Novor.ai – Increased precision and accuracy utilizing an AI model for de novo sequencing

Qixin Liu[1]; Mingjie Xie[1]; Dennis Trede[2]; Tharan Srikumar[3]; Jonathan Krieger[3]; Marc-Antoine Beauvais[2], Bin Ma[1]; George Rosenberger[4]

[1]Rapid Novor Inc., Kitchener, ON; [2]Bruker Daltonics GmbH & Co. KG, Bremen, Germany; [3]Bruker Ltd., Milton, ON; [4]Bruker Switzerland AG, Faellanden, Switzerland

Introduction:

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. 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. De novo sequencing with the Novor algorithm provided a fast and accurate amino acid level identifications. To further improve the accuracy and precision of de novo sequencing, a completely redesigned AI model was developed and trained. Other AI models for de novo sequencing have been developed, including Casanovo, DeepNovo, PointNovo and ∏-PrimeNovo, none have been optimized for timsTOF data or utilize the CCS information.

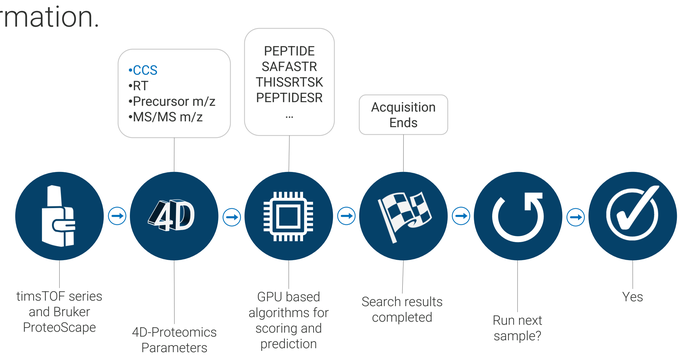


Fig. 1: Run & Done: The BPS platform allows data to be streamed during acquisition for processing on a separate workstation, allowing for real-time data analysis, so that when acquisition completes, users have data for review.

Fig. 2: Amino acid (A) and peptide (B) precision recall graphs for MHC-I dataset. (C) The percent of correct amino acids and peptides assigned by BPS Novor (dark blue) and Novor.ai (light blue) compared to ground truth results taken from combined ProLuCID + TIMS2Rescore analysis. (D-E) The percent of correct amino acids and peptides assigned by BPS Novor (dark blue) and Novor.ai (light blue) compared to ground truth results taken from combined ProLuCID + TIMS2Rescore analysis for the MHC-II RCC dataset from Hoenisch Gravel et. al.

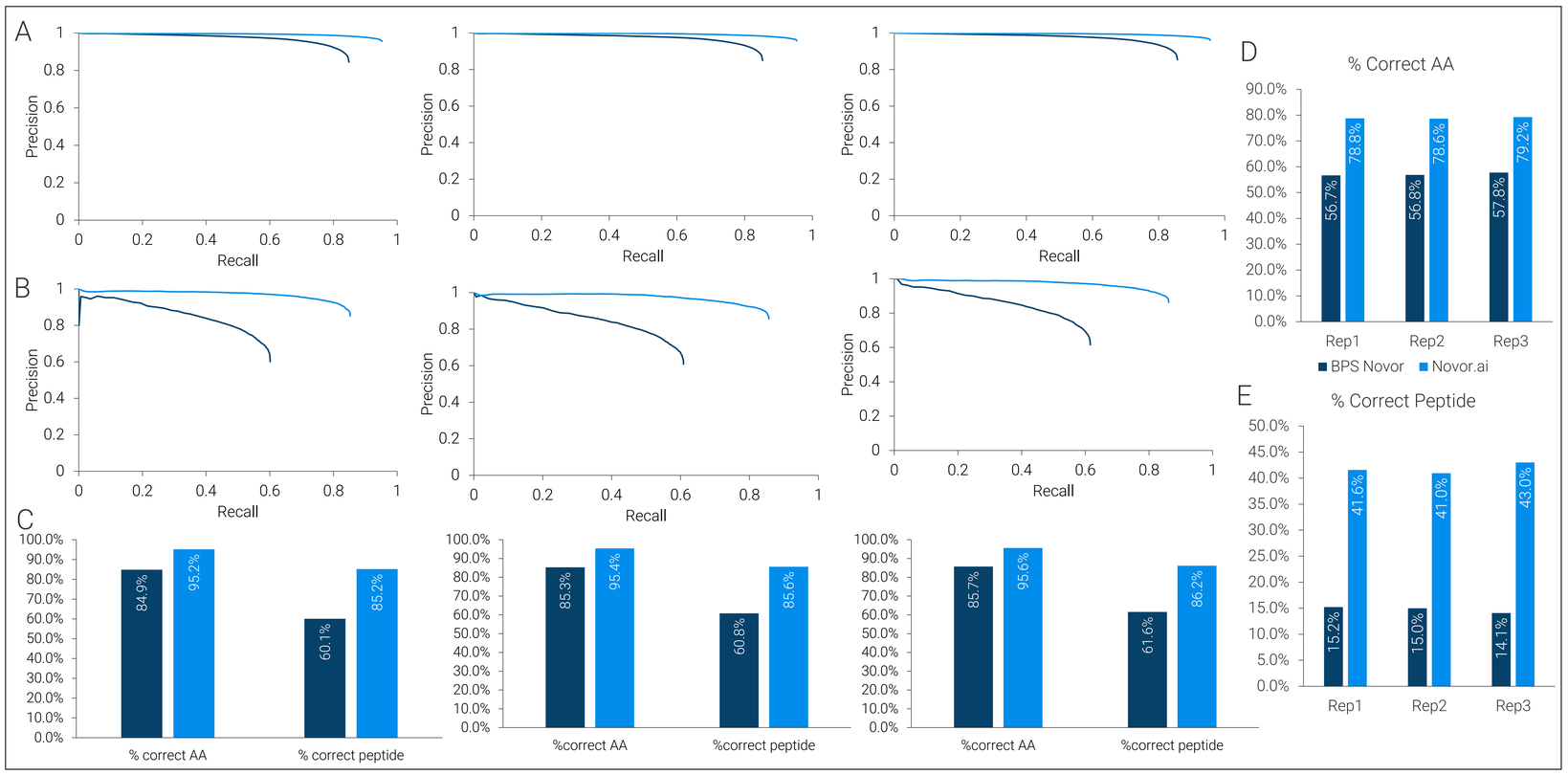
Results:

