

Improving *de novo* sequencing methods and post translational modification screening tools for the analysis of complex protein mass spectra

Mariangela Kosmopoulou¹; George Alevizos¹; Georgia Orfanoudaki¹; Dimitris Papanastasiou¹; Detlev Suckau²; Boris Krichel^{3, 4, 5, 6, 7}; Hsin-Ju Chan⁷; Charlotte Uetrecht^{8, 9}; Ying Ge^{3, 4, 7}

¹Fasmatech Science & Technology, Chalandri, Greece; ²Bruker Daltonics GmbH & Co. KG, Bremen, Germany; ³Department of Cell and Regenerative Biology, University of Wisconsin-Madison, Madison, WI; ⁴Human Proteomics Program, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI; ⁵CSSB Centre for Structural Systems Biology, Deutsches Elektronen-Synchrotron DESY & Leibniz Institute of Virology (LIV) & University of Lübeck, Hamburg, Germany; ⁶Institute of Chemistry and Metabolomics, University of Lübeck, Lübeck, Germany; ⁷Department of Chemistry, University of Wisconsin-Madison, Madison, WI; ⁸CSSB Centre for Structural Systems Biology, Deutsches Elektronen-Synchrotron DESY & Leibniz Institute of Virology (LIV) & University of Lübeck, Notkestraße, Hamburg, Germany; ⁹Institute of Chemistry and Metabolomics, University of Lübeck, Hamburg, Germany

Introduction

• Top-down mass spectrometry (TDMS) is greatly empowered by radical-driven dissociation methods, providing detailed sequence and modification-related information, complementary to collisional activation. The mass spectral complexity, however, requires sophisticated software tools to retrieve the underlying information with sufficient confidence.

• The OmniScape™ software is shown to be extremely resilient in processing congested TD mass spectra in a reliable manner.

• Here, we report on the two most recent advancements, namely an improved *de novo* sequencing algorithm and a new tool for ultra-fast post translational modification (PTM) screening.

Methods

The new *de novo* algorithm and the “PTM Screening workflow” for identifying and localizing protein modifications is developed in C++ language.

PTM screening is based on the total number of possible proteoforms for a given protein family. A brute force method is applied for a smaller search space, while a heuristic approach is adopted for an extreme number of possible protein configurations of 2^n , even with $n > 100$, achieving convergence within a few seconds.

De novo sequencing is based on deisotoping using the OmniWave™ algorithm. Calculated sequence tags are submitted to MS-BLAST (Sunyaev lab, Harvard) for homology-based protein identification and the list of scored proteins can be further interrogated using the “Confirmation workflow”.

In this work, the old and new *de novo* algorithms are evaluated by comparing the corresponding *de novo* sequence tags and assessing these relative to the expected peptide sequences originating from known proteins.

Results

The new features of the software are evaluated using TD mass spectra of AMPK-β and carbonic anhydrase II (CA II), using ScimaX MRMS and maXis II ETD instruments (Bruker), respectively.

Screening billions of proteoforms within seconds

The PTM screening workflow is optimized to process an extremely high number of protein configurations and identify the most probable one.

