Novel Data Analysis Approaches for Cross-linking Mass Spectrometry Proteomics and Glycoproteomics

By
Lei Lu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Pharmaceutical Sciences)

at the
UNIVERSITY OF WISCONSIN-MADISON
2020

Date of final oral examination: August 27th, 2020

The dissertation has been examined by the following members of the Final Oral Committee:
Lloyd M Smith, Professor, Department of Chemistry
Lingjun Li, Professor, School of Pharmacy and Department of Chemistry
Charles T Lauhon, Associate Professor, School of Pharmacy
Paul C Marker, Professor, School of Pharmacy
Aaron Hoskins, Associate Professor, Department of Biochemistry

Acknowledgements

I would like first to express my sincere gratitude to my advisor Dr. Lloyd M Smith, without whom this work would not be possible. When I joined the group in 2017, I did not have enough confidence about research and thought earning a PhD degree would be a torture. In the past three years I have been thrilled to work with varied projects in the area of algorithm and software development for proteomics data analysis. This is all because Dr. Smith provided me with full support and freedom to initiate whatever meaningful projects I was interested in and to discover my potential in scientific research. I have greatly enjoyed my biweekly talks with Dr. Smith. He tried hard to teach me 'weird' idioms and writing skills, although I can hardly remember his suggestions after our talks. What I do remember clearly is the attitude toward science he has and the way he thinks about science. He always encouraged me to try 'impossible' ideas and to pursue my curiosity and excitement. After three years, I have acquired the same philosophy he holds, that 'Science is mostly driven by tools'. Dr. Smith cares about his students on both a professional and a personal level. Now I am confident that I am well prepared for the future and I will forever be thankful for this opportunity to work under his mentorship.

I would also like to thank my committee members, Dr. Lingjun Li, Dr. Charles Lauhon, Dr. Paul Marker, and Dr. Aaron Hoskins, for their time and feedback at the preliminary examinations, annual committee meetings, and this dissertation. Your suggestions are always invaluable in helping me to become an independent scientist in my future career. Especially, I would like to thank Dr. Li for her thoughtful advice on my PhD career and life, and Dr. Lauhon for his generous help.

I would also like to thank my laboratory teammates who have helped my graduate career and contributed to my projects. Dr. Michael Shortreed introduced me to the proteomics data analysis world and mentored me in the 'fail fast' philosophy. Dr. Shortreed is always available to help the graduate students from 6:00 am every day. During my PhD work, I discussed every detail of my projects with him, and he is always so patient and provides me with potential solutions. Dr. Brian Frey helped me with my chemical synthesis project and provided suggestions for others. Zach Rolfs taught me about bottom-up proteomics search algorithms. We discussed our interesting discoveries in the office every day, and he was always the first to listen and comment on my new thoughts. Rob Millikin helped me write the first draft of my crosslink proteomics data analysis paper and helped to maintain the software module and revised a lot of my code. Katie Henke taught me the lab techniques to prepare bottom-up proteomics samples. Mark Scalf and Leah Schaffer taught me how to prepare top-down proteomics samples. Mark also taught me much about our MS instruments and helped run many samples. Stefan Solntsev showed me how to make MetaMorpheus into a robust software. I also benefitted a lot from interacting with Yunxiang Dai, Rachel Miller, Dr. Rachel Knoener, John Pavek, Isabella Whitworth and Yuling Dai. I would like also to thank my teammate Dr. Baobin Li in a previous laboratory, who taught me protein engineering techniques and helped me acquire experience with protein structural analysis.

I would like to give my thanks to my research collaborator Dr. Nick Riley. The exciting work with Nick in Dr. Carolyn Bertozzi's research group has made me even more passionate about proteomics research. I benefitted immensely from many long discussions with Nick about our joint projects and learned a great deal from the collaboration.

I would also like to acknowledge my first research advisor, Dr. Tiangang Liu for his mentorship at Wuhan University. The experience prepared me with expertise in many biological techniques and helped me to discover my interest in doing research.

Finally, I would like to thank my family who has given me unconditional love and support throughout graduate school. Most importantly, I would like to thank my wife, Xiaofang, who has accompanied me for the past eight years. We discuss interesting research and share happiness with one other. I don’t think I could have gotten through the hard times without her. She gives me full support, makes me more confident, and helps me to take myself a little less seriously. I am so grateful for the love and encouragement from all my family.

Table of Contents

Novel Data Analysis Approaches for Cross-linking Mass Spectrometry Proteomics and Glycoproteomics

Lei Lu
Under the supervision of Professor Lloyd M Smith
At the University of Wisconsin-Madison

Abstract

Bottom-up proteomics has emerged as a powerful technology for biological studies. The technique is used for a myriad of purposes, including among others protein identification, post-translational modification identification, protein-protein interaction analysis, protein quantification analysis, and protein structure analysis. The data analysis approaches of bottom-up proteomics have evolved over the past two decades, and many different algorithms and software programs have been developed for these varied purposes. In this thesis, I have focused on improving the database search strategies for the important special applications of bottom-up proteomics, including cross-linking mass spectrometry proteomics and O-glycoproteomics.

In cross-linking mass spectrometry proteomics, a sample of proteins is treated with a chemical cross-linking reagent. This causes peptides within the proteins to be cross-linked to one another, forming peptide doublets that are released by treatment of the sample with a protease such as trypsin. The data analysis tools are designed to identify the cross-linked peptides. In O-glycoproteomics, the peptides that are released by protease digestion of the protein sample can be modified with any of or even multiple distinct O-glycans, and the data analysis tools should be able to identify all of the glycans and the modification sites at which they are located. In both cases, traditional database searching strategies which try to match the experimental spectra to all potential theoretical spectra is not practical due to the large increases in search space. Researchers suffered from a lack of efficient data analysis tools for these two applications. Here we successfully devised new search algorithms to address these problems, and impemented them in two new software modules in our laboratories' bottom-up software engine MetaMorpheus (Crosslinking data analysis via MetaMorpheusXL and O-glycoproteomics data analysis via O-Pair Search).

The new search strategies used in the software program are both based on ion-indexed open search, which was first developed for large scale proteomic studies in the programs MSFragger and Open-pFind. The ion-indexed open search was optimized for cross-linking mass spectrometry proteomics and O-glycoproteomics in this study, and combined with other algorithms. In O-glycoproteomics, a graph-based algorithm is used to speed up the identification and localization of O-glycans. Other useful features have been added in the software program, such as enabling analysis of both cleavable cross-links and non-cleavable cross-links in the cross-link search module, and calculating localization probabilities in the O-glyco search module. Further optimizations including machine learning methods for false discovery rate (FDR) analysis, retention time prediction and spectral prediction could further improve the current best search approaches for cross-link proteomics and O-glycoproteomics data analysis.

Chapter 1 provides an overview of bottom-up proteomics data analysis methods and outlines how ion-indexed open search could be useful for special bottom-up proteomics studies. **Chapter 2** describes the development of a cross-linking mass spectrometry proteomics search module, resulting in efficiency improvements for both cleavable and non-cleavable cross-link proteomics data analysis. **Chapter 3** describes the development of an O-glycoproteomics search module; by combining the ion-indexed open search algorithm with the graph-based localization algorithm, the O-pair Search is more than 2000 times faster than the currently widely used software program Byonic. In **Chapter 4**, a novel top-down data acquisition method is described. **Chapter 5** provides conclusions and future directions.

Chapter 1

Introduction and Research Summary

Bottom-up proteomics[1,2] uses mass spectrometry to analyze the sequence of peptides. First, proteins are extracted from biological samples, then proteins are digested with proteases to generate peptides which generally contain 5-60 amino acids, the peptides are fragmented and analyzed by mass spectrometry, and the data are analyzed by matching the experimental spectra with theoretical spectra.

There are different types of data acquisition methods in proteomics research. Data dependent acquisition (DDA) is the major data acquisition method for discovery proteomics. In the mass spectrometer, the peptides are scanned to obtain their precursor masses in an MS1 spectrum. The instrument then randomly selects the most abundant ions for fragmentation. For each selected precursor, it will produce product ions and generate an MS2 spectrum, which will be matched to a database[3].

The idea of database searching is to match experimental spectra with theoretical spectra[3]. For the traditional database searching algorithm, we generate theoretical spectra for all potential peptides from a protein library (**Fig. 1**). For each theoretical peptide, we obtain the theoretical mass and a spectrum containing its theoretical fragment ions. In the search process, we try to match each experimental spectrum with all theoretical spectra having the same precursor mass. The best match is selected from many theoretical spectrum matches for each experimental spectrum and will be considered as an identification.

The traditional database searching strategy works well for general purpose peptide identification. However, if the search database is too large, this traditional database search strategy is not efficient and requires a excessively long time/large amount of computational power to analyze the data due to the increased number of comparisons between experimental spectra and theoretical spectra. In other words, the number of theoretical spectra to be considered as candidates is increased when the size of the database is large. For cross-linking mass spectrometry proteomics, theoretical peptide doublets need to be created for execution of the traditional database search strategy, which will exponentially increase the peptide database size. O-glycoproteomics and other variable modification settings could also dramatically increase the database size (Details will be explained in Chapter 2 and Chapter 3). Researchers developed the ion-indexed open search strategy to address this large scale database problem, as implemented in MSFragger[4], Open-pFind[5] and TagGraph[6].

In the ion-indexed open search, a lookup table is designed and peptide identifications are represented under each fragment mass (**Fig. 2**) from the original peptide database. By matching each peak in an MS2 experimental spectrum to the fragment ion-index, all peptides in the database are scored simultaneously. After the first-round matching, a fine scoring follows to find the best matched theoretical spectrum from the top candidates from the first-round match. The computational complexity of finding a list of peptide candidates is linearly proportional to the number of peaks in the spectrum.

Ion-indexed open search improves the efficiency of database search for large scale studies or special studies where the size of database increases. We proposed that the ion-indexed open search strategy could be used for cross-link proteomics data analysis and O-Glycoproteomics software program development. The idea is to find peptide candidates from the first-round matching of ion-indexed open search, and then create and match real candidates (cross-linked peptides or O-glycopeptides). Other researchers also applied ion indexed open search for varied purposes[7,8].

In cross-linking mass spectrometry proteomics, the peptide candidates identified from the ion-indexed open search are paired with each other to form theoretical doublets (**Fig. 3**). The experimental spectrum is matched with all the doublets that satisfy the precursor tolerance equation: mass of the precursor = mass of alpha peptide + mass of beta peptide + mass of the crosslinker. The best matched doublet will be considered the identified cross-linked peptides. We successfully developed MetaMorpheusXL[9,10], which applied this algorithm and is one of the most efficient cross-linking mass spectrometry proteomics data analysis tools. It is also able to analyse both cleavable and non-cleavable crosslinks.

In O-glycoproteomics, the peptide candidates identified from the ion-indexed open search are paired with glycans from a glycan database and form theoretical glycopeptides (**Fig. 4**). The experimental spectrum is matched with all theoretical glycopeptides. Then we apply an optimized graph-based algorithm for O-glycan localization analysis (Details will be explained in Chapter 4). By combining the ion-indexed open search strategy with graph-based localization, we developed the O-Pair Search software program module[11], which is 2000 times faster than the current standard program Byonic[12]. Our software program is also the first to perform localization probability calculations for O-glycopeptides.

(1) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R. Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **1997**, *69* (4), 767–776.

(2) Washburn, M. P.; Wolters, D.; Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–247.

(3) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. mass Spectrom.* **1994**, *5* (11), 976–989.

