Defining, Measuring, and Evaluating Operational Excellence for Three DIA-MS Informatics Solutions

Gautam Saxena[1]; AMP-PD Investigators and Sponsors; Jennifer E Van Eyk[3]
[1]GoldenHaystack Lab, Bethesda, MD; [2]AMP-PD, USA; [3]Cedars-Sinai Medical Center, Los Angeles, CA

Introduction:

For nearly 30 years, MS-based discovery proteomics has been promising to outshine genomics. However, at least four major challenges have historically prevented our community from realizing the expected potential benefits: (a) the dynamic range of the LC-MS instruments was not sufficiently large; (b) the throughput of the LC-MS instruments often prevented larger-scale studies (i.e., 100s or 1000s of samples); (c) >99% of labs used informatics solutions that only accessed < ~10% of the data in the MS files (despite the much higher biological value of the peptides present in the remaining ~90%); and (d) the informatics was too slow, disjointed, difficult to use, and of questionable robustness to realistically analyze multiple large-scale projects in a reasonable timeframe and showcase the results to scientists primarily outside of our community.

However, in the last several years, technical advances have largely addressed the first three concerns: the LC-MS vendors have released order-of-magnitude improvements in both sensitivity and dynamic range for their latest instrument platforms while simultaneously increasing throughput by at least 3x; about half-a-dozen small companies have been founded with the ability to enrich the lower abundant proteins for plasma and CSF samples; and, in early 2025, we introduced an algorithm called GoldenHaystack (GH) that can quantify not just (a) the ~10% of peptides in DIA-MS datasets that are in protein library (i.e., FASTA) search spaces but also (b) the biologically invaluable remaining ~90% of peptides in DIA-MS files that may not have an initially known sequence or PTMs. Briefly, GH accomplishes this quantitation of nearly all the peptides in the MS by leveraging the fact that MS2 XIC fragments from different peptides that exactly coelute in one subset of samples do not exactly coelute in another subset of samples, and so for projects of sufficient size (e.g., ≥ ~30 samples) MS2 XIC fragments can be computationally matched across a project's set of samples, then computationally separated, and then quantified (Fig 1.), regardless of whether the peptide signals are subsequently optionally identified from a FASTA file; more details can be found in the original preprint and ASMS 2025 poster.

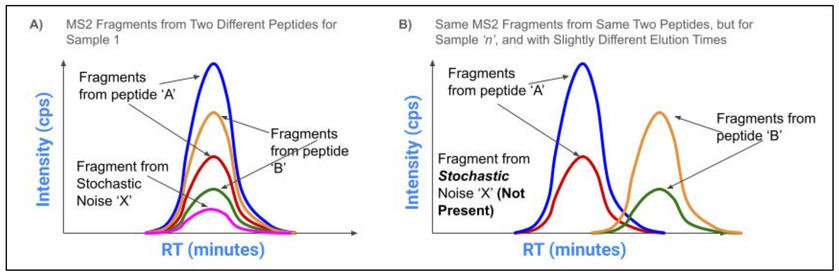

Fig. 1: Illustration of core insight behind GH algorithm

Objective(s):

None of those above independent technical developments to characterize the plasma proteome by MS matter however if the informatics solution is almost unusably slow, overly complicated to use, lacking in scalable visualization tools (to ensure auditability/transparency of informatics conclusions), and/or of questionable robustness. In other words, the informatics now needs to achieve operational excellence that is measurable. We therefore set out to first define and then measure DIA-MS informatics operational excellence (a) for our GH algorithm and, to the extent possible, (b) for two popular DIA-MS informatics solutions.

