - Wrong peak selected
  - Ratios calculated from peak height vs peak area
- Future improvements
  - Leverage MS2 to ensure right peak is selected
  - Implement peptide modification GUI to the tool

Conclusions

- We have a tool to interrogate large amounts of data in real time
- We can retrospectively ask scientific questions
- We barely scratched the surface of what is possible
- Better use of metadata/ELN

Future Work

- Adding MS2 capability to ensure the right peak is selected
- Adding GUI for peptide modification search
- Released glycans – search chromatography dimension without MS
- NLP for contextualizing the data

Acknowledgements

Biologics Analytical Organization

Alexandre Ambrogelly

Structure Characterization Team

Yana Lyon

Sean Shen

Mike Li

Analytical Development

Tawnya Flick