(4) Kong, A. T.; Leprevost, F. V; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **2017**, *14* (5), 513–520.

(5) Chi, H.; Liu, C.; Yang, H.; Zeng, W.-F.; Wu, L.; Zhou, W.-J.; Wang, R.-M.; Niu, X.-N.; Ding, Y.-H.; Zhang, Y. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **2018**, *36* (11), 1059–1061.

(6) Devabhaktuni, A.; Lin, S.; Zhang, L.; Swaminathan, K.; Gonzalez, C. G.; Olsson, N.; Pearlman, S. M.; Rawson, K.; Elias, J. E. TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat. Biotechnol.* **2019**, *37* (4), 469–479.

(7) Chen, Z.-L.; Meng, J.-M.; Cao, Y.; Yin, J.-L.; Fang, R.-Q.; Fan, S.-B.; Liu, C.; Zeng, W.-F.; Ding, Y.-H.; Tan, D. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **2019**, *10* (1), 1–12.

(8) Mao, J.; You, X.; Qin, H.; Wang, C.; Wang, L.; Ye, M. a new searching strategy for the identification of O-linked glycopeptides. *Anal. Chem.* **2019**, *91* (6), 3852–3859.

(9) Lu, L.; Millikin, R. J.; Solntsev, S. K.; Rolfs, Z.; Scalf, M.; Shortreed, M. R.; Smith, L. M. Identification of MS-Cleavable and Non-Cleavable Chemically Crosslinked Peptides with MetaMorpheus. *J. Proteome Res. 0* (ja), null.

(10) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced global post-translational modification discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17* (5), 1844–1851.

(11) Lu, L.; Riley, N. M.; Shortreed, M. R.; Bertozzi, C. R.; Smith, L. M. O-Pair Search with MetaMorpheus for O-glycopeptide Characterization. *bioRxiv* **2020**.

(12) Bern, M.; Kil, Y. J.; Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinforma.* **2012**, *40* (1), 13–20.

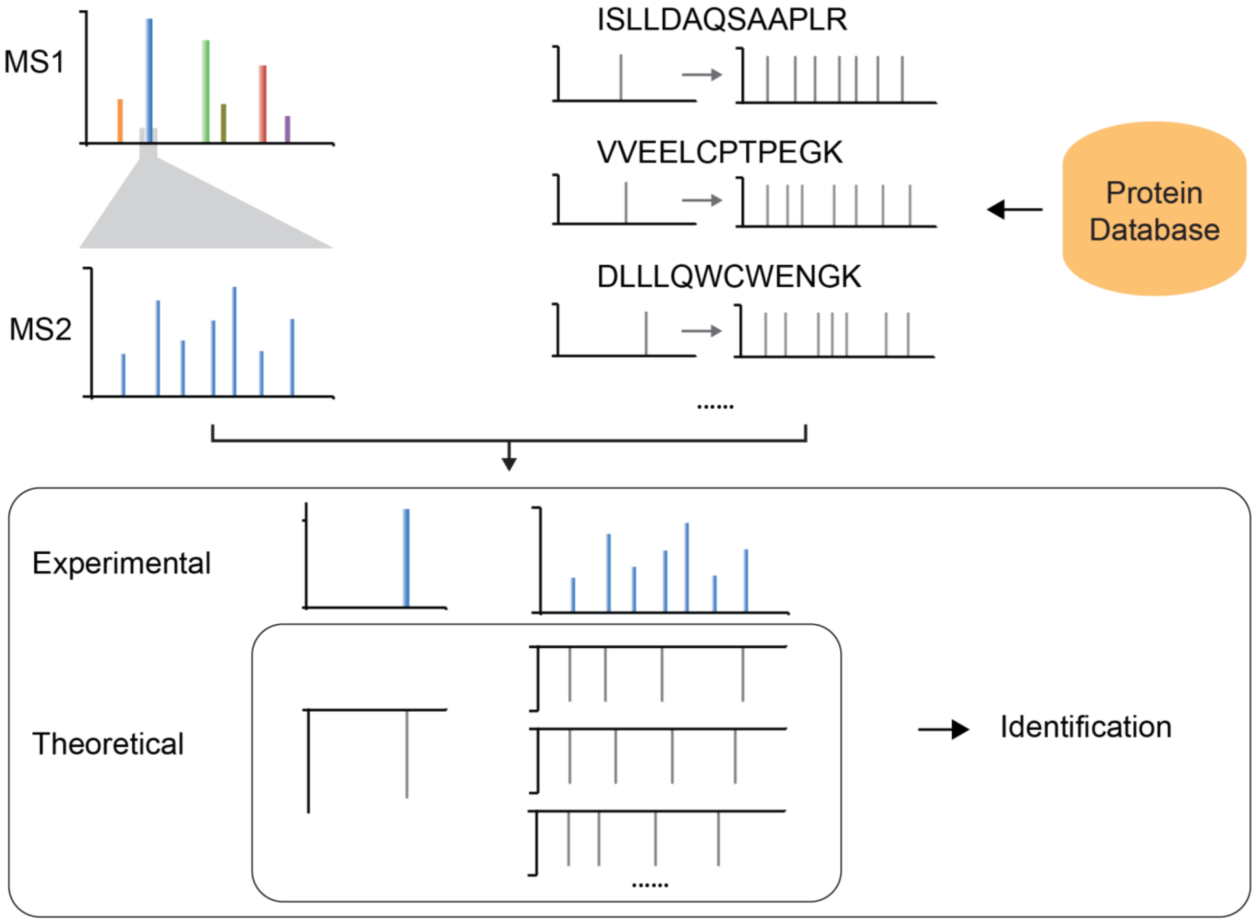


Figure 1. Traditional database search strategy. Theoretical spectra are generated based on the protein database. Each experimental MS2 spectrum will be used to match all potential theoretical spectra with the tolerated precursor mass. The best matched theoretical peptide is the target peptide.

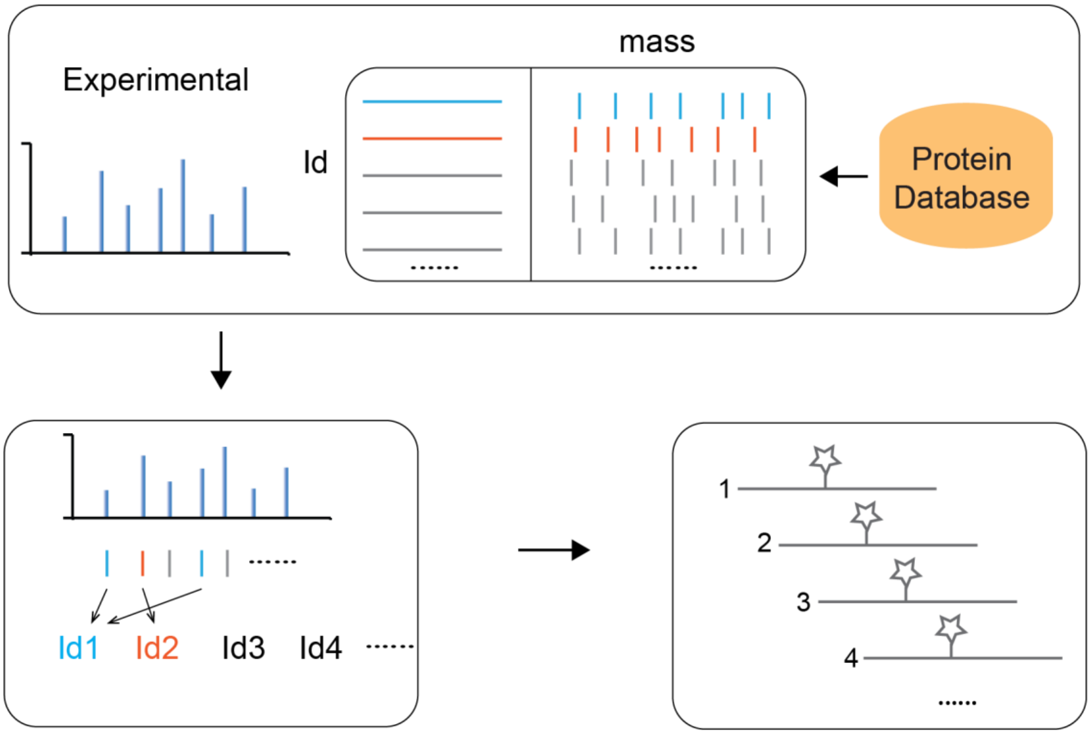


Figure 2. Ion-indexed open search strategy. A peptide-fragment lookup table is created from the protein database. In one experimental spectrum, each peak is matched to the lookup table, and the peptide candidates will be obtained simultaneously.

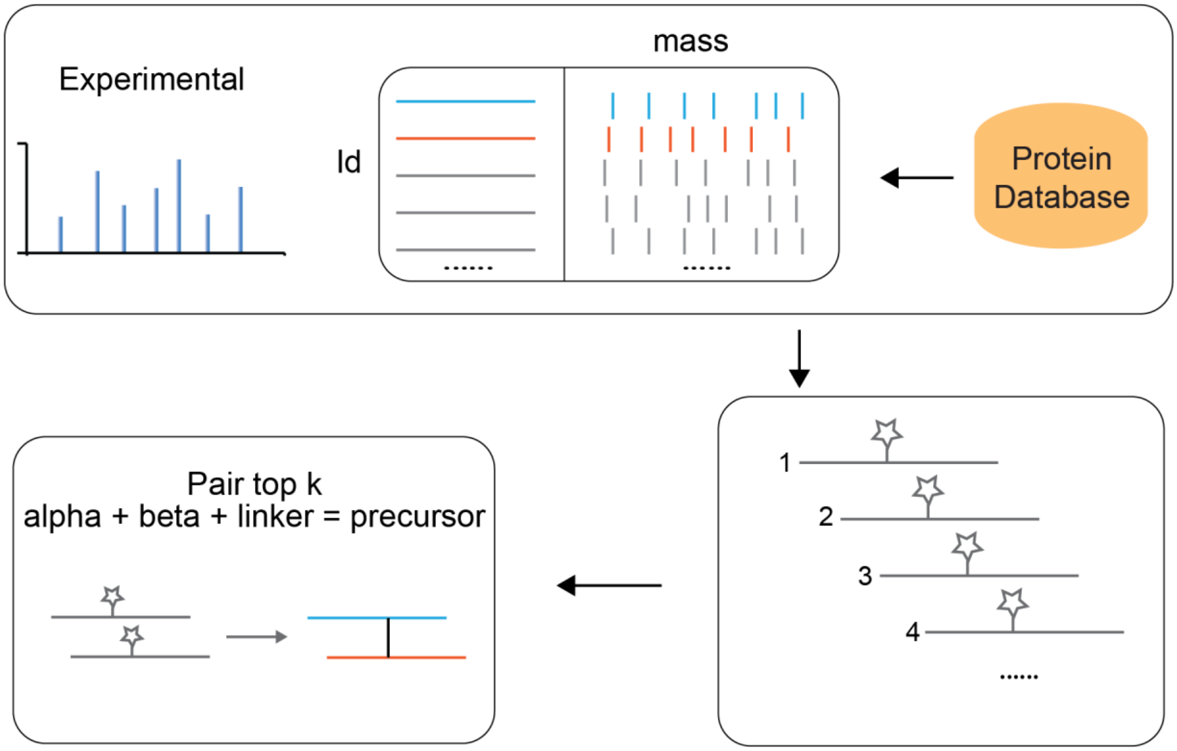


Figure 3. Cross-link search with ion-indexed open search method. Peptide candidates are used to pair with each other and all those theoretical doublets that satisfy the precursor equation will be matched with the experimental spectrum.

