A novel cloud-native pipeline enabling **deep, unbiased** proteomics at **extreme scale**

Seth Just*, Amir Alavi, Andrew Nichols, Jian Wang, **Iman Mohtashemi**, Theodore Platt, Serafim Batzoglou

Enabling large-scale proteomics analysis with cluster computing

Liquid Chromatography coupled to Mass Spectrometry (LC-MS) is a ubiquitous proteomics technology due to its speed, sensitivity, and flexibility. While instrumentation hardware continues to improve, corresponding **increases in translating LC-MS data to insight have lagged**. Although substantial progress has been made in data processing algorithms, all common tools are incapable of supporting experiments larger than a few hundred LC-MS injections. This **limits statistical power of studies** and prevents interrogation of large proteomic data repositories. Significant development and optimization efforts have been invested in the last two decades to ensure reliability and performance of existing tools, but these optimizations unintentionally limit adaptation of these solutions to handle extremely large datasets.Processing thousands of LC-MS files to generate peptide-spectrum-matches (PSMs) can be accomplished in parallel, as shown in our prior work. However, translating raw LC-MS data into biological insight across hundreds or thousands of samples is challenging with available proteomics data analysis tools. Here we combine multiple approaches and demonstrate a proteomics data analysis pipeline that is capable of scaling to thousands of LC-MS datasets with quick turnaround and low cost.

Pipeline overview

Parallel Search: PSMs are generated from each individual injection in parallel by standard DDA or DIA search engines (*e.g.,* MSFragger or EncyclopeDIA)

Apache Spark: Search results are loaded into a Spark Dataframe for efficient access to extreme-scale datasets in a cluster computing environment

Match Rescoring: A rescoring model is trained (using the Percolator[1] approach) and evaluated to produce a list of scored PSMs

FDR Estimation: PSM- or peptide-level *q*-values are estimated by a distributed implementation of the mixture-maximum (MixMax) algorithm[2] (to support peptide- and spectrum-centric searches)

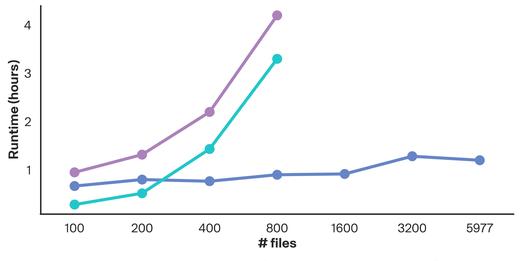


Figure 1. Scalability of Legacy Workflows

Typical scalability of currently available proteomics data analysis pipelines is exemplified by the open-source Alphapept pipeline. Individual file processing on an AWS ECS cluster auto-scales well for feature finding and PSM matching. However, for steps that require data aggregation, vertical scalability becomes a limiting factor due to size/memory constraints of a single machine, requiring an alternative approach.

Scalable, modular implementations of key algorithms for faster insight into big proteomics data

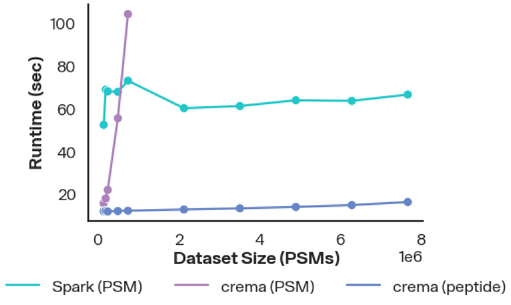


Figure 2. Distributed FDR Estimation Gives Performance at Scale

Our Spark-based MixMax implementation (teal) estimates *q*-values in near-constant time, while an optimized single-node implementation (`crema`[3]; purple) shows quadratic slowdown, limiting ultimate scale. Peptide-level *q*-values are more easily estimated using Spark to collect PSMs, then running `crema` (blue).

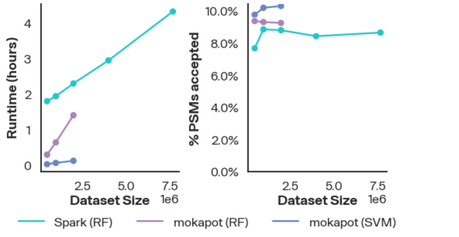
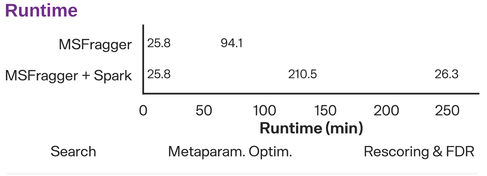


Figure 3. Distributed PSM Rescoring

Spark-based PSM rescoring (random forest; teal) allows greater scale than single-node RF (purple) or SVM (blue).



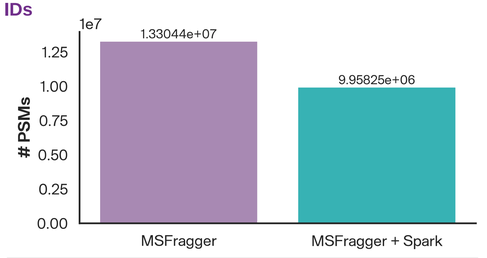


Figure 4. Results of Processing 1,390 Files

Plasma samples were processed with the Proteograph™ Assay. Resulting LC-MS data were searched by MSFragger[4] using a combined "entrapment" FASTA (Human/Yeast/Arabadopsis) and processed by either MSFragger/Philosopher[5] (purple) or our Spark pipeline (random forest; teal) with a 1% PSM-level FDR threshold.

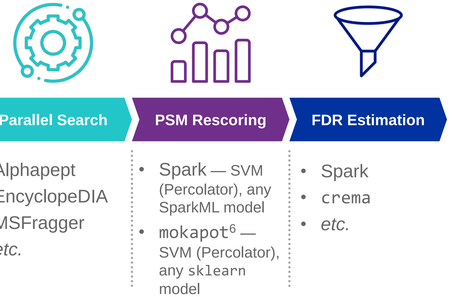


Figure 5. Modular Design Supports Multiple Algorithms and Workflows. Designing around interfaces decouples workflows from any specific algorithms or implementations. Many implementations for each step of processing can be freely interchanged within a single workflow. Multiple workflows can combine modules to achieve various goals.

Future Directions

Library creation: Scalable processing enables mining of repository-scale datasets to produce highly sensitive libraries.

Model building & transfer: rescoring models are trained at large scale, then applied to smaller datasets, boosting sensitivity and consistency.

Extreme-scale quantification: Align PSMs across files and perform transition refinement and/or cross-run matching, followed by parallel signal extraction.

Conclusion

- **Deep, scalable proteomics** experiments require **next-generation data processing** tools.
- We have developed a **modular pipeline** for efficient and **scalable** processing of extreme-scale proteomics experiments that can support **any** algorithmic approach.
- Published algorithms can be re-implemented using **modern, scalable infrastructure** and achieve similar performance to current single-node tools.

References

[1] Granholm *et al.* 2012
[2] Keich *et al.* 2015
[3] github.com/Noble-Lab/crema
[4] Nesvizhskii *et al.* 2017
[5] Leprevost *et al.* 2020
[6] Fondrie et *al.* 2021