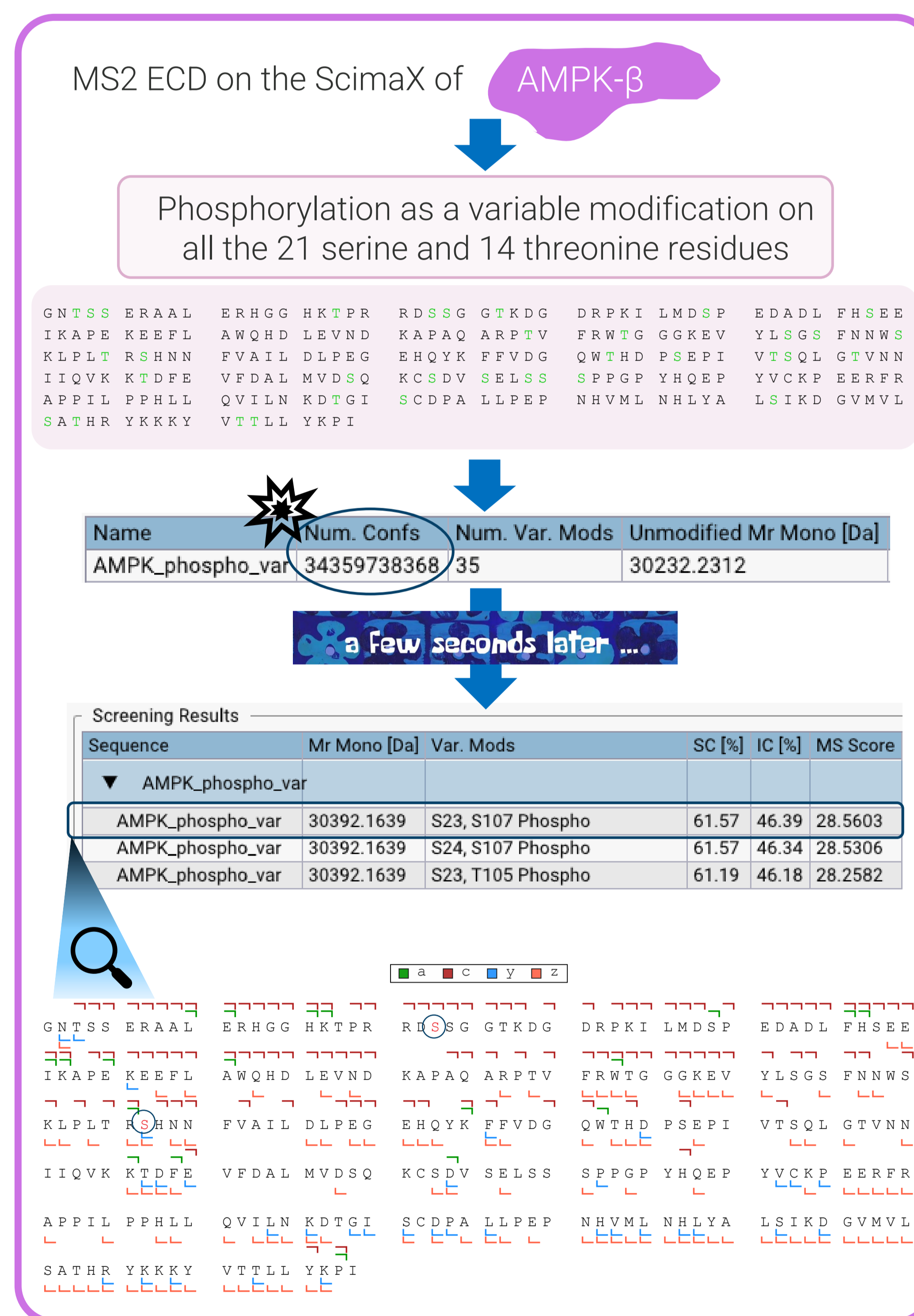


Fig. 1 PTM Screening workflow applied to 34 billion possible proteoforms, identifying the most abundant proteoform and its corresponding sequence map.

The PTM Screening workflow in OmniScape offers an effective method for PTM identification when co-isolated ions undergo fragmentation. This scenario can arise with co-isolated positional isomers or chimeric MS/MS spectra, which are generated due to the quadrupole mass filter's limited resolving capabilities.

Faster and more accurate *de novo* Sequencing



Fig. 2 Confirmation result for the ETD experiment of CA II 33+ performed on the maXis II ETD.