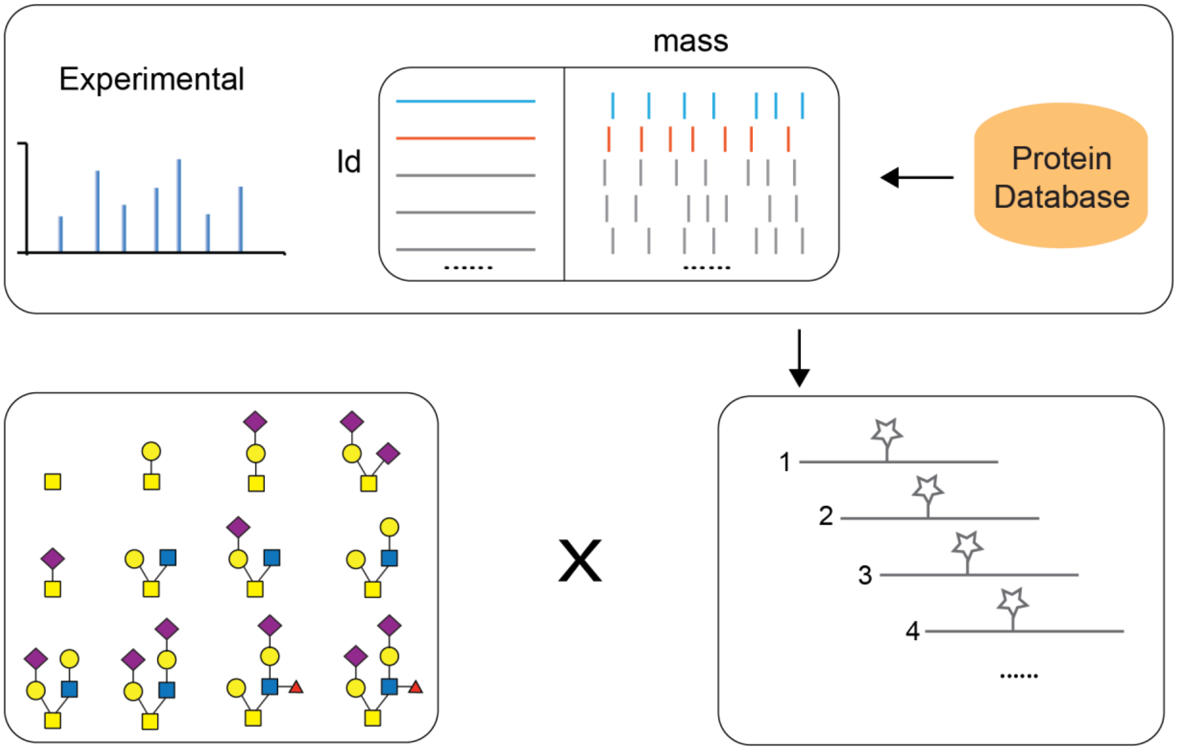
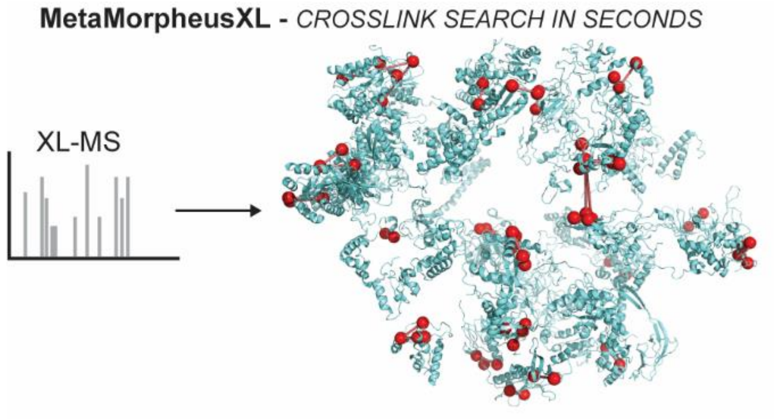


Figure 4. O-Glycopeptide search with ion-indexed open search method. Each peptide candidates will be used to combine glycan candidates to form theoretical glycopeptides. These glycopeptides will be matched with the experimental spectrum and the best match will be selected for localization analysis.

Chapter 2

Identification of MS-Cleavable and Non-Cleavable Chemically Crosslinked Peptides with MetaMorpheus



Adapted from Lu, L., Millikin, R. J., Solntsev, S. K., Rolfs, Z., Scalf, M., Shortreed, M. R., & Smith, L. M. (2018). Identification of MS-cleavable and noncleavable chemically cross-linked peptides with MetaMorpheus. Journal of Proteome Research, 17(7), 2370-2376.

ABSTRACT

Protein chemical crosslinking combined with mass spectrometry has become an important technique for the analysis of protein structure and protein-protein interactions. A variety of crosslinkers are well developed, but reliable, rapid, and user-friendly tools for large-scale analysis of crosslinked proteins are still in need. Here we report MetaMorpheusXL, a new search module within the MetaMorpheus software suite that identifies both MS-cleavable and non-cleavable crosslinked peptides in MS data. MetaMorpheusXL identifies MS-cleavable crosslinked peptides with an ion-indexing algorithm, which enables an efficient large database search. The identification does not require the presence of signature fragment ions, an advantage compared to similar programs such as XlinkX. One complication associated with the need for signature ions from cleavable crosslinkers such as DSSO (disuccinimidyl sulfoxide) is the requirement for multiple fragmentation types and energy combinations, which is not necessary for MetaMorpheusXL. The ability to perform proteome-wide analysis is another advantage of MetaMorpheusXl compared to such programs as MeroX and DXMSMS. MetaMorpheusXL is also faster than other currently available MS-cleavable crosslink search software programs. It is imbedded in MetaMorpheus, an open-source and freely available software suite that provides a reliable, fast, user-friendly graphical user interface that is readily accessible to researchers.

INTRODUCTION

Crosslinking mass spectrometry (XL-MS) has been widely used to determine protein structure and identify protein-protein interactions.[1–5] Protein chemical crosslinking uses a small-molecule bridge to form a covalent bond between two proximal amino acids, with the crosslinker's length determining the distance at which the crosslink can form. This bond-length restriction can be used to characterize protein structure or protein-protein interactions. Compared to other protein structure determination methods (X-ray crystallography, nuclear magnetic resonance and cryo-electron microscopy) that depend on challenging sample preparation procedures, protein structure determination using chemical crosslinking is much more straightforward. Crosslinked proteins are enzymatically digested (e.g., with trypsin) into peptides and analyzed by LC-MS/MS. Wang *et al.* [6] successfully employed XL-MS in conjunction with cryo-electron microscopy and computational modeling to fully resolve dynamic structures of the human 26S proteasome. In addition, they detected dynamic states of the proteasome subunits Rpn1, Rpn6 and Rpt6 and identified several new proteasome-interacting proteins. Chen *et al.* [7] used XL-MS to interpret the architecture of yeast RNA Polymerase-TFIIF complex (TFIIF is a transcription initiation factor). Despite these successes, the XL-MS search itself is still in need of significant improvement.

One limitation of some crosslink search programs (e.g. X!Link[8] and Xlink-Identifier[9]) is that they employ a database of all possible theoretical peptide dimers. The number of theoretical peptide-peptide combinations increases quadratically with database size; if a protein database contains n peptides, the possible number of peptide-peptide combinations is n(n+1)/2, which defines the search space.[2] Generally, this type of search algorithm is inefficient and not feasible for large databases because of the significant time and computational power required, and the greater chance of incorrect false-positive identifications.[10]

MS-cleavable crosslinkers such as DSSO can make the job of peptide identification more straightforward compared to non-cleavable crosslinkers. When fragmented, MS-cleavable crosslinked peptides yield two pairs of signature ions (αS/βL and αL/βS, where α (alpha) and β (beta) refer to the two crosslinked peptides and S and L refer to the "short" and "long" pieces of the fragmented crosslinker molecule) (**Figure 1a**). These ion pairs have a signature mass difference between them, which is used to help identify crosslinked peptides. The signature ions allow the search algorithm to distinguish the ion series associated with each of the individual crosslinked peptides. [11] MS-cleavable crosslinked peptides also typically generate more fragment ions than peptides connected with a non-cleavable crosslinker, thereby improving identification.[12]

Many programs have been designed for non-cleavable crosslink studies (for example xQuest[3], Plink[13], Protein Prospector[14], xi[15], ECL[16] and Kojak[17]), which separately search the alpha and beta peptides by treating one of them as a modification on the other. These programs are limited, however, to non-cleavable crosslinkers.

MeroX[18] and DXMSMS[19] were developed to be able to search cleavable crosslinks by searching theoretical dimers. MeroX avoids searching all theoretical dimers by using a DiBond algorithm[18] to reduce the number of dimer candidates, but it is still limited to only be able to search small databases. In 2015, Liu *et al.* [4] developed a search strategy (XlinkX 1.0) that searches fragmentation spectra for these signature ion pairs. The XlinkX 1.0 algorithm first finds the masses of each of the two crosslinked peptides by identifying all four signature ion peaks, followed by a standard fragment-based search to determine their sequences. However, due to the lack of signature fragment ions in many experimental spectra, this strategy leaves many crosslinked peptides unidentified. In response to these difficulties, researchers have had to use complicated, optimized fragmentation methods to maximize the chances of observing all four signature ions.

Since then, the developers of XlinkX have improved their program to avoid requiring the presence of all four signature ions in a spectrum.[12] However, the approach implemented in XlinkX 2.0 still relies on the detection of at least one signature ion of high intensity. Another problem with requiring the presence of signature ions is that some crosslinkers with bond strengths comparable to peptide amide bonds (such as DSBU, Disuccinimidyl Dibutyric Urea) make the generation of signature ions difficult.

