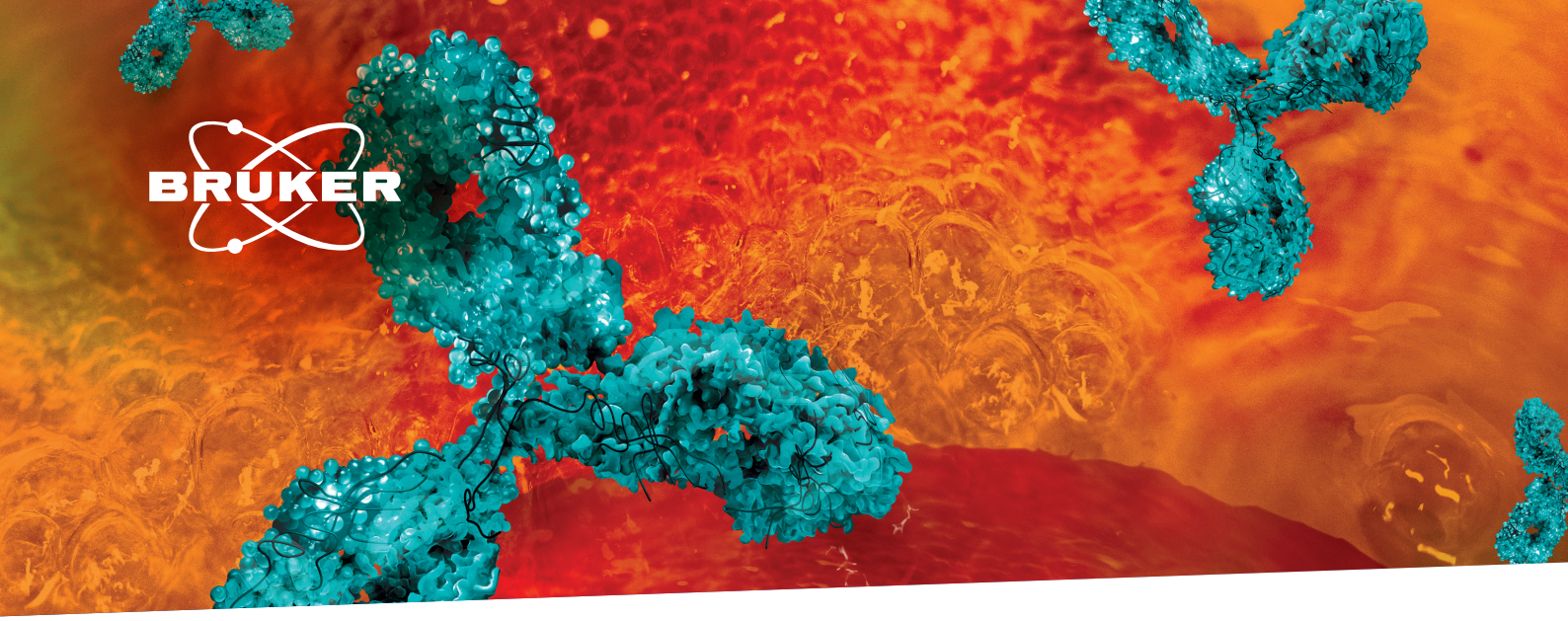


PaSER™ Novor: Real-time *de novo* sequencing for 4D-Proteomics™ applications

Peptide *de novo* sequencing, where primary amino acid sequence information is derived independent of a database, has become an invaluable tool in many research areas, especially in the field of immunopeptidomics.

Abstract

To further discoveries in immunopeptidomics, metaproteomics and other applications where *de novo* sequencing is needed, we developed PaSER Novor. Our solution provides the quickest, most accurate *de novo* sequencing results for timsTOF data with all the advantages of Run & Done.