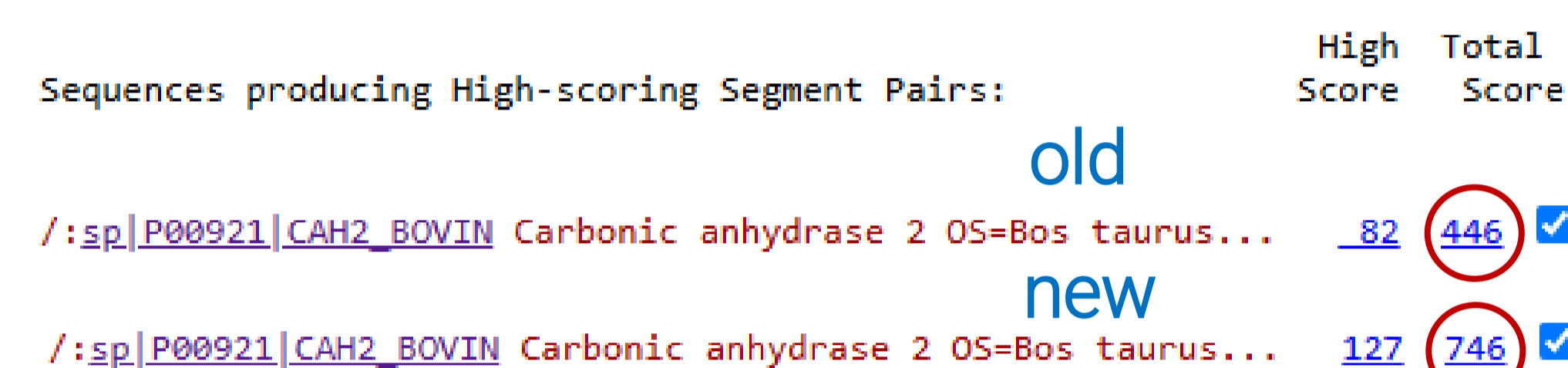


Fig. 3 MS-BLAST results based on sequence tags calculated using the previous and the new *de novo* algorithm.

CA II was the top-score protein using both the old and the new *de novo* algorithms (Fig. 3); however, the score for the latter was much improved due to the a new scoring system developed for addressing the quality of the isotopic distributions, thus providing higher-confidence sequence tags. Fig. 4 shows that the new algorithm produced longer sequence tags compared to the old one, where only fractions of the true sequence tags were found, lowering the score in MS-BLAST.

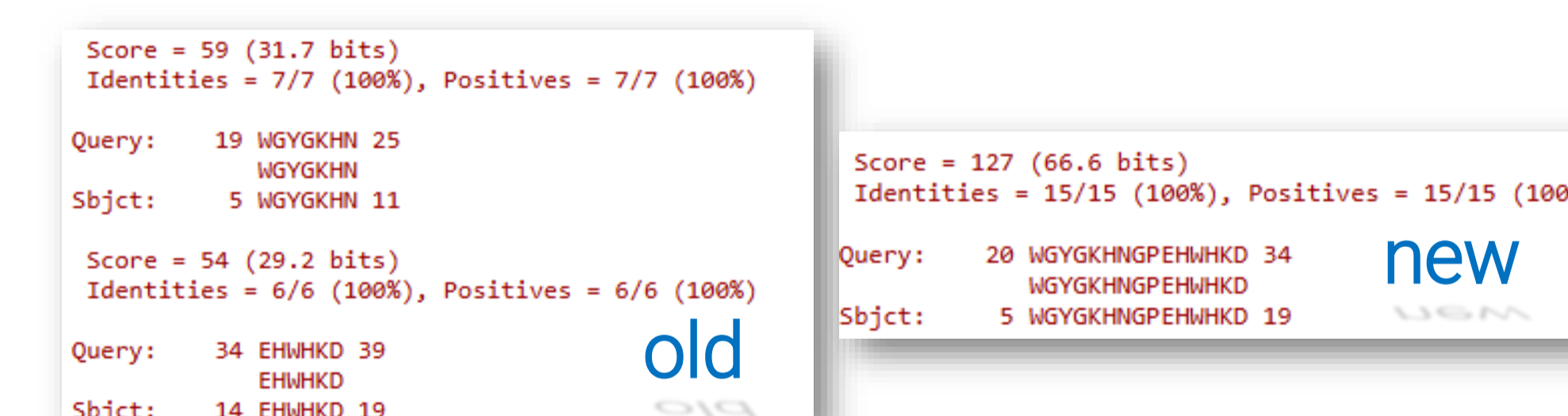


Fig. 4 Sequence tags for the CA II range W₄-D₁₈ identified using the old and the new *de novo* algorithms.

The updated method can also identify fragment ion types providing peptide directionality. For the sequence tag corresponding to L46-R57, the identification of 3 a-type ions determined the N-terminal trajectory for this tag (Fig. 5).