The present work describes a novel search program for the detection of both MS-cleavable and non-cleavable crosslinked peptides and it is the first software program reported with both capabilities. The search strategy has been implemented in the computer program MetaMorpheus[20], which has a user-friendly graphical user interface (GUI). Novel crosslinker molecules are easily added if desired. A fragment-ion index scheme makes the search computationally efficient.[17,20–22] Additionally, the MetaMorpheus software suite contains multiple other functions useful to proteomics researchers such as traditional bottom-up search algorithms, mass calibration[22], post-translational modification (PTM) discovery[22,23], and label-free quantification.[20]

METHODS

Crosslink Search Algorithm

MetaMorpheusXL identifies alpha and beta crosslinked peptides with an ion-indexed open search algorithm (as outlined in **Figure 1b**). First, all fragment ions are indexed according to their *m/z* prior to searching in order to increase the speed of the search (see Supporting Information for a description of the indexing algorithm).[20,21] Then fragmentation spectra are searched against a target and decoy database[24] (decoys are generated by reversing sequences of each target protein) with an unlimited precursor mass tolerance (an “open-mass” search[21]) and all candidate peptides for each spectrum are held in memory. Second, all candidate peptides for each spectrum from the first step are paired in an attempt to find a combined mass (the two peptides plus the crosslinker mass) that matches the precursor ion mass. If such a mass match is found, that peptide pairing is considered a CSM (crosslinked peptide spectrum match) candidate. The next step is scoring these CSMs and that depends whether the crosslinker is MS-cleavable or not. If cleavable, the algorithm searches the spectrum for any signature fragment ions that could arise from the peptide pair. All possible crosslink site pairs for one CSM are considered during the generation of theoretical fragment ions; the pairing with the most fragment ion matches between theoretical and experimental is considered to emanate from correct crosslinking. This information thus informs the position of the crosslink within the pair. After attempting to match all theoretical ions from the peptide pair, the score of a CSM is the summed count of both peptides’ observed fragment ions, plus any signature fragment ions if the crosslinker was MS-cleavable. Next, the CSMs are ranked by score. The false-discovery rate (FDR) is estimated using the target-decoy strategy[24] (see Supporting Information for a description of MetaMorpheusXL’s FDR estimation). Candidate CSMs that meet a suitable FDR threshold (e.g. 1%) are considered an identified pair of crosslinked alpha and beta peptides. MetaMorpheusXL's output is a tab-delimited text file, which is readable by Percolator[17,25], a semi-supervised learning program, to potentially increase the number of target CSMs below a desired FDR. MetaMorpheusXL also formats its output into pepXML, enabling easy visualization of crosslink peptide search results with publicly available software such as ProXL[26], Kojak Spectrum Viewer and TransProteomicPipeline (TPP).[27]

Sample Preparation

BSA (Bovine serum albumin; 1µg/µL; Sigma) was dissolved in PBS (phosphate-buffered saline) buffer. Ribosomes (13.3 µM; NEB) were diluted to 1 µg/µL with HEPES buffer. Freshly prepared 50 mM MS-cleavable crosslinker DSSO (disuccinimidyl sulfoxide, Thermo Scientific) dissolved in DMSO was added to a final concentration of 1 mM. After incubating at RT for 60 min, the reaction was quenched by adding Tris Buffer to 40 mM. The samples were digested with a modified eFASP procedure as described.[28] Briefly, the crosslink reaction samples were washed with 8 M Urea, 0.1% DCA using a 30 kDa cut-off Ultrafree filter (Millipore). The samples were reduced with 20 mM DTT for 30 min, alkylated with 20 mM iodoacetamide for 60 min, and digested with 1 µg trypsin per 40 µg protein overnight at 37 ℃. The peptide digests were dried *in vacuo*, resuspended in 0.1% TFA, and desalted with C_{18} OMIX ZipTip (Agilent). The final peptides were dissolved in 95:5 H_2O/ACN with 0.2% formic acid.

Mass Spectrometry

Samples (~2 µg protein each injection) were analyzed via HPLC (NanoAcquity, Waters)-ESI-MS/MS (Q Exactive HF, ThermoFisher Scientific). The HPLC separation employed a 15 cm * 365 µm fused silica capillary micro-column packed with 3 µm diameter, 100 Å pore size C_{18} beads (Magic C_{18}; Bruker), with an emitter tip pulled to approximately 2 µm using a laser puller (Sutter instruments). Peptides were loaded on-column at a flow-rate of 400 nL/min for 30 min, then eluted over 120 min at a flow-rate of 300 nL/min with a gradient from 5% to 35% acetonitrile in 0.1% formic acid. The gradient is then ramped to 70% acetonitrile in 0.1% formic acid over 5 min and held for 5 min, then reduced to 2% acetonitrile in 0.1% formic acid over 5 min and held for 15 min. Full-mass profile scans are performed in the Orbitrap between 375 and 1,500 *m/z* at a resolution of 120,000, followed by MS/MS HCD (higher energy collisional dissociation) scans of the ten highest intensity parent ions with z > 2 at 30 CE (relative collision energy) and 15,000 resolution, with a mass range starting at 100 *m/z*. Dynamic exclusion was enabled with an exclusion window of 30 s.

Analysis of MS/MS Spectra

Single protein data of DSSO crosslinked BSA, DSSO crosslinked *E. coli* ribosome, and BS3 (bis(sulfosuccinimidyl)suberate, a non-cleavable crosslinker) crosslinked protein complex data of yeast Pol II (ProteomeXchange Dataset Identifier PXD004749) were analyzed.[7] After data-dependent acquisition, tandem mass spectral data were first calibrated using MetaMorpheus' mass calibration function[22] (see Supporting Information for a description of the calibration algorithm). The generated .mzML files were searched by MetaMorpheusXL (MetaMorpheus version 0.0.237). The small database used for DSSO crosslinked *E. coli* ribosome contained the 52 known protein sequences of ribosomal complex and another 41 protein sequences for proteins known to interact with the *E. coli* ribosome. We also searched the ribosome data against the complete *E. coli* proteome database which contains 4443 proteins. The search took ~6.5 min and resulted in 35% fewer CSMs below 1% FDR than the CMSs identified from searching against the small database containing 93 proteins. Detailed results are shown in Supplementary Figures S-2 (b). The database used for BS3 crosslinked yeast Pol II contained the 12 protein sequences of the Pol II complex. The DSSO data were also searched with XlinkX 2.0 and the BS3 data with Kojak 1.5 for comparison. The distances between each lysine-lysine pair of identified crosslinked peptides from *E. coli* ribosome and yeast Pol II were further validated by mapping to known structures with a custom python script.

RESULTS AND DISCUSSION

MetaMorpheusXL does not require the observation of signature ions

MetaMorpheusXL identifies alpha and beta peptides based on peptide fragment ions in an "open-mass" search; the observation of signature ions is not required for its MS-cleavable crosslink search. This is an advantage of MetaMorpheusXL over XlinkX, which requires high-intensity signature ions to be present in the spectrum. Other software programs such as MeroX and DXMSMS also do not require the observation of signature ions, but these programs lack the ability to search large databases.

To evaluate MetaMorpheusXL's performance for MS-cleavable crosslink searches, we first analyzed DSSO crosslinked BSA data. At 1% FDR, 513 CSMs were identified by MetaMorpheusXL, which correspond to 108 unique crosslinked peptide pairs; XlinkX 2.0 identified 104 CSMs with 39 unique crosslinked peptide pairs, 35 of which had also been found in MetaMorpheusXL (**Figure 2a**). MetaMorpheusXL found 5 times as many CSMs (3 times as many unique crosslinked peptide pairs) as XlinkX 2.0 in 1/10th the computational time (**SI Figure S-2a**).

Of the 513 CSMs identified by MetaMorpheusXL, 34 contained all 4 signature ions; 85 contained 3 signature ions, 170 contained 2 signature ions, 139 contained 1 signature ion and 85 contained no signature ions (**Figure 2b**). A majority of CSMs thus had fewer than four signature ions. CSMs containing zero signature ions were not detected by XlinkX 2.0, but many were detected by MetaMorpheusXL, as the peptide fragment ions provided sufficient information for characterization.

We also examined the intensity distribution of the signature ions. XlinkX 2.0 requires one signature ion to be among the most intense fragment peaks for each CSM. However, the intensities of signature ions depend on the type of crosslinker, MS instrumentation, and acquisition parameters. With the MS method we used for DSSO crosslinked BSA (HCD with 30 CE), we found that although roughly 50% of the most intense signature ions are among the top 15 most intense peaks in a spectrum, about 25% are not in the top 30 peaks, and 2.5% are not in the top 100 (**Figure 2c**). Because MetaMorpheusXL is not dependent on matching signature ions first, it provides a wider variety of choices for crosslinkers and acquisition parameters compared to other algorithms.

DSSO crosslinked ribosome analysis

We also generated DSSO crosslinked *E. coli* ribosomal complex data to further validate that MetaMorpheusXL could be used for protein complexes. The *E. coli* ribosome contains 52 proteins, and a wealth of detailed structural data exists to assist in validating the search results. Experimental replications were not performed here. In addition, no sample enrichments or prefractionation steps were employed.

At 1% FDR, MetaMorpheusXL found 262 CSMs including 46 inter- and 216 intra-CSMs (inter: crosslinked peptides from different proteis; intra: crosslinked peptides from a single protein); XlinkX 2.0 identified 49 CSMs including 28 inter-and 21 intra-CSMs. The pepXML output of the ribosome crosslinks were visualized by ProXL[26] (**SI Figure S-3**). MetaMorpheusXL detected 5 times more CSMs than XlinkX 2.0, similar to the results for DSSO crosslinked BSA data. In total, MetaMorpheusXL identified 77 unique crosslinked peptide pairs, while XlinkX 2.0 identified 31 (**Figure 3a**).

We further investigated the results by mapping the identified crosslink residues to the known structure of the ribosomal complex. Distances between the Cα carbon of crosslinked residues were expected to be within 30 Å based upon the crosslinker spacer arm length and structural flexibility considerations. In total, 171 of the 262 CSMs identified by MetaMorpheusXL could be mapped to the *E. coli* ribosome structure (PDB: 3jcd). 4 unique crosslinked peptide pairs (from 11 CSMs) had distances greater than the 30 Å restriction; the distances of the rest of the CSMs fell within the expected 30 Å (**Figure 3b**, **Figure 4**). Of the 4 unique crosslinked peptide pairs, two of them (32.7 Å and 33.9 Å) were very close to 30 Å. The other two contained an intra-crosslink with distance of 40.1 Å, and an inter-crosslink with distance of 38.2 Å. The structure suggested that these four crosslinks may occur due to flexibility in the protein structure (**Figure 4**). Additionally, because they were identified multiple times and shared crosslinked residues with other crosslinked peptides, they are likely to be true crosslinked peptides. For the results produced by XlinkX 2.0, 21 of the 31 unique crosslinked peptide pairs could be mapped to the ribosome structure, 2 had distances larger than the 30 Å restriction, and the large distances 56.9 Å and 139 Å are not likely to be transient crosslinks. More intra-CSMs could be detected than inter-CSMs, consistent with the known structure of the ribosome (the 52 proteins of the ribosome being dispersed around the ribosomal RNA).

We further analyzed the distribution of number of signature ions observed from identified CSMs and the relationship between the number of CMSs and intensity ranks of the most intense signature ions from the MetaMorpheusXL results. Similar to the intensity distribution of signature ions for BSA, the majority of CSMs lack 2 or 3 signature ions and about 5% of CSMs have no signature ion matches (**Figure 5a**). The rank distribution of the most intense signature ions also followed a similar pattern to that obtained for BSA: 75% of the most intense signature ions found in the most intense 30 peaks of a spectrum (**Figure 5b**).

Because MetaMorpheusXL uses fragment-ion indexing prior to searching, it was faster than XlinkX 2.0 for both the BSA data and the ribosomal complex data (**SI Figure S-2a**).

Non-cleavable crosslink search with MetaMorpheusXL

MetaMorpheusXL can also be used for identification of peptides with non-cleavable crosslinkers. A high-quality dataset of yeast Pol II complex acquired by Chen *et al.* was used for this study.[7] The dataset contains 4 raw files from peptide digests of the 12 subunits of yeast Pol II (523 kDa) crosslinked with crosslinker BS3.

MetaMorpheusXL identified five types of PSMs in this dataset: intra-protein crosslinks, inter-protein crosslinks, loop-links, dead-end-links and single peptide PSMs. MetaMorpheusXL identified 2075 PSMs (peptide spectrum matches) with ~1% FDR from the yeast Pol II data. About 76% of the PSMs were single peptide PSMs, which is consistent with other crosslinking studies, and about 11% of the PSMs were loop-links (crosslinked residues are in the same peptide) or dead-end-links (crosslinker reacted with one residue). In total 277 CSMs were identified (168 intra-CSMs and 109 inter-CSMs), including 97 unique crosslinked peptide pairs.