Keywords:
Proteomics, PaSER, timsTOF, Immunopeptidomics, Peptide *de novo* sequencing, metaproteomics

Introduction

Parallel search engine in real-time or PaSER was developed to provide state-of-the-art, real-time database searching for the timsTOF instrument family. Immediate peptide assignments improve speed of analysis and allow for assessment of method development, instrument performance, sample management and decisions to be made immediately. Since conception, PaSER aims to be a comprehensive proteomics data analysis platform that can integrate third-party tools while utilizing the concept of data streaming to realize fully customizable real-time processing workflows including on-the fly decision making based on the data generated. While database searching is the preferred solution when canonical proteins are being investigated, it is difficult to execute for immunopeptidomics, metaproteomics, and other applications where enzyme specificity is lacking. Additionally, such analysis can be incomplete or restricted and computationally expensive, requiring long processing times. To address these challenges, we've integrated a timsTOF optimized *de novo* sequencing engine from Rapid Novor Inc., called PaSER Novor.

PaSER Novor is capable of processing on average >1000 spectra/second, lending itself for real-time *de novo* sequencing on the timsTOF platform. PaSER Novor consistently outperforms standard Novor and Software A on a variety of datasets. At the amino acid level, PaSER Novor achieved 51.6% recall, whereas standard Novor (pre-training) achieved 44.5%, and Software A

Rui Zhang[1], Qixin Liu[1], Mingjie Xie[1], Dennis Trede[2], Tharan Srikumar[3], Jonathan Krieger[3], Bin Ma[1], George Rosenberger[4]; [1]Rapid Novor Inc., Kitchener, Ontario, Canada; [2]Bruker Daltonics GmbH & Co. KG, Bremen, Germany; [3]Bruker Ltd., Milton, Ontario, Canada; [4]Bruker AG, Fällanden, Switzerland.

achieved 45.8% recall, respectively as assessed with 7 different sample types. Taken together, PaSER Novor is a fast, precise and accurate engine that can allow real-time *de novo* sequencing of timsTOF data on the PaSER platform.

Material and Methods

PaSER Novor was trained on a variety of timsTOF acquired data, where ground truth was taken from ProLuCID-GPU (Xu *et al.*, 2015) database search results filtered to 1% PSM FDR with DTASelect (Tabb *et al.*, 2002). The data included experiments with fixed collision energy measurements of deeply fractionated, GluC, Pepsin, Elastase, Chymotrypsin and Trypsin digested K562 lysates. Collectively, >1,780,000 PSMs were part of the training dataset utilized to optimize PaSER Novor's decision tree-based scoring functions (Ma, 2015). Training Novor on non-tryptic digests allowed learning of a generalized model, particularly suited for sequencing of non-enzymatically digested peptides.

On a PaSER workstation, the MS/MS stream is processed by PaSER Novor with results written to an output stream and to disk. We utilized offline functionality to compare PaSER Novor against other *de novo* tools across multiple datasets, including various enzyme digests, mixed species and immunopeptidomics timsTOF data as well as Novor (version 3.0; pre-timsTOF-training). We limited Novor (pre-training) and PaSER Novor to 32 compute threads to reflect the resource available to the *de novo* module from Software A. MGF files were utilized to remove the confounding effects of any pre-processing and to allow direct comparison between algorithms based on scan number matching. Amino acid level and peptide level precision and recall for each dataset were computed as in Ma, 2015.