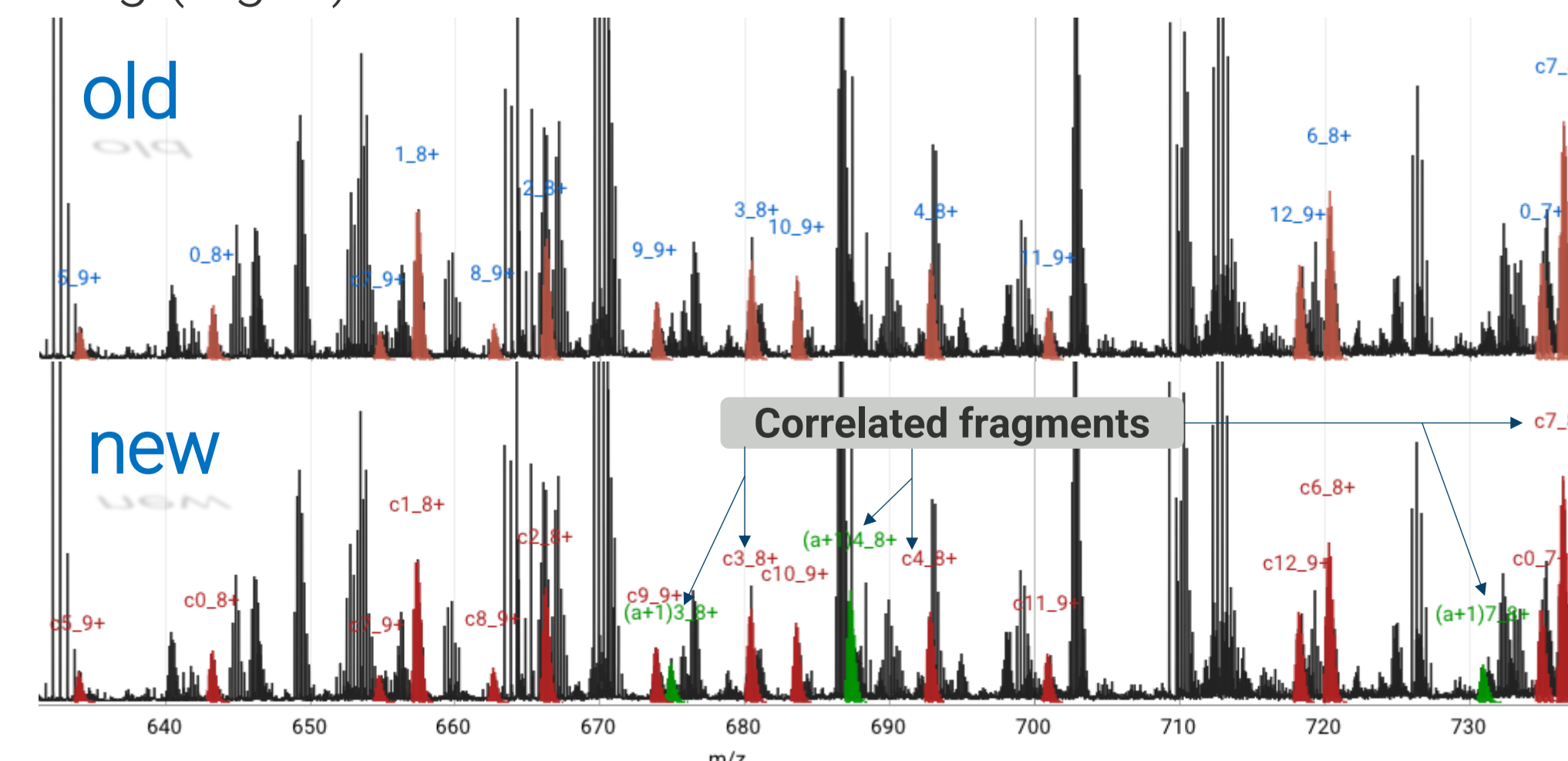


Fig. 5 Correlating fragment ion-type improves the score and protein identifications.