The search speed of MetaMorpheusXL for non-cleavable crosslinks was also assessed. Kojak is one of the most efficient software tools for crosslink studies. The algorithms of Kojak and MetaMorpheusXL yielded comparable search times (data not shown).

To assess the crosslinks identified with MetaMorpheusXL, we analyzed the yeast Pol II data in parallel with Kojak, and compared with the original result from the paper which used a program named Xi.[7] Kojak identified 315 CSMs (109 unique crosslinked peptide pairs) at 1% FDR, the paper reported 287 CSMs (105 unique crosslinked peptide pairs) with high confidence, and MetaMorpheusXL identified 277 CSMs (97 unique crosslinked peptide pairs). Among the unique crosslinked peptide pairs identified by all three programs, 60 were common to all three analyses (**Figure 6a**).

We then validated the crosslinked residues by comparison with the published crystal structure data (PDB: 1wcm, **Figure 6b**). In the MetaMorpheusXL result, 227 of all identified CSMs fell in regions consistent with the published structure, while 5 (2.2%) were not within the 30 Å cutoff. The level of structural agreement was better than for Kojak (13 of 284, 4.6% not within cutoff) or Xi (11 of 214, 5.1% were not within cutoff).

CONCLUSIONS

MetaMorpheusXL is a new search algorithm designed for large scale studies of chemically crosslinked peptides. The approach increases the number of identified MS-cleavable crosslinked peptides compared to existing software. The algorithm is readily compatible with any crosslinker and will benefit developers and researchers who want to test the performance of different crosslinkers. MetaMorpheusXL has publicly available source code (https://github.com/smith-chem-wisc/MetaMorpheus). The software is user friendly and the search results can easily be pipelined to downstream software (e.g. Percolator *et al.*).

ACKNOWLEDGEMENTS

We thank Brian L. Frey, who helped edit the manuscript. We thank Anthony J. Cesnik and Leah V. Schaffer, who also contributed daily input and guidance to the improvement of MetaMorpheus. We thank Prof. Lingjun Li’s lab for supplying EThcD (Electron-Transfer/Higher-Energy Collision Dissociation) fragmented crosslinked BSA data.

REFERENCES

(1) Liko, I.; Allison, T. M.; Hopper, J. T.; Robinson, C. V. Mass spectrometry guided structural biology. *Curr. Opin. Struct. Biol.* **2016**, *40*, 136–144.

(2) Barysz, H.; Malmström, J. Development of large-scale cross-linking mass spectrometry. *Mol. Cell. Proteomics* **2017**, 1–32.

(3) Rinner, O.; Seebacher, J.; Walzthoeni, T.; Mueller, L.; Beck, M.; Schmidt, A.; Mueller, M.; Aebersold, R. Identification of cross-linked peptides from large sequence databases. *Nat. Methods* **2008**, *5* (4), 315.

(4) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* **2015**, *12* (12), 1179–1184.

(5) Yu, C.; Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **2018**, *90* (1), 144–165.

(6) Wang, X.; Cimermancic, P.; Yu, C.; Schweitzer, A.; Chopra, N.; Engel, J. L.; Greenberg, C.; Huszagh, A. S.; Beck, F.; Sakata, E.; et al. Molecular details underlying dynamic structures and regulation of the human 26S proteasome. *Mol. Cell. Proteomics* **2017**, *16* (5), 840–854.

(7) Chen, Z. A.; Jawhari, A.; Fischer, L.; Buchen, C.; Tahir, S.; Kamenski, T.; Rasmussen, M.; Lariviere, L.; Bukowski-Wills, J.-C.; Nilges, M.; et al. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **2010**, *29* (4), 717–726.

(8) Young, J. L.; Lackner, L. L.; Nunnari, J. M.; Phinney, B. S. Shotgun cross-linking analysis for studying quaternary and tertiary protein structures. *J. Proteome Res.* **2007**, *6* (10), 3908–3917.

(9) Du, X.; Chowdhury, S. M.; Manes, N. P.; Wu, S.; Mayer, M. U.; Adkins, J. N.; Anderson, G. A.; Smith, R. D. Xlink-identifier: an automated data analysis platform for confident identifications of chemically cross-linked peptides using tandem mass spectrometry. *J Proteome Res* **2011**, *10* (3), 923–931.

(10) Ji, C.; Li, S.; Reilly, J. P.; Radivojac, P.; Tang, H. XLSearch: A probabilistic database search algorithm for identifying cross-linked peptides. *J. Proteome Res.* **2016**, *15* (6), 1830–1841.

(11) Kao, A.; Chiu, C.; Vellucci, D.; Yang, Y.; Patel, V. R.; Guan, S.; Randall, A.; Baldi, P.; Rychnovsky, S. D.; Huang, L. Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes. *Mol. Cell. Proteomics* **2011**, *10* (1), M110.002212.

(12) Liu, F.; Lössl, P.; Scheltema, R.; Viner, R.; Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* **2017**, *8* (May), 15473.

(13) Yang, B.; Wu, Y.-J.; Zhu, M.; Fan, S.-B.; Lin, J.; Zhang, K.; Li, S.; Chi, H.; Li, Y.-X.; Chen, H.-F.; et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **2012**, *9* (9), 904–906.

(14) Trnka, M. J.; Baker, P. R.; Robinson, P. J. J.; Burlingame, A. L.; Chalkley, R. J. Matching Cross-linked Peptide Spectra: Only as Good as the Worse Identification. *Mol. Cell. Proteomics* **2014**, *13* (2), 420–434.

(15) Fischer, L.; Chen, Z. A.; Rappsilber, J. Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *J. Proteomics* **2013**, *88*, 120–128.

(16) Yu, F.; Li, N.; Yu, W. ECL: an exhaustive search tool for the identification of cross-linked

peptides using whole database. *BMC Bioinformatics* **2016**, *17* (1), 1–8.

(17) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; Maccoss, M. J.; Davis, T. N.; Moritz, R. L.; Michael, J.; Davis, T. N.; Moritz, R. L. Kojak: Efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* **2015**, *14* (5), 2190–2198.

(18) Götze, M.; Pettelkau, J.; Fritzsche, R.; Ihling, C. H.; Schäfer, M.; Sinz, A. Automated assignment of MS/MS cleavable cross-links in protein 3d-structure analysis. *J. Am. Soc. Mass Spectrom.* **2014**, *26* (1), 83–97.

(19) Petrotchenko, E. V.; Borchers, C. H. ICC-CLASS: Isotopically-coded cleavable crosslinking analysis software suite. *BMC Bioinformatics* **2010**, *11* (1), 64.

(20) Millikin, R. J.; Solntsev, S. K.; Shortreed, M. R.; Smith, L. M. Ultrafast Peptide Label-Free Quantification with FlashLFQ. *J. Proteome Res.* **2018**, *17* (1), 386–391.

(21) Kong, A. T.; Leprevost, F. V; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **2017**, *14* (5), 513–520.

(22) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17* (5), 1844–1851.

(23) Cesnik, A. J.; Shortreed, M. R.; Sheynkman, G. M.; Frey, B. L.; Smith, L. M. Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy. *J. Proteome Res.* **2016**, *15* (3), 800–808.

(24) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214.

(25) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923–925.

(26) Riffle, M.; Jaschob, D.; Zelter, A.; Davis, T. N. ProXL (protein cross-linking database): A platform for analysis, visualization, and sharing of protein cross-linking mass spectrometry data. *J. Proteome Res.* **2016**, *15* (8), 2863–2870.

(27) Hoopmann, M. R.; Mendoza, L.; Deutsch, E. W.; Shteynberg, D.; Moritz, R. L. An Open Data Format for Visualization and Analysis of Cross-Linked Mass Spectrometry Results. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (11), 1728–1734.

(28) Erde, J.; Loo, R. R. O.; Loo, J. A. Improving proteome coverage and sample recovery with enhanced FASP (eFASP) for quantitative proteomic experiments. *Methods Mol. Biol.* **2017**, *1550*, 11–18.

(29) Vizcaíno, J. A.; Côté, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **2012**, *41* (D1), D1063--D1069.

FIGURES

1)

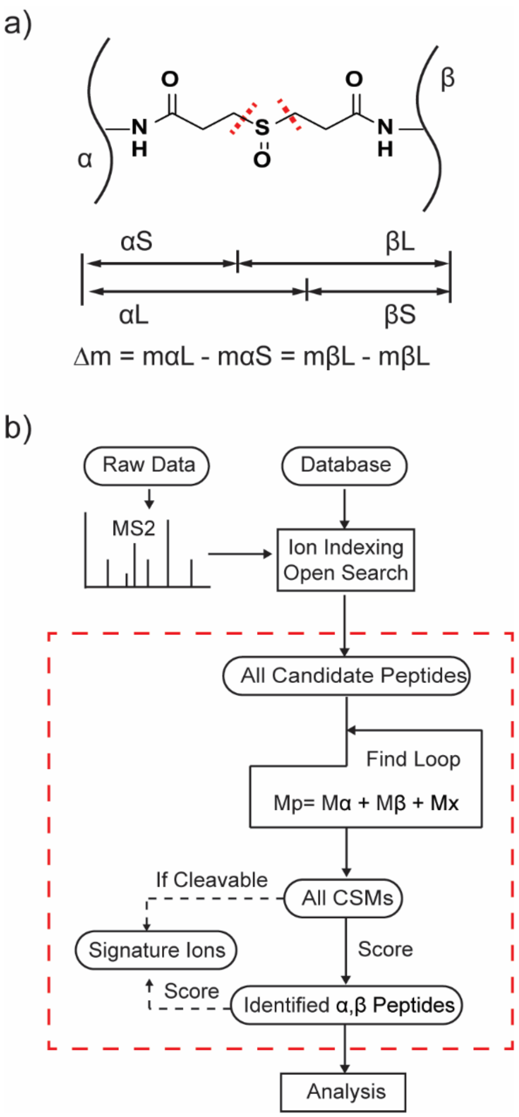


Figure 1. MetaMorpheusXL for MS-cleavable crosslink search. (a) Four signature fragment ions are generated from DSSO crosslinked peptides (α and β peptides) using CID/HCD (CID: collisional induced dissociation). If the crosslinker is cleaved at the left of the sulfoxide moiety, it will generate the αS and βL fragments; if crosslinker is cleaved at the right, it will generate the αL and βS fragments. The signature mass difference (Δm) between αL and αS, βL and βS is the same; Δm is 31.97 Da for DSSO. (b) Workflow of MetaMorpheusXL (detailed explanation in Methods

Section). In the 'Find Loop', the candidate PSMs are matched pairwise to attempt to satisfy the equation: $M_{precursor} = M_{alpha} + M_{beta} + M_{crosslinker}$. ('M' - mass. 'x' - crosslinker)

2)

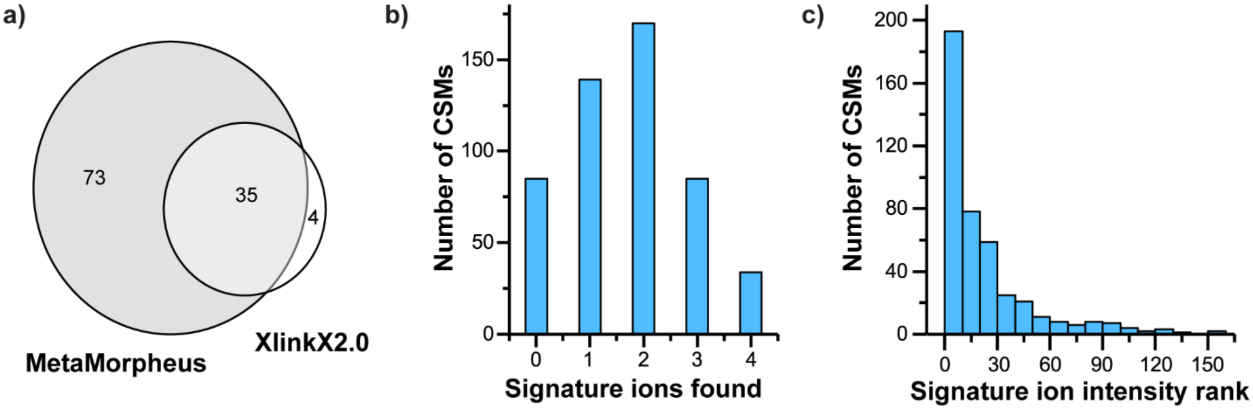


Figure 2. MS-cleavable crosslink search of BSA. (a) Comparison of identified unique crosslinked peptide pairs of BSA by MetaMorpheusXL and XlinkX 2.0. (b) The distribution of number of signature ions observed in the MS2 spectra from identified CSMs. (c) The relationship between number of CSMs and the most intense signature ions' intensity ranks.

