From Acquisition to Results – Low memory usage and integrated AI/ML based rescoring for dda-PASEF immunopeptidomics searches in minutes

Justin Sing[1], Sander Willems[2], Vijay Gnanasambandan[3],George Rosenberger[4], Dennis Trede[4], Matthew Albano[3], Jonathan Krieger[5], Tharan Srikumar[5] and Hannes Rost[1]

[1]University of Toronto, Department of Molecular Genetics, Toronto Ontario, [2]Bruker Belgium SA, Kartuizersweg 3, 2550 Kontich, Belgium, [3]Bruker Scientific LLC, Billerica MA USA, [4]Bruker Daltonik GmbH, Fahrenheitstraße 4, 28359 Bremen, Germany, [5]Bruker Ltd., 2800 Highpoint Dr. Suite#206, Milton, ON L9T 6P4, Canada

Introduction:

As the field of mass spectrometry establishes itself in high-throughput, short-gradient data acquisition, post-acquisition bioinformatics has been playing catch-up. Herein we describe the modification of the open-source Sage search engine's database indexing step to allow for memory usage control of the search engine, particularly for non-specific searches as commonly used in immunopeptidomics, or for large databases common to metaproteomics. Through testing we successfully show the ability to obtain identical results showing the same biological conclusions while reducing memory footprint by 50%, allowing non-specific searches to be done on a standard workstation where it could not be completed previously due to memory availability. This reduction of memory footprint is a user-configurable option and can be optimized according to the user's hardware. To benefit the entire proteomics community this code has also been committed back to the open-source Sage repository (https://github.com/lazear/sage/pull/154).

Secondly, to obtain the best data, peptide candidates are often rescored and re-ranked using additional feature vectors in a separate step commonly referred to as re-scoring. There are several tools to rescore data including TIMS2RESCORE, AlphaPeptDeep, Oktoberfest etc., that all require running the data through a second pipeline. To circumvent this secondary pipeline, we have integrated the AlphaPeptDeep LC, IM and fragmentation models directly within Sage, along with semi-supervised learning approach, which utilizes classifiers such as XGBoost, Gradient Boosting Decision Trees (GBDT), and Support Vector Machine (SVM) that are then applied to optimize target-decoy separation, leading to more confident peptide identifications. Additionally, this Sage-Redeem workflow also implements transfer learning for fine-tuning at run time on the data being processed.

Results:

We first evaluated the performance of this workflow using a standard trypsin digested K562 lysate samples, 30 SPD acquired on timsTOF HT in-house, utilizing an entrapment FASTA file of Uniprot Human Reviewed sequences (20414 entries) as well a 1:5 entrapment sequences file (102070 entries). Top 5 candidates were reported from the database search and utilized for rescoring. The XGBoost classifier was utilized for target-decoy separation in Redeem. When DB chunking was enabled, a chunk size of 1000 was utilized. RAM usage was monitored and did not exceed 20GB as opposed to reaching >110GB in standard operation with an unspecific digest.

We next evaluated a collection of immunopeptidomics samples from Hoenisch Gravel et. al (https://doi.org/10.1038/s41467-023-42692-7). Datafiles from MHC class I enrichment from Kidney, Liver and Lung tissue were selected as representatives. The datafiles were searched with 'DB chunking' enabled and top 5 candidates being reported by from the database search and being utilized for rescoring. The ideal comparison would be to utilize AlphaPeptDeep, however the rescoring module is not yet available.

Taken together, two powerful extensions to an already powerful open-source software are presented here. The DB chunking extension enables the large search spaces, such as unspecific digestion conditions or large FASTA files … (from metaproteomics), to be searchable with modest hardware. By controlling the size of the chunks, users can optimize search time against available RAM, reducing the data-processing bottleneck for hard-to search samples and alleviate the need for expensive computers to do such processing. The second extension allows semi-supervised PSM rescoring (with sample specific fine-tuning), increasing the sensitivity, with a minimal increase in processing time. Both extension allow Sage-Redeem to be a single efficient tool for bottom-up proteomics analysis.