Methods:

For a 2583 cerebrospinal fluid (CSF) sample of a case control design, where case were individuals with Parkinson's Disease (PD) and control were individual with similar clinical features but with no PD diagnosis, CSF samples were digested with trypsin and were analyzed on the Orbitrap Exploris 480 (ThermoFisher) using DIA-MS. We measure for GH, DIA-NN 1.9.1 and MSFragger: (a) wall clock time for end-to-end informatics processing (i.e., from raw file to final AI/ML results); (b) wall clock time when a pipeline is re-run with either a single LC-MS file added/removed or a single search parameter changed; (c) wall-clock time for re-running the informatics pipeline after a failure occurs in the middle or towards the end of the pipeline; (d) reproducibility of results (i.e., degree to which running the same pipeline twice produces identical results); and (e) wall-clock time to visualize XICs and spectra for both expected MS2 fragments as well as user-queried MS2 fragments.

Results:

GH processed end-to-end the 2583 raw file PD project in less than 7 hours on a single large computer of 180 CPU cores (360 virtual threads when hyperthreading was turned on) (Fig. 2); for re-running of the pipeline for a change in search parameters, it took less than 1 hour, largely because there was only a single mgf file generated by the GH algorithm for a search engine to search irrespective of the number of LC-MS files (Fig. 3); for failures midway or towards the end of the pipeline, the net additional time was negligible (< 5 minutes) to "continue from point of initial failure" onwards; for reproducibility, running the pipeline twice produced 100% identical results; for separating study conditions using more than a single analyte (i.e., a "panel" approach), an AI module specifically configured for proteomics analysis was built into GH and ran within the previously described ~7 hour end-to-end window and displayed both a final high level precision-recall AUC conclusion (Fig. 4) and a human explainable model using simple trees (Fig. 5); and, for visualization of XICs, it took seconds to produce the plots, even on projects of 100s of samples, and users could also type in m/z fragments of interest (even if the m/z fragments were not matched to any peptide's theoretical m/z fragments) and have those corresponding MS1&2 XICs displayed for a user-selected set of samples (Fig. 6). Users could also access all GH visualization and other functionality from their internet browser, so they did not need a powerful or properly configured computer. For MS-Fragger and DIA-NN, on the identical 180 cpu core server, processing times were typically either orders-of-magnitude larger or it was not possible to run the pipeline in any reasonable timeframe (Fig. 1) and we observed that DIA-NN rarely used more than ~20% of the CPU cores on the 180 cpu core server; DIA-NN showed variation in results when running the exact same pipeline twice with the same input; neither DIA-NN nor MS-Fragger included any built-in AI module to separate study conditions; and Skyline (a tool typically used for visualizing DIA-NN results) could not support visualization of the MS1&2 XICs for this sized project, nor did it have the ability to display MS1&2 XICs for user-supplied m/z fragments.

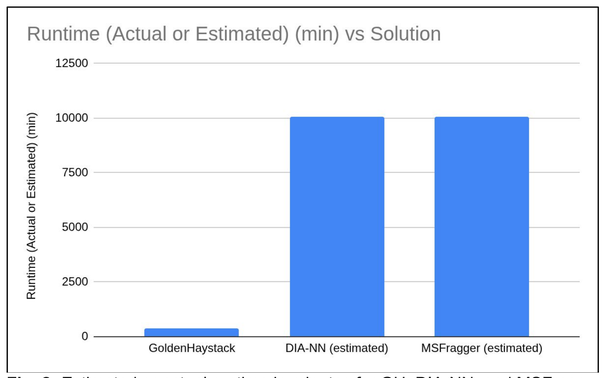

Fig. 2: Estimated or actual runtime in minutes for GH, DIA-NN, and MSFragger. All programs were run on the same 180 cpu core (360 threads) Linux server. If an "estimated" runtime, the numbers reflect the minimum expected time to completion. The real time is most likely substantially greater than the minimum estimated time. DIA-NN typically did not use more than ~20% of available cores. Search parameters included 2 missed cleavages and variable modifications of oxidation(M) and phosphorylation(STY).