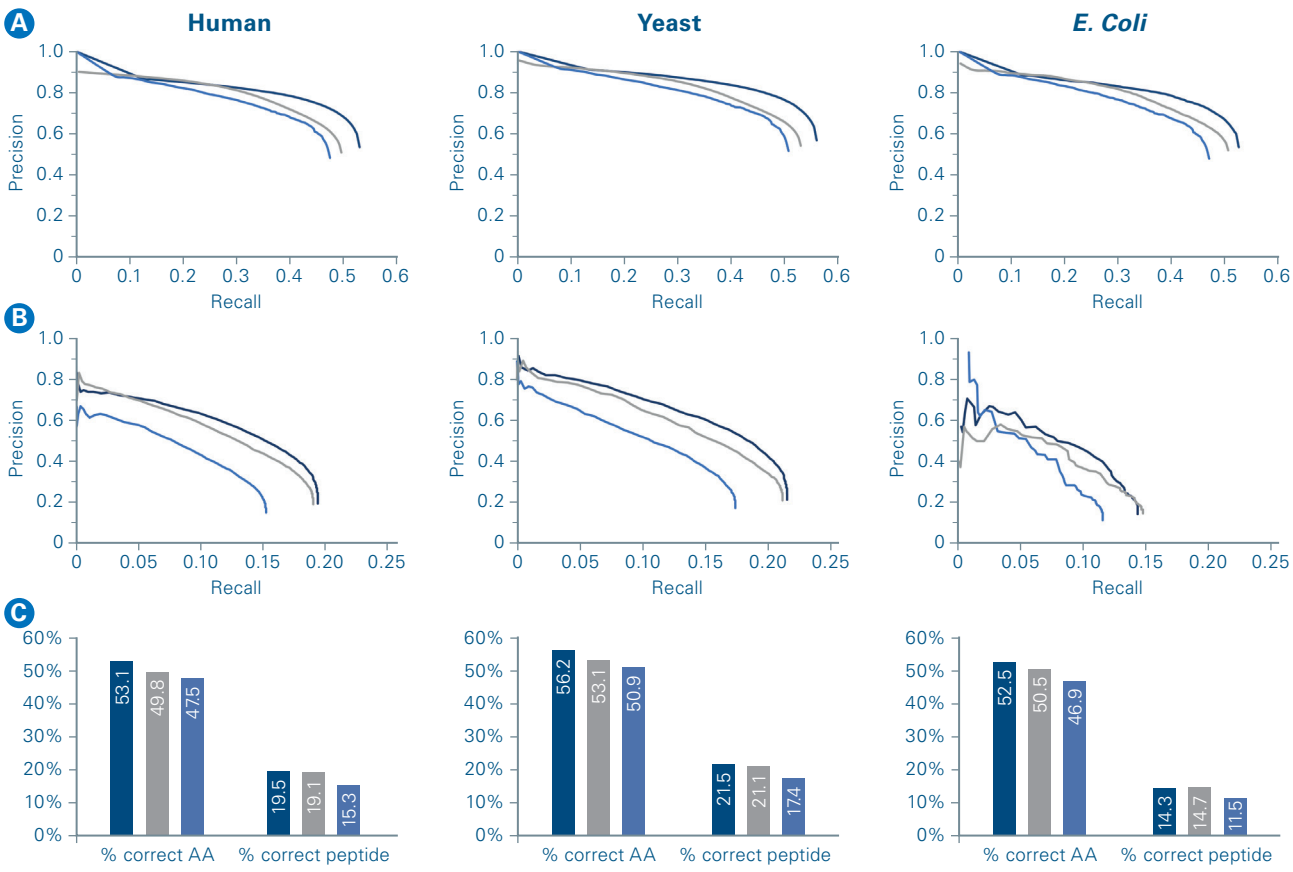


Figure 1
Amino acid (A) and peptide (B) precision recall graphs for a trypsin digested mix of Human, Yeast and *E.coli* sample. The percent of correct amino acids and peptides assigned by each algorithm is also shown (C), where all results are segmented by their species as well.

Results

We integrated the timsTOF trained *de novo* sequencing engine, PaSER Novor, into the PaSER platform, extending the capabilities of Run & Done beyond database search-based algorithms. To validate the accuracy of the timsTOF specific training we compared the results from PaSER Novor against Novor (pre-training) as well as another commonly utilized *de novo* sequencing tool built into Software A.

Species independent accuracy

To help validate the performance of PaSER Novor, we first analyzed a mixed species run (Prianichnikov *et al.*, 2020, PXD014777) that was not utilized in the prior training. This sample consisted of a mix of trypsin digested Human (65%), Yeast (15%) and *E. coli* (20%), evaluating the effects species specificity might have on PaSER Novor performance. Figure 1A depicts the amino acid precision-recall curves, Figure 1B shows the peptide precision-recall curves and Figure 1C shows the percentage of correct amino acid or peptide assignments by the three *de novo* algorithms PaSER Novor, Software A and Novor (pre-training). In all cases, the dataset was split by species. PaSER Novor showed on average an increase of 5% in correct amino acid identifications vs. software A and 11% vs. Novor algorithm pre-trained. As expected, there were no observable differences between spectra originating from different species. However, there is a substantial difference in the processing time between Novor and Software A, even when only accounting for the *de novo* modules, where PaSER Novor can support real-time *de novo* sequencing at the acquisition speed of PASEF, while Software A cannot. Small differences observed in summarized amino acid-level algorithm accuracy can have a more profound impact on the sequencing performance for individual spectra. One such difference is highlighted in Figure 2. Here, an example of PaSER Novor's (Figure 2A) and Software A's (Figure 2B) assignments vs. ProLuCID's for the same MS2 spectrum is shown. While both algorithms failed to completely sequence the complex spectra, PaSER Novor correctly identified 12/14 amino acids, whereas Software A only correctly identified 3/14 amino acids.