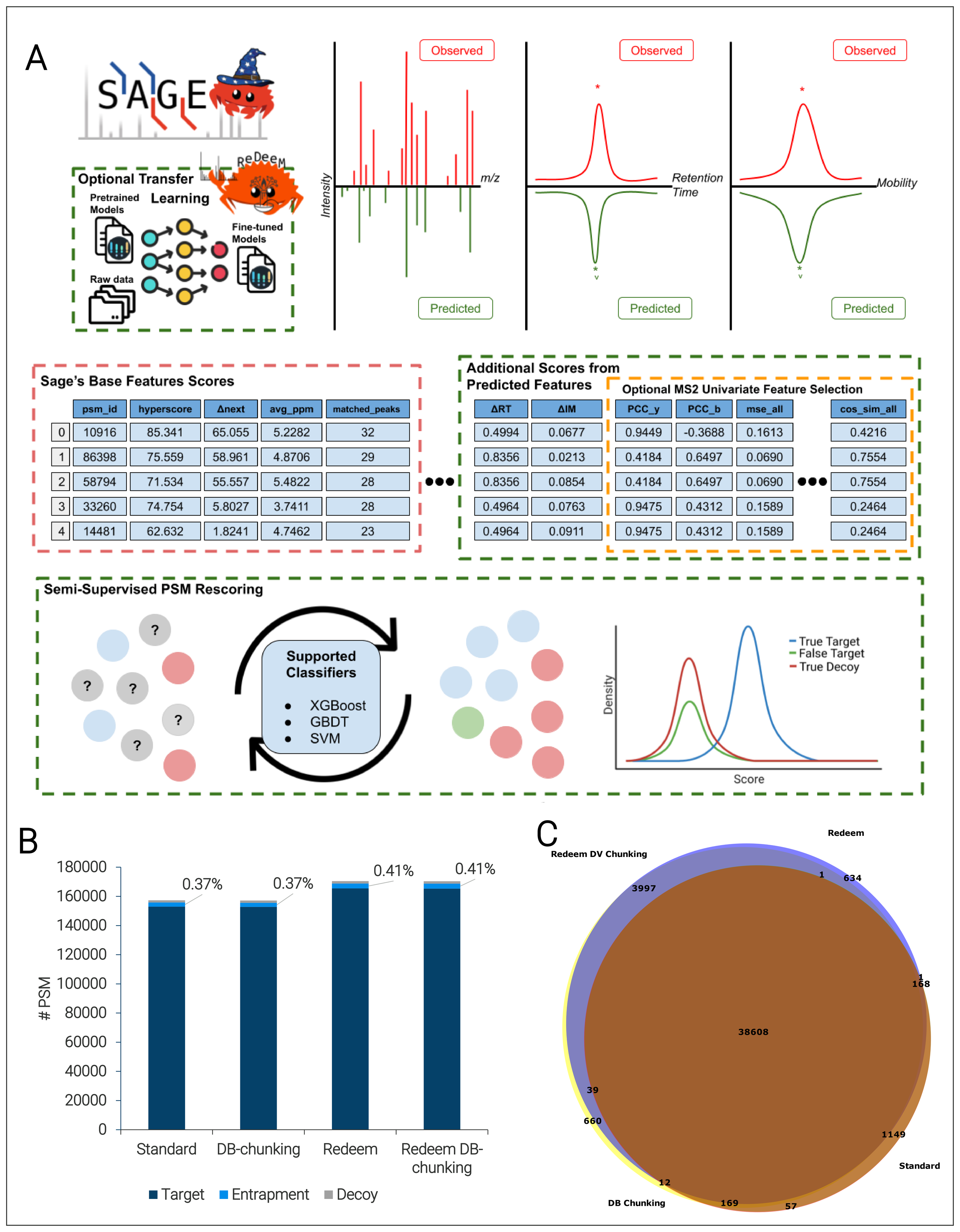
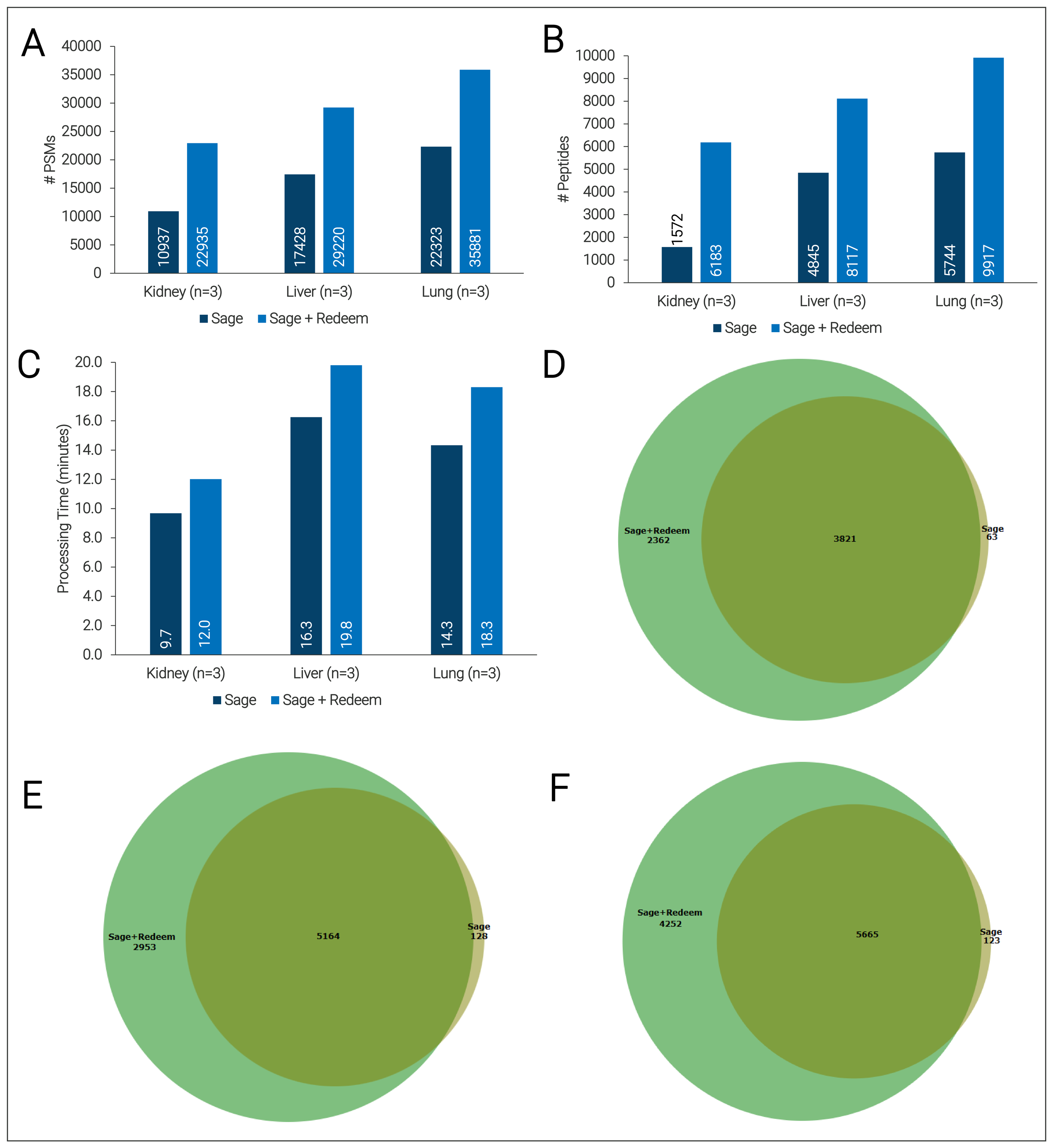