3)

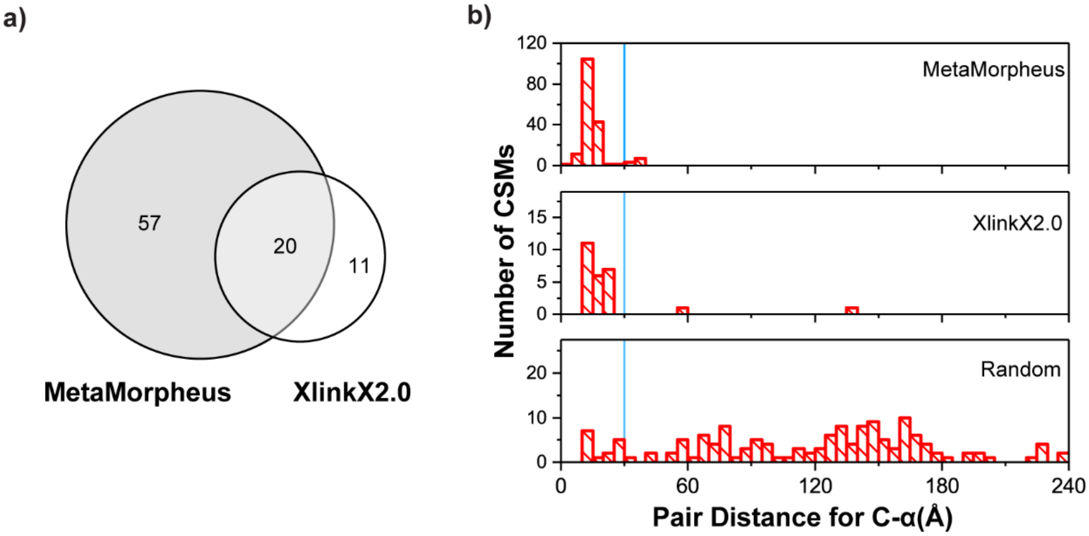


Figure 3. Analysis of DSSO crosslinked Ribosome. (a) Comparison of identified unique crosslinked peptide pairs of *E. coli* ribosome by MetaMorpheusXL and XlinkX 2.0. (b) Cα-Cα distance distribution for experimentally observed lysine-lysine pairs from MetaMorpheusXL, XlinkX 2.0 and a random distribution. The blue lines denote the 30 Å cutoff.

4)

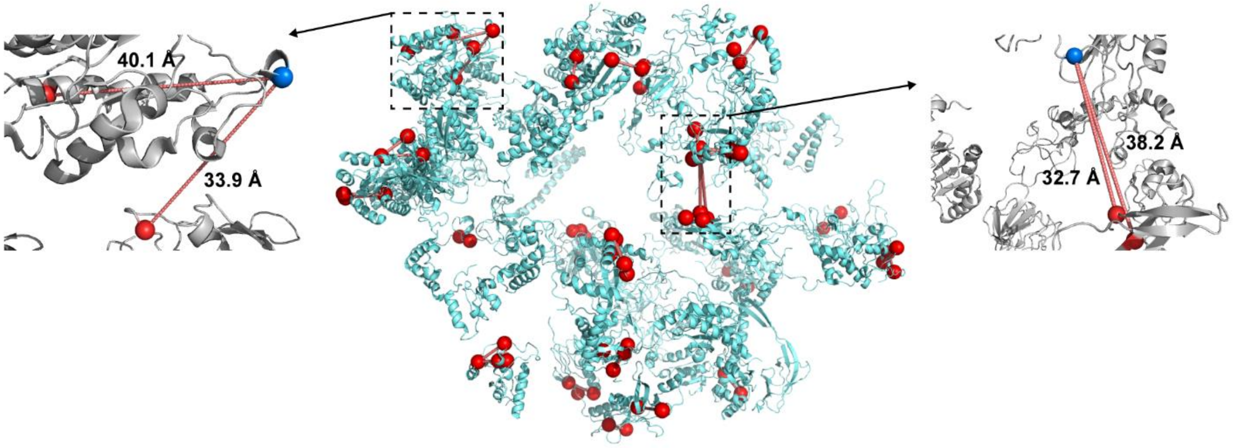


Figure 4. Structure of DSSO crosslinked *E. coli* ribosome (PDB: 3jcd) with crosslinked lysine residues and distances shown. Crosslinks are identified using MetaMorpheusXL at a 1% FDR cutoff. The whole *E. coli* ribosome structure is shown in the middle, red lines indicate Cα-Cα distances between the crosslinked lysine (marked as red spheres). Four crosslink outliers are also shown; at left panel with an inter-crosslink with Cα-Cα distance of 40.1 Å on P0A7V0(36) and P0A7V0(58), an intra-crosslink with Cα-Cα distance of 33.9 Å on P0A7V0(36) and P0A7W7(69).; at right panel with an intra-crosslink with Cα-Cα distance of 38.2 Å on P0A7L0(141) and P0A7N9(58), and another intra-crosslink with Cα-Cα distance of 32.7 Å on P0A7L0(141) and P0A7N9(9).

5)

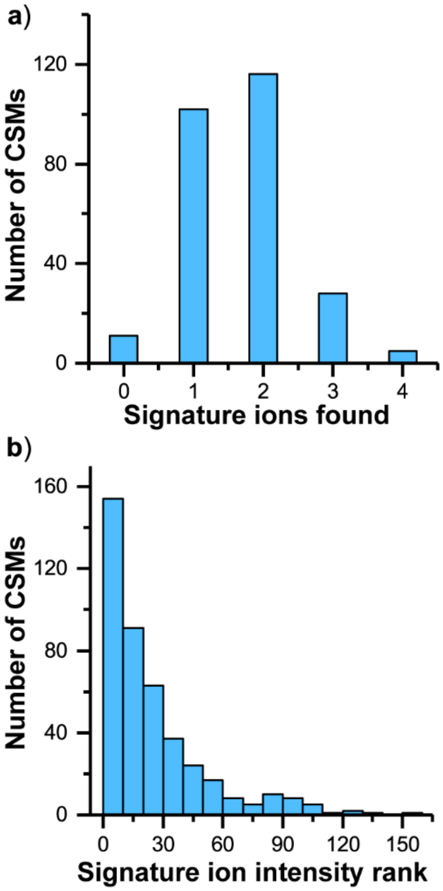


Figure 5. Validation of signature ions from DSSO crosslinked *E. coli* ribosome. (a) The distribution of number of signature ions observed in the MS2 spectra from identified CSMs. (b) The relationship between number of CSMs and the most intense signature ions' intensity ranks.

6)

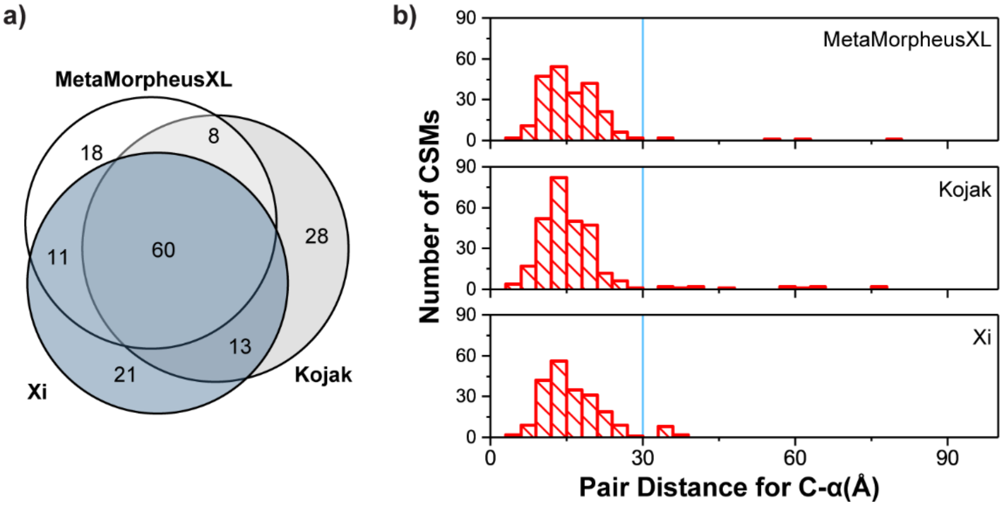


Figure 6. Analysis of non-cleavable crosslinked yeast Pol II complex results. (a) Comparison of identified unique crosslinked peptide pairs of yeast Pol II complex by Kojak, Xi and MetaMorpheusXL. (b) Cα-Cα distance distribution for experimentally observed lysine-lysine pairs from MetaMorpheusXL, Kojak and Xi. The blue lines denote the 30 Å cutoff.