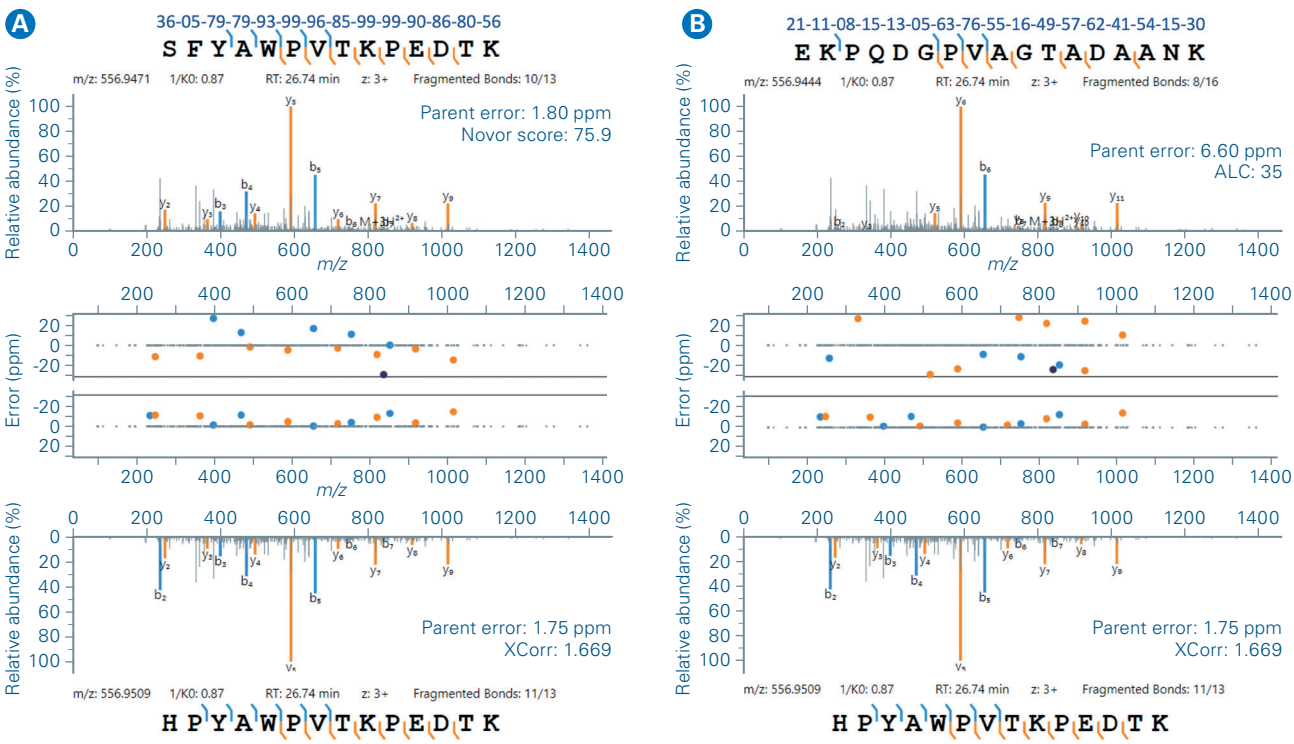


Figure 2

Two mirror plots depicting the same MS2 spectrum, and the sequence provided by PaSER Novor (Ⓐ upper) and Software A (Ⓑ upper) vs. the ground truth as identified by ProLuCID (Ⓐ &Ⓑ lower).