Fig. 1: (A) Overview of the integrated Sage-Redeem workflow for fast database search and semi-surprised PSM rescoring. To validate that the addition of DB chunking and Redeem did not adversely affect false discovery rate (FDR) or false discovery proportion (FDP), we searched 3 trypsin digested human lysate samples, acquired on timsTOF HT with 30SPD gradient, against entrapment FASTA file with 1:5 ratio of target proteins to entrapment proteins. PSMs were filtered to q-value 0.01 and the number of Target, Entrapment and Decoy hits tabulated (B). No significant increased in FDP was observed due to the addition of the DB chunking approach, Redeem or the combination. (C) Venn diagram shows that the peptides identified by DB chunking were nearly identical to those identified by the standard approach.

Fig. 2: MHC calls I enrichment from Kidney, Liver and Lung tissues (n=3) from Hoenish Gravel et. al was processed with Sage or Sage with Redeem. Sage with Redeem shows a small increase to processing time (C), with large gain in the number of PSMs (A) and peptides (B) being identified. Venn diagram of the identified peptides in the (D) kidney, (E) liver and (F) lung datasets showed good overlap.



Conclusion:

- Database splitting and indexing allows for efficient memory utilization for large search spaces
- Integrated feature generation utilizing retention time, ion mobility and fragmentation models for rescoring reduces overall processing time
- Sage-Redeem is a single efficient tool for sensitive database searching

Technology