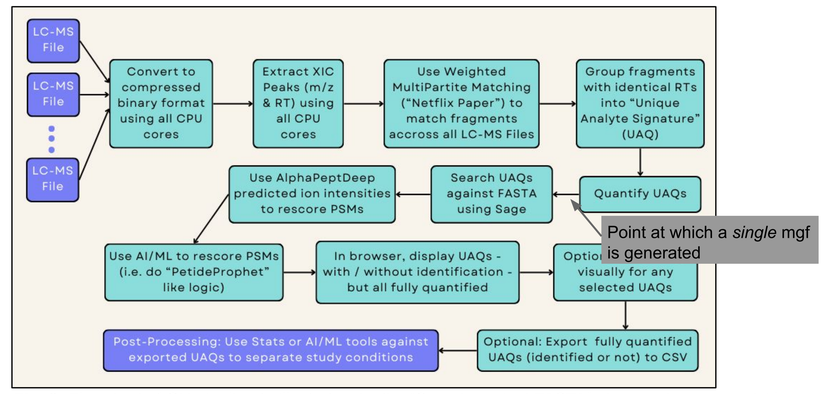

Fig. 3: Flowchart of GH pipeline to illustrate why the "Search against FASTA" step can be rerun quickly using different search parameters, since the step that generates the UAQs generates a *single* mgf-like file for Sage to search even if there are 100s of LC-MS files from the previous steps.

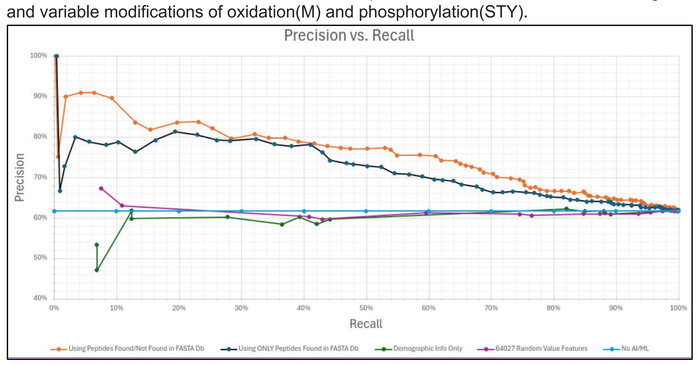

Fig. 4: GH's built-in AI routine can analyze >100s of samples and >>10,000s of quantified peptides (identified or not) that may *together* create a small parsimonious model (e.g., ~200 peptides) that can separate even difficult-to-distinguish-by-lay-person study conditions, and then display a summarizing precision-vs-recall AUC plot under various scenarios.

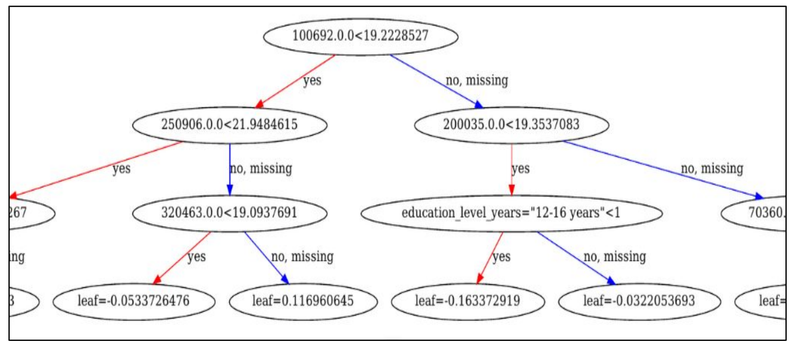

Fig. 5: The AI model is human *explainable* as a series of summable 'n' trees (where 'n' is typically <<100), one of which is shown above, e.g.: "the likelihood of having PD is slightly lower for those whose education level is 12-16 years (i.e., college educated) if they also *simultaneously* have a log2 peptide abundance of > ~19.22 and < ~19.35 for peptides '100692.0.0' and '200035.0.0' respectively."