Enzyme independent accuracy

PaSER Novor was trained on data from a variety of non-tryptic enzymes, in addition to tryptic datasets. This resulted in a generalized model, allowing applications in a variety of fields. We next validated PaSER Novor against 4 datasets from digests of Elastase, Pepsin, GluC and Chymotrypsin. Regardless of the enzyme type, PaSER Novor and Novor are able to achieve processing speeds >1000 spectra/sec (average of 1464 spectra/sec across all 7 datasets). In contrast, Software A's processing speed is ~20x slower, with an average speed of 79 spectra/sec. Even at this speed, the re-trained PaSER Novor algorithm showed gains of 9-29% vs. Novor (pre-training) and 15-25% vs. Software A with regards to correct amino acid assignments. This translated to 38-54% gains vs. Novor (pre-training) and 50-117% vs. Software A with regards to correct peptide assignments (Figure 3C).

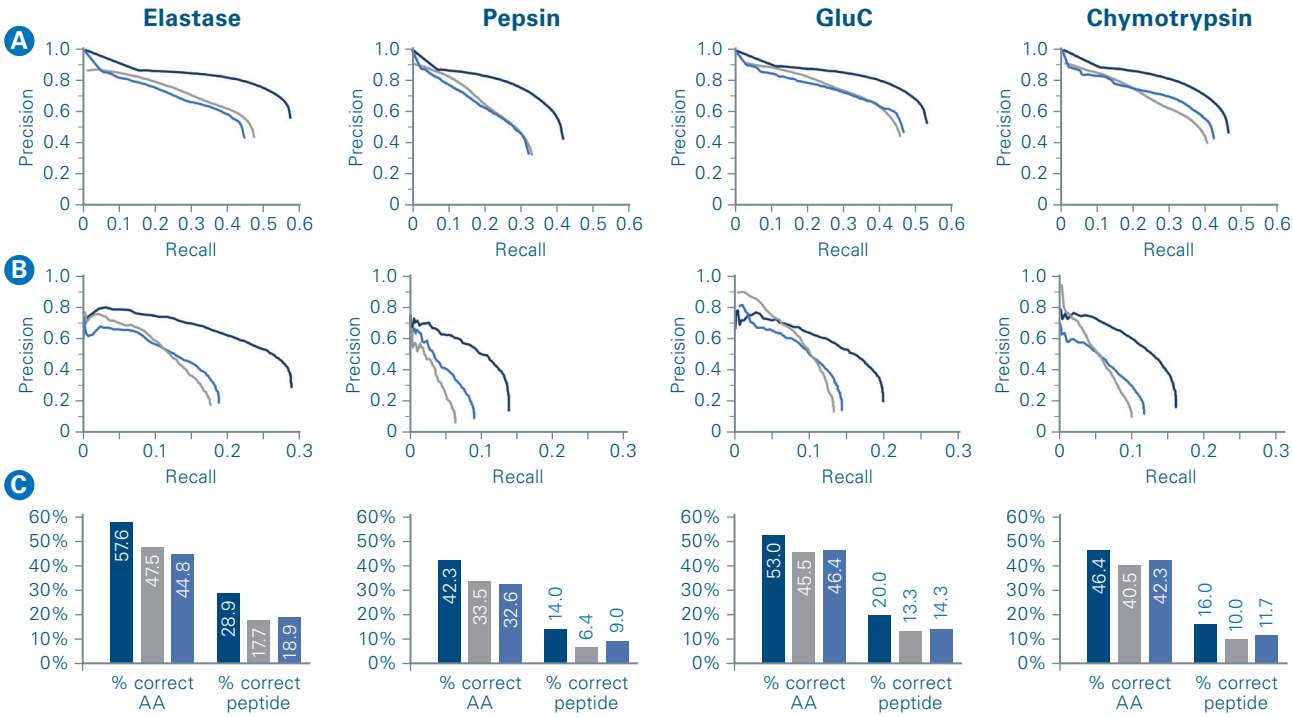


Figure 3

Amino acid (A) and peptide (B) precision recall graphs for Human lysate digested with one of 4 enzymes, Elastase, Pepsin, GluC or Chymotrypsin. The percent of correct amino acids and peptides assigned by each algorithm is also shown (C).