Supplemental Information

Supplementary Methods

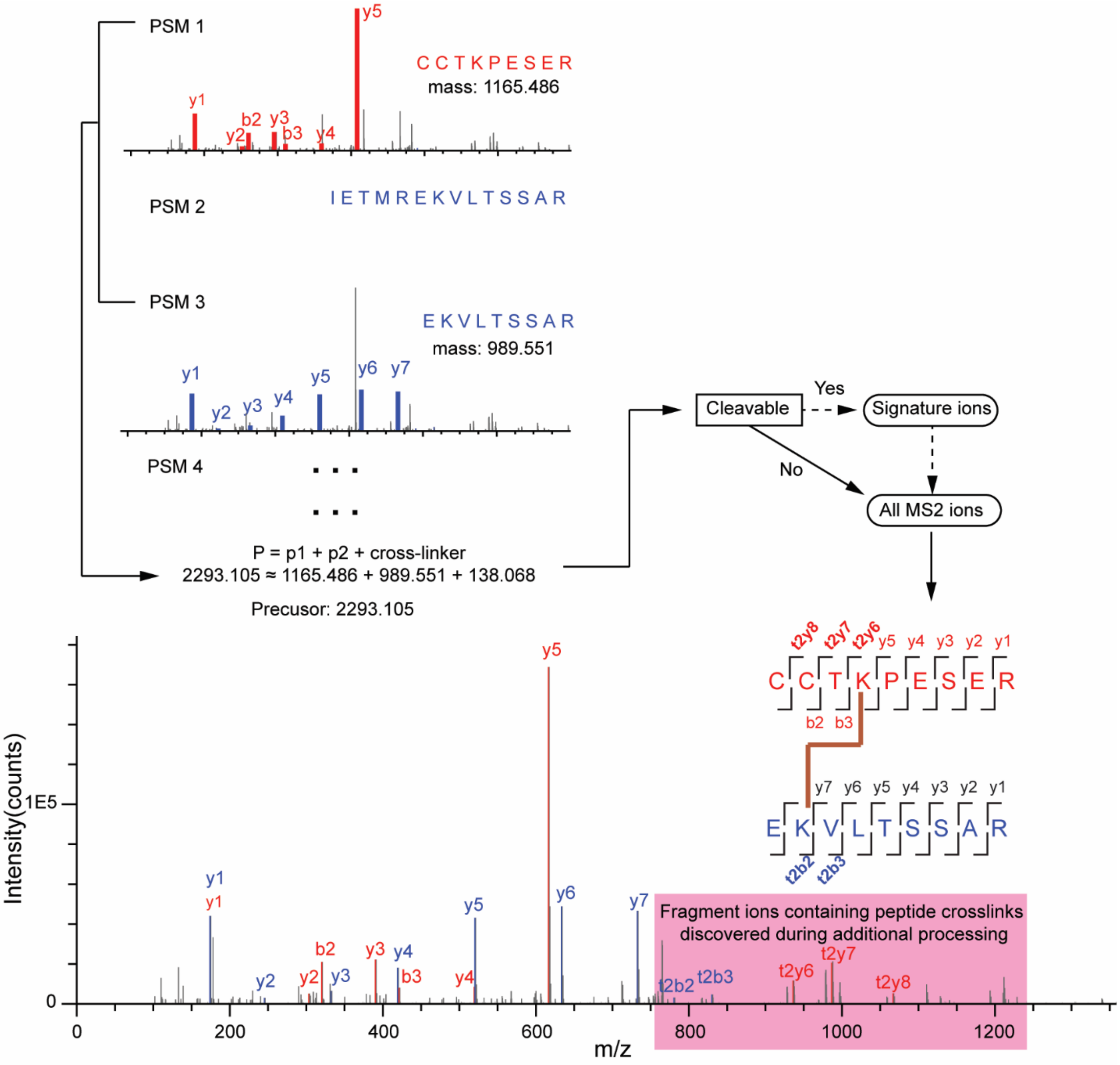
1. Ion-indexing: An "open-mass" search (i.e., when the precursor mass does not limit the space of theoretical peptides for fragment matching) with an ion-indexing strategy is used in MetaMorpheusXL. In this algorithm, the protein database is digested *in silico* and the digestion products (theoretical peptides) are written to a peptide index with each unique peptide being identified by an integer value ("ID"). The peptides are ordered by mass and each is fragmented *in silico*. For each theoretical fragment, its peptide's ID is stored in a lookup table according to the fragment mass (rounded to the nearest mDa). Experimental fragments are matched to theoretical peptides by finding the experimental fragment's mass in the lookup table. The peptide IDs in the fragment mass bin are filtered by the desired precursor mass tolerance (for completely open-mass searches, this tolerance is infinity and all peptides in the bin are counted as having matched to that experimental fragment ion).

2. Calibration: The mass accuracy of MS1 and MS2 spectra gathered during a proteomics experiment can vary significantly over the course of a single run and over the course of several runs. Systematic drift, random noise, and changes in temperature and other environmental conditions can contribute to this variation. Therefore, spectral mass calibration prior to the final analysis can improve peptide identification accuracy. MetaMorpheus uses a machine-learning algorithm to calibrate both MS1 and MS2 spectra. The process begins with a preliminary search of the uncalibrated file to identify a set of confident peptide spectral matches. Mass spectral peaks of confident PSMs are the calibration points, accompanied by several additional values, including: the difference between observed *m/z* and theoretical *m/z* (the "*m/z* error"), the absolute *m/z*, the retention time, the total ion current, and the ion injection time. All of these values serve as input to a random forest machine-learning algorithm that performs a regression analysis to model the *m/z* error as a function of the above explanatory variables. This function is used to shift the *m/z* of all peaks in all scans in the run. The calibrated spectra file is then used for a complete proteomics analysis.

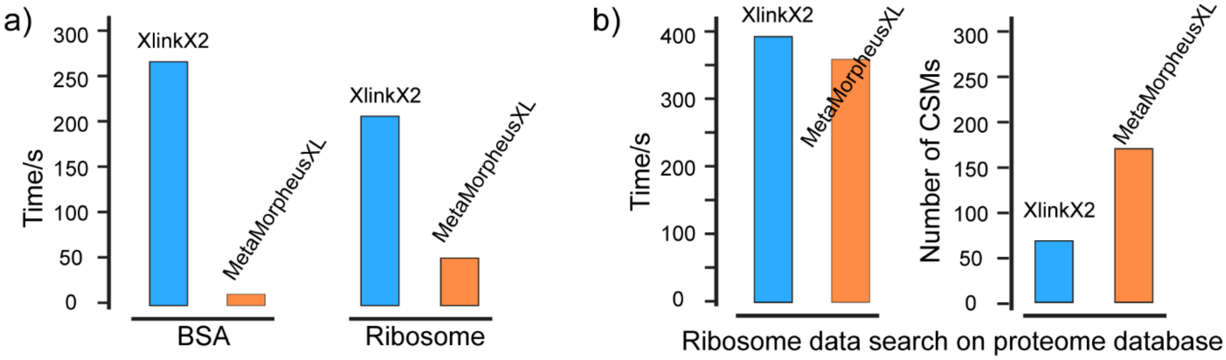
3. FDR estimation: The q-value for each CSM is determined by calculating the ratio of the count of CSMs assigned to target by the count of CSMs assigned to decoy with scores greater than or equal to the current CSM (q-value = (target count)/(decoy count)). In MetaMorpheusXL, a CSM is assigned as a target only when both peptides of the crosslink pair are present in the target database. When either member or both of a crosslink pair are present in the decoy database, the CSM is assigned as a decoy. This results in an imbalance in the total number of target and decoy pairs, which makes it more likely for a CSM to be assigned as a decoy than a target compared to a typical target-decoy search. Therefore, it is possible for the q-value to exceed 1.

4. Example explanation of MetaMorpheusXL's workflow: SI Figure S-1 An illustration of the algorithm used by MetaMorpheusXL to identify a peptide crosslink. This example uses a high-quality MS2 spectrum from a BSA sample crosslinked with DSS. The MS2 spectrum was obtained from a precursor species with a mass of 2293.105 Da. MetaMorpheusXL first finds all candidate peptides by matching primary fragment ions using an indexed-ion open search method (see the *Ion-Indexing* section of this supplement). All possible peptide matches are then paired with each other to generate candidates for crosslink pairs. A candidate pair is valid if the summed mass of the two peptides and the crosslink molecule matches the precursor mass. For example, PSM 1 and PSM 3 in the top panel of **SI Figure S-1** are considered a valid candidate pair because the summed mass of the two peptides and the crosslinker (1165.486 Da + 989.551 Da + 138.068 Da) are within the tolerance of precursor mass (2293.105 Da ± 10 ppm). Theoretical ions containing crosslinker-specific modifications are then generated for each crosslink candidate pair and matched to the spectrum; the highest-scoring (the CSM with the most matching fragments) is retained.

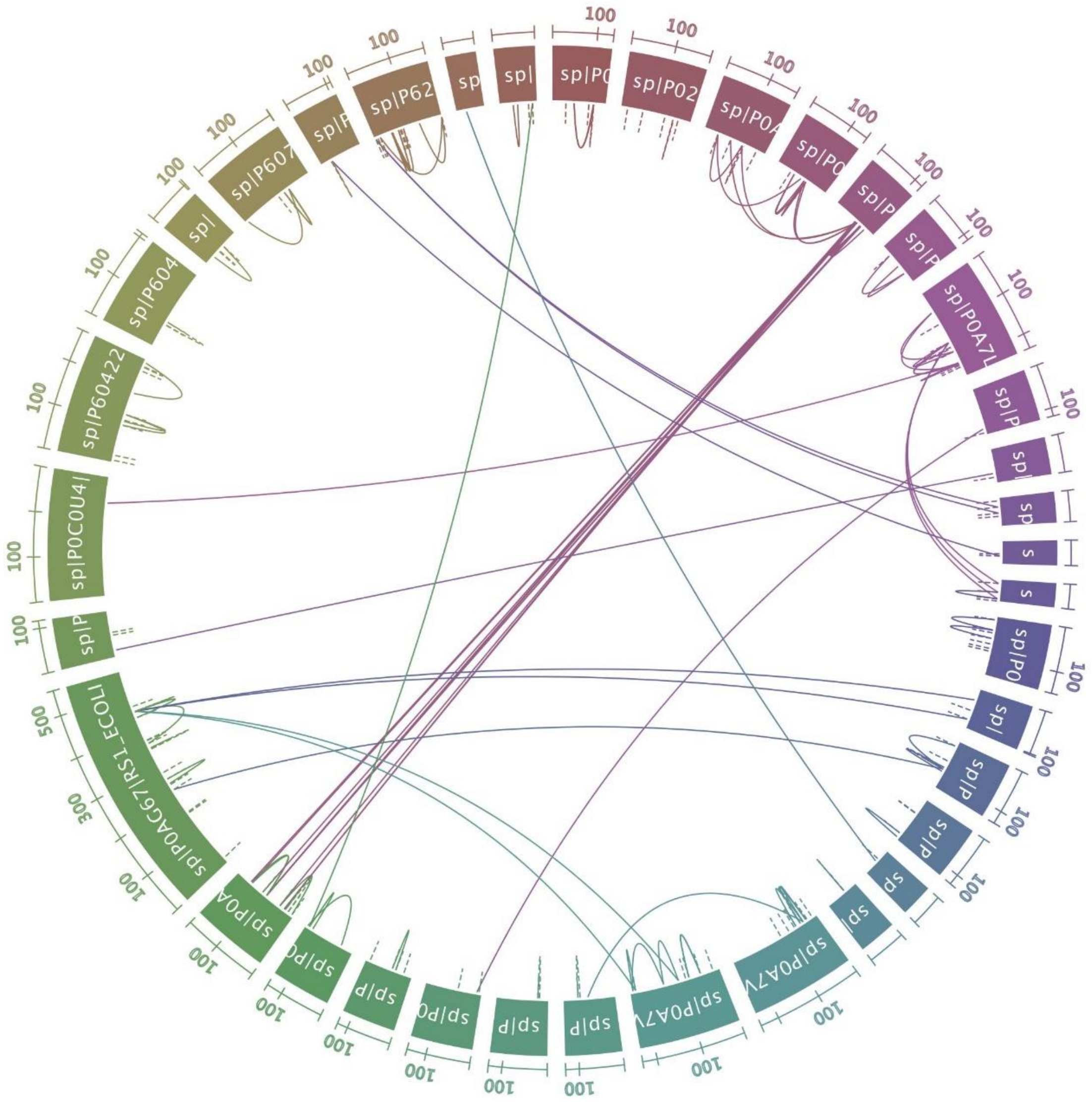
SI Figure S-1. Example of a crosslinked peptide identified by MetaMorpheusXL. The MS2 spectrum’s precursor mass is 2293.105 Da. Preliminary peptide matches are generated with an open-mass search. The candidate PSM masses are paired, with a valid pair satisfying the equation $M_{precursor} = M_{alpha} + M_{beta} + M_{crosslinker}$. The highest-scoring pair that satisfies this constraint was PSM 1 (CCTKPESER) paired with PSM 3 (EKVLTSSAR). Fragment ions containing peptide crosslinks are discovered during an additional processing and the CSM score is increased by one for each additional fragment ion matches.