In this study, we evaluate the performance of Novor.ai model. We focused on an MHC class I and class II dataset that was recently published for validation. These samples were not previously seen by the model as part of any training or testing sets. In this dataset, Novor.ai model showed an average increase of 19% in correct amino acid identifications vs BPS Novor, the currently available timsTOF optimized algorithm. This increased amino acid accuracy was also reflected at the peptide level, where an average increase of 16% was observed.

Compared to MHC-I dataset, an average increase of 26% was observed in the MHC-II data for the percentage of correctly assigned peptide sequences. Taken together, Novor.ai provides a significant boost in the accuracy of peptides sequences assigned by de novo sequencing. Broader applicability in other application areas need to be evaluated further.

In comparison to BPS Novor, Novor.ai is slower, but we believe further model parameter optimization and model architecture refinement will decrease the processing time of Novor.ai further.

Novor.ai was not compared with other AI models for de novo sequencing, but it is planned. Many of these AI models are primally trained on Orbitrap data and minimally require transfer learning, if not re-training of the models for timsTOF data for an even-handed comparison against Novor.ai.



References:

Hoenisch Gravel, N. et al. TOFIMS mass spectrometry-based immunopeptidomics refines tumor antigen identification. Nat Commun 14, 7472 (2023).

Methods:

The original Novor algorithm was extended by replacing its decision-tree based scoring function with an AI-based one. Unlike other AI-based de novo sequencing tools, Novor.ai does not predict the sequence directly from the AI model. Instead, the AI model is only used to build the scoring functions. A separate dynamic programming algorithm as described in the original Novor paper was used to compute a sequence that maximizes the score. This approach ensures that the de novo sequences satisfy the precursor mass error tolerance of a high-resolution instrument, which both is a desired property and helps improve the overall sequencing accuracy. The model was trained with 7.3 million MS/MS spectra sampled from several different sources, including the NIST spectral library, MassIVE-KB spectral libraries, SysteMHC Atlas, and a custom MHC peptide-spectrum library generated with a timsTOf instrument. The overrepresentation of the MHC peptides in the library ensures that model's great performance on MHC peptides, whereas the inclusion of non-MHC peptides ensures the model's generalizability. The testing data included MHC class I and class II samples each was run in three replicates from Hoenisch-Gravel et. al.. Peptide-spectra matches identified by using the TIMS2rescore workflow in BPS within 1% FDR were used as the ground truth for the benchmark. In addition, peptide sequences that appear in the training datasets were excluded from the testing.

Conclusion:

- Novor.ai is an AI model based de novo sequencing algorithm
- Novor.ai was trained on >7.3 million MS2 spectra
- Novor.ai already outperforms BPS Novor with regards to precision and accuracy
- Novor.ai can be further optimized for speed and accuracy

Technology

Conflict of Interest Disclosure:

Q.L., M.X., B.M. are employees of Rapid Novor, Inc. Q.L., M.X., B.M. are co-founders of Rapid Novor, Inc. D.T., T.S., J.K., G.R. are employees of subsidiaries of Bruker Corp. Novor is a product of Rapid Novor, Inc. BPS Novor is a product of Rapid Novor, Inc. sold and distributed by subsidiaries of Bruker Corp.

© 2026 Bruker