State-of-the-art performance for Immunopeptidomics

Immunopeptidomics is the study of naturally processed peptides bound in complex with cell surface molecules and presented to the immune system. There has been a resurgence of interest in immunopeptidomics fueled by improvements in instrument sensitivity and speed (such as the timsTOF SCP) coupled to bioinformatic workflows that frequently rely on *de novo* sequencing to improve peptide-spectrum-matches (PSMs). To evaluate the performance of PaSER Novor for these types of data, we utilized a subset of data from Feola *et al.*, 2021, retrieved via ProteomeXchange (PXD022194). As anticipated based on the results of the multi-enzyme analysis, PaSER Novor performed well with these datasets, with a noticeable increase of ~24% vs. Novor (pre-training) for correct amino acids and >25% increase for correct peptides (Figure 4). This clearly delineates the utility of optimizing the algorithm for a given platform, such as the timsTOF family. PaSER Novor also showed more modest gains of 3-10% vs. Software A for correct amino acids. To define the consensus binding motifs for all 9-mer peptides (with

Score >70) Gibbs clustering analysis was performed. Comparable results were observed between the analysis from Feola *et al.*, 2021 (Figure 4E, modified from Figure 3 in publication) and our PaSER Novor results (Figure 4F) as well as the Software A (Figure 4G) results. Two distinct groups were formed and showed the same preferences for reduced amino acid complexity for residues at positions P2 and Ω as published.

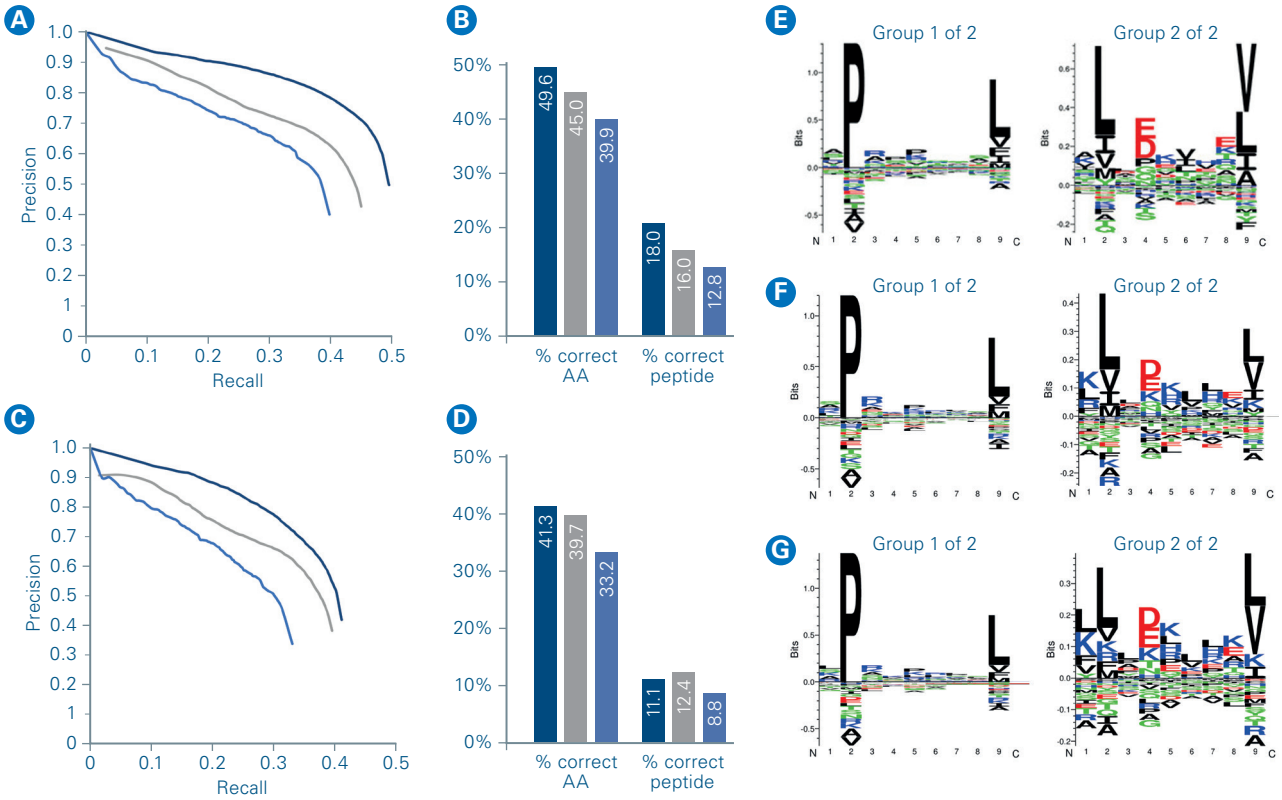
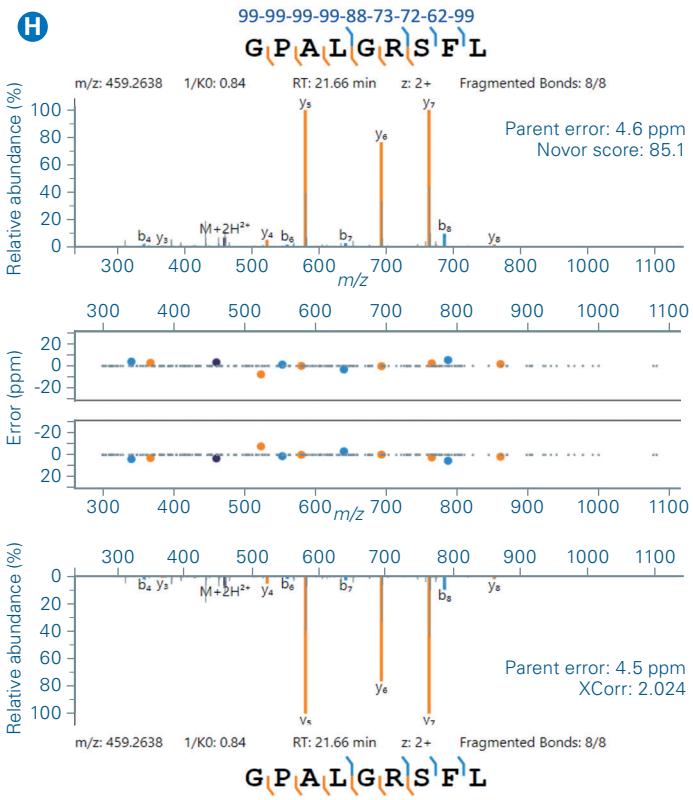