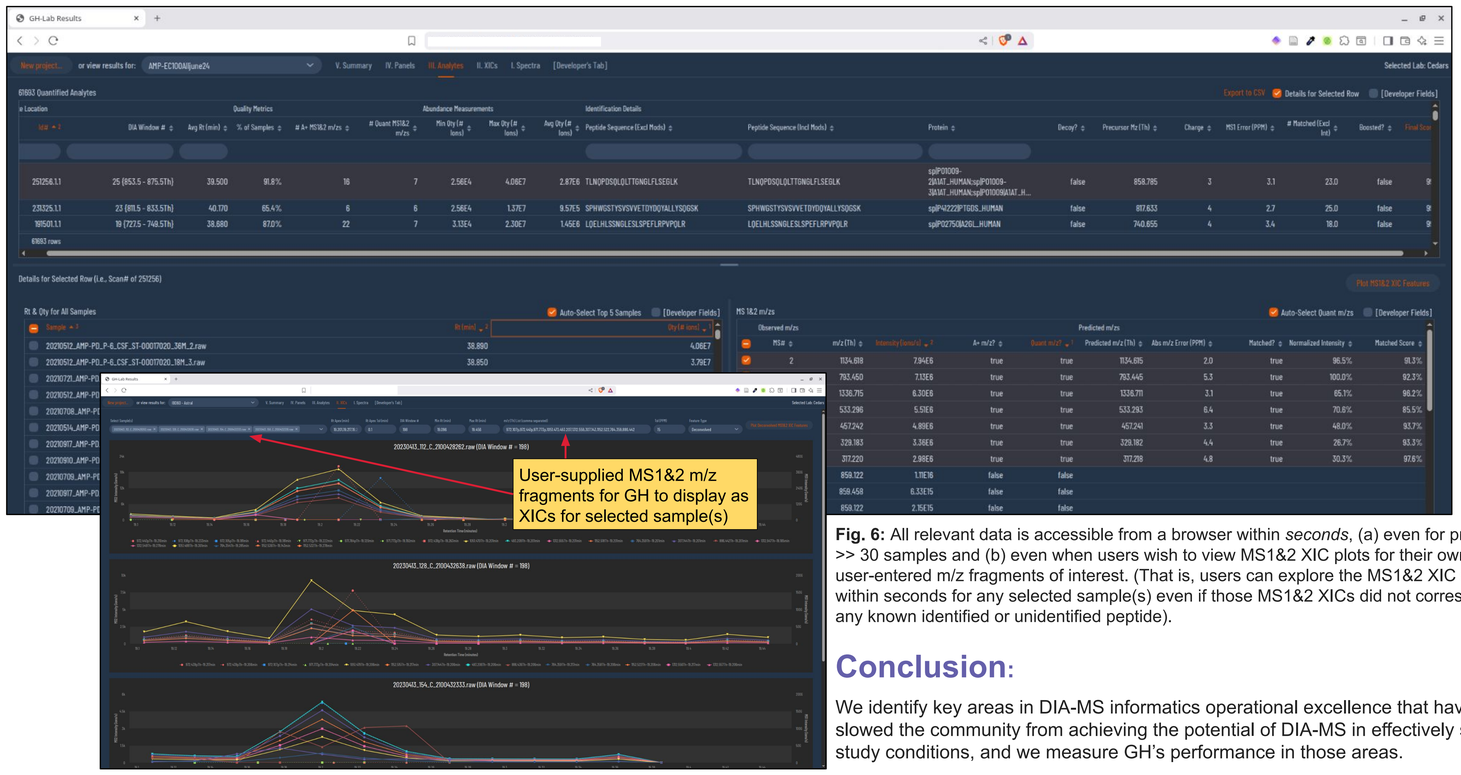

Fig. 6: All relevant data is accessible from a browser within *seconds*, (a) even for projects of >> 30 samples and (b) even when users wish to view MS1&2 XIC plots for their own user-entered m/z fragments of interest. (That is, users can explore the MS1&2 XIC space within seconds for any selected sample(s) even if those MS1&2 XICs did not correspond to any known identified or unidentified peptide).

Conclusion:

We identify key areas in DIA-MS informatics operational excellence that have traditionally slowed the community from achieving the potential of DIA-MS in effectively separating study conditions, and we measure GH's performance in those areas.