A graph for all isotopic distributions is constructed in the new *de novo* algorithm reflecting the distances corresponding to different relationships. These relationships include complementary neutral masses, water and ammonia neutral losses, hydrogen atom shifts and fragment-type mass differences. The greater the number of relationships associated with a specific isotopic distribution, the higher its score. Additionally, isotopic distributions detected in multiple charge states obtain a higher score.

A small part of the graph produced for the ETD experiment on CA II is shown in Fig. 6, corresponding to the L48-V49 CA II region. The sequence tag directionality is determined by correlating fragment ion-types and charge states. a and c-ion types are assigned due to their participation in the highest scored sequence tag.

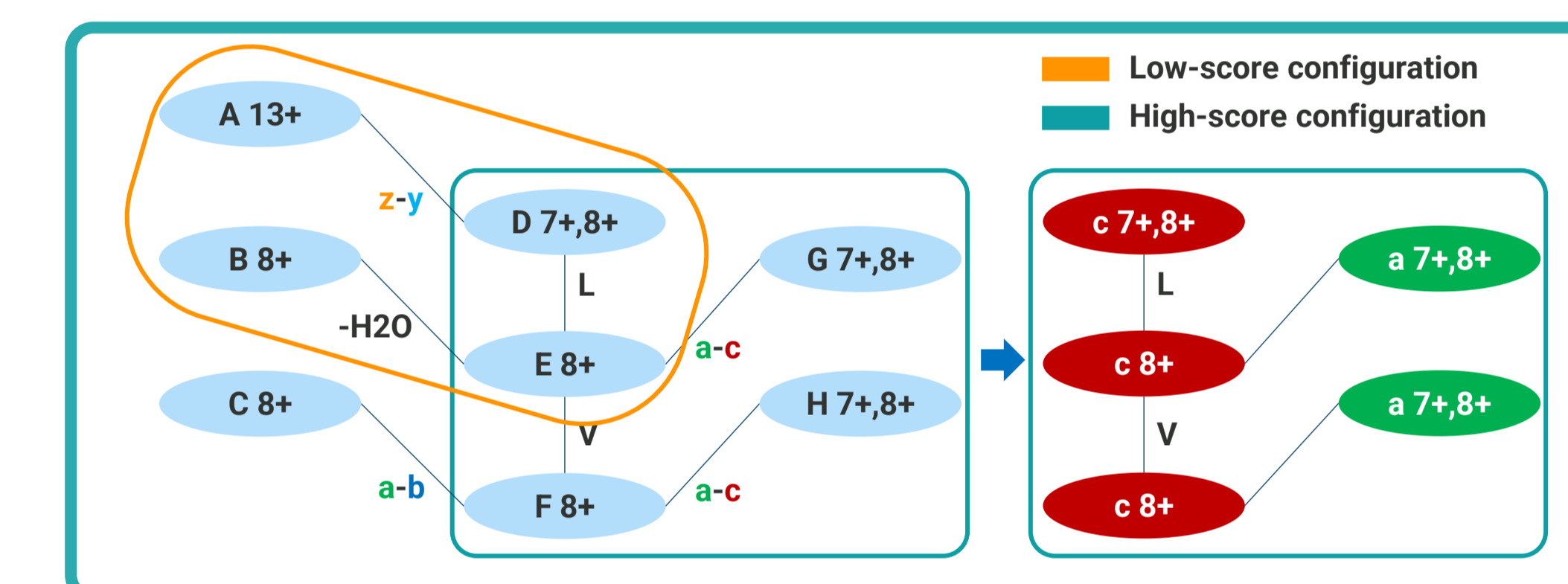


Fig. 6 A subgraph of isotopic distributions based on distances corresponding to different relationships.

Summary

- ✓ The new PTM Screening workflow significantly accelerates the localization of multiple PTMs by screening through Billions of proteoforms within seconds.
- ✓ The enhanced performance of the new *de novo* algorithm is reflected in the MS-BLAST results, returning higher confidence protein identifications, while increasing the processing speed by >10x.

Conclusion

- New ultrafast PTM Screening workflow
- *De novo* sequencing of TDMS data with
- high speed and accuracy

De novo & PTM Screening in OmniScape