Figure 4
Amino acid (Ⓐ & Ⓒ) precision recall graphs for two MHC class I eluted samples from PXD022194. (Ⓑ & Ⓓ) The percent of correct amino acids and peptides assigned by each algorithm for these two samples respectively. Gibbs clustering analysis for the 9-mer peptides as shown in the publication Ⓔ; Feola *et al.*, 2021, identified with score >70 by PaSER Novor Ⓕ and Software A Ⓖ. An example 9-mer peptide identified by PaSER Novor and ProLuCID Ⓗ from this dataset.



Conclusion

- A fast, accurate and precise peptide *de novo* sequencing algorithm has been integrated into PaSER, providing Run & Done capabilities to additional 4D-Proteomics applications.
- PaSER Novor does not show a noticeable bias for digestion specificity or species and is 20x faster than competing products.
- Combined with PASEF technology on the timsTOF platform, PaSER Novor provides enhanced sensitivity for real-time *de novo* sequencing for a variety of applications including immunopeptidomics.

References

[1] Ma B, 2015. *Novor: Real-Time Peptide de Novo Sequencing Software.* J Am Soc Mass Spectrom **26**:1885–1894.

[2] Prianichnikov N, *et al*, 2020. *MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics.* Molecular & Cellular Proteomics **19**:1058–1069.

[3] Xu T, *et al*, 2015. *ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity.* Journal of Proteomics **129**:16–24.

[4] Tabb DL, McDonald WH, Yates JR, 2002. *DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics.* J Proteome Res **1**: 21–26.

[5] Feola S, *et al*, 2021. *PeptiCHIP: A Microfluidic Platform for Tumor Antigen Landscape Identification.* ACS Nano **15**:15992–16010.

For Research Use Only. Not for use in clinical diagnostic procedures.

Bruker Switzerland AG

Fällanden · Switzerland
Phone +41 44 825 91 11

Bruker Scientific LLC

Billerica, MA · USA
Phone +1 (978) 663-3660

ms.sales.bdal@bruker.com – www.bruker.com

Learn more at **www.bruker.com/paser**