SI Figure S-2. Computation time comparison between MetaMorpheusXL and XlinkX 2.0. (a) Computation time comparison for searching BSA and ribosome data using small theoretical databases. (b) Computation times and numbers of CSMs identified when searching ribosome data against the entire *E. coli* proteome database, which contains 4443 proteins. MetaMorpheusXL took 6.4 min when restricted to the top 500 peptide matches per MS2 spectrum and identified 173 inter- and intra-CSMs combined, 35% less than the 262 inter- and intra-CSMs identified using the restricted database. Only 3 identified proteins were not ribosomal or ribosome-related. XlinkX 2.0 took 6.5 min and identified 66 CSMs, among which 21 proteins were not ribosomal or ribosome-related. Searches with MeroX and DXMSMS using the whole *E. coli* proteome database took too long to evaluate the results.

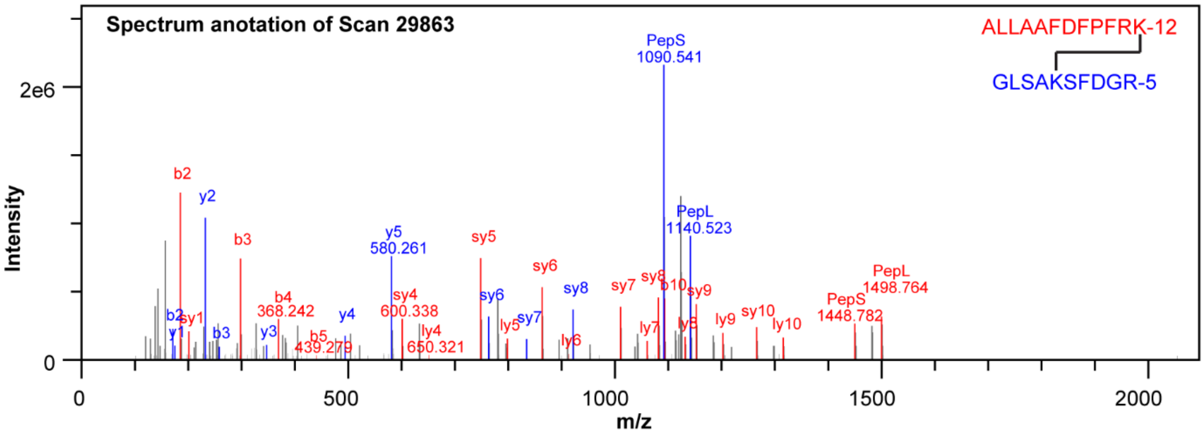


SI Figure S-3. Circle plot from ProXL displaying crosslinks of DSSO-crosslinked ribosome proteins. The bars represent proteins, lines represent crosslinks and dashed lines represent dead-ends.

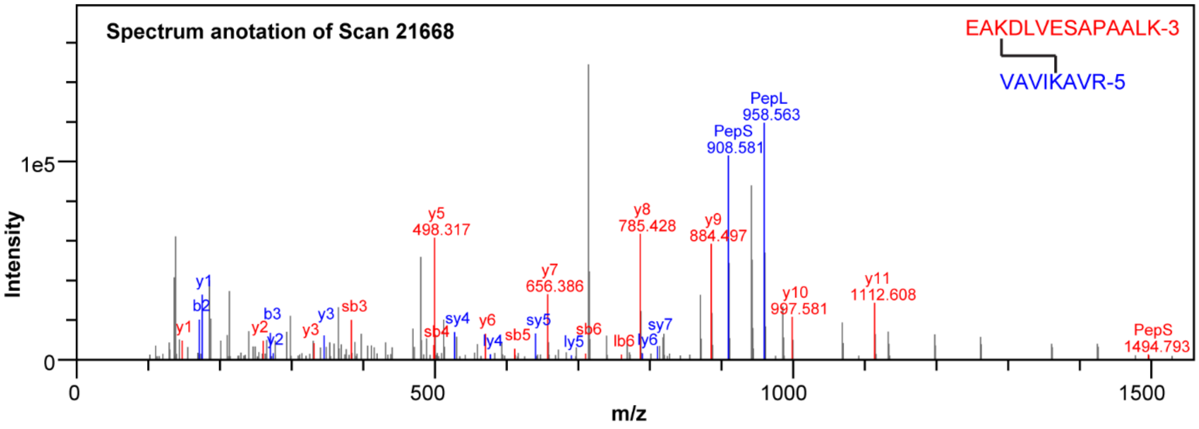


SI Figure S-4. Examples of annotated spectra from *E. coli* ribosome data with different numbers of identified signature ions (4 to 0) and from different score ranges (high to low). The “PepS” or “PepL” annotations indicate signature ions containing the short or long pieces of the fragmented MS-cleavable crosslinker molecule. “PepS2” indicates a doubly-charged signature ion with a short crosslinker piece. The “lb4” and “sb4” refer to b4 ions with long or short cleavage products, respectively. (a) This CSM contains 4 signature ions with a high score. (b) This CSM contains 3 signature ions. (c) This CSM contains 2 signature ions, both of which are from the alpha peptide. (d) This CSM contains 1 signature ion from its alpha peptide. (e) This CSM contains no signature ions. (f) This CSM contains no signature ions and is low-scoring.

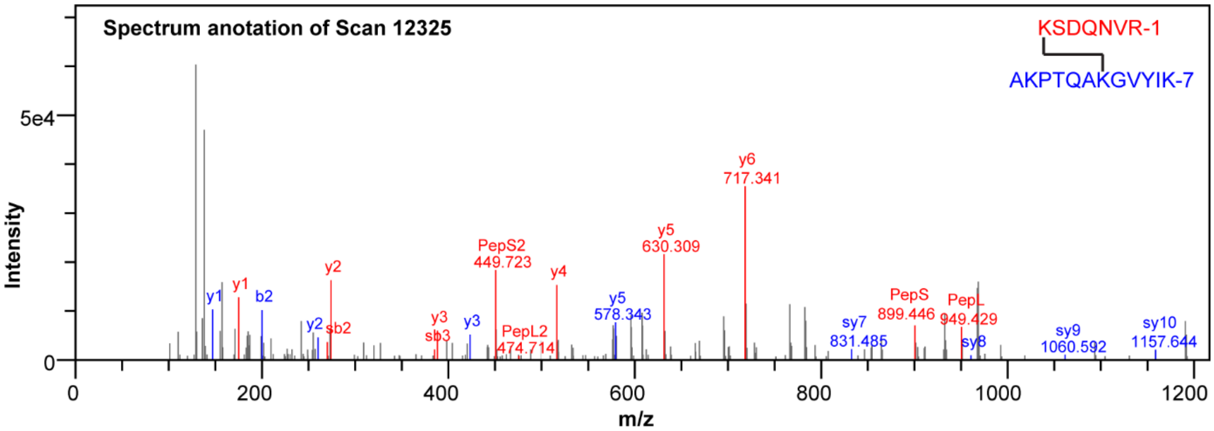
a)

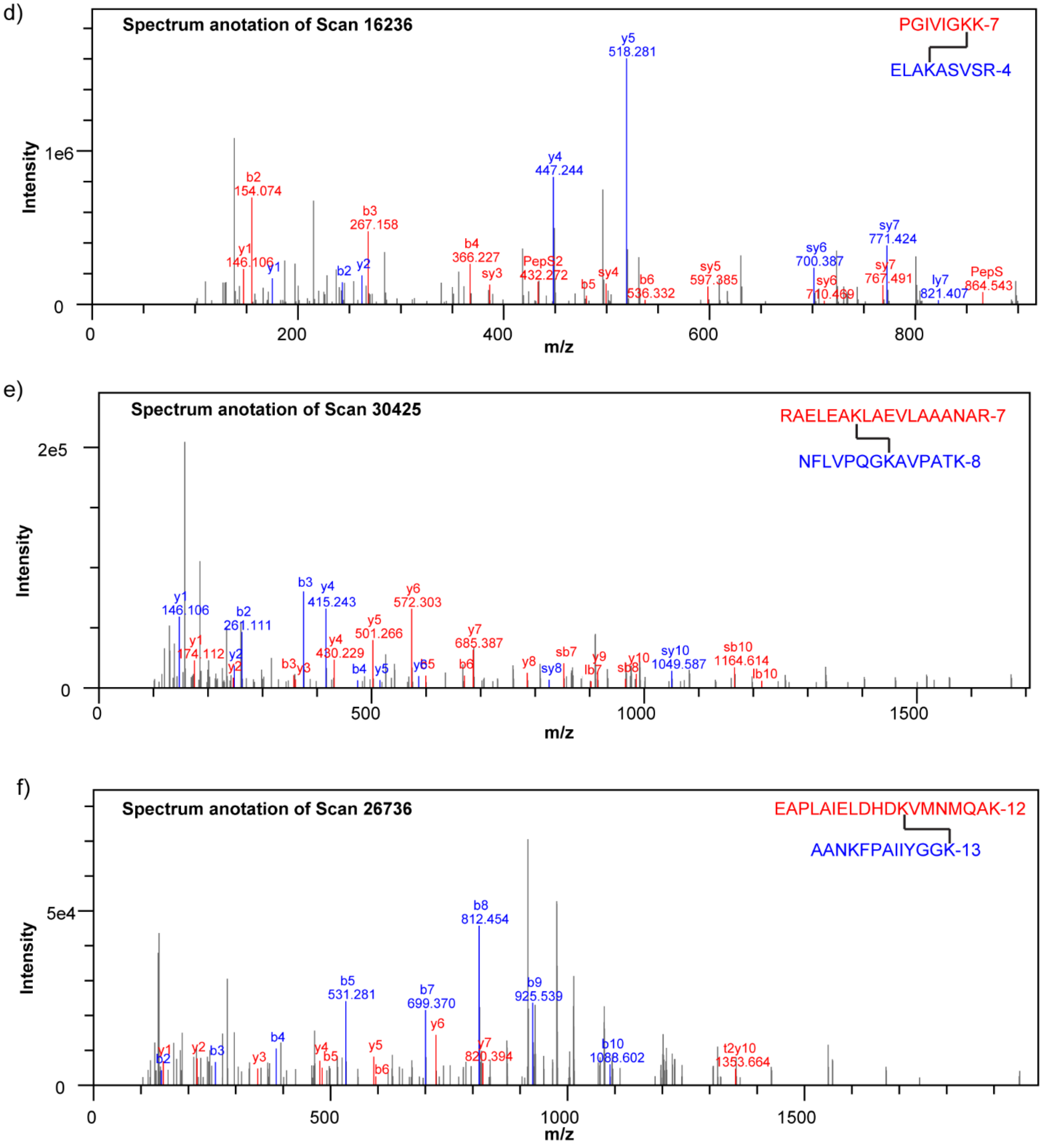

Spectrum anotation of Scan 29863
ALLAAFDFPFRK-12
GLSAKSFDGR-5
PepS
1090.541
PepL
1140.523
b2
y2
b3
y5
580.261
sy5
sy6
sy8
sy7
sy8
b10
sy9
PepL
1498.764
PepS
1448.782
b4
368.242
sy4
600.338
y4
b2
sy1
y1
b3
y3
b5
439.279
ly4
650.321
sy6
ly5
sy7
ly6
ly7
y8
ly9
sy10
ly10
Intensity
2e6
0
0
500
1000
1500
2000
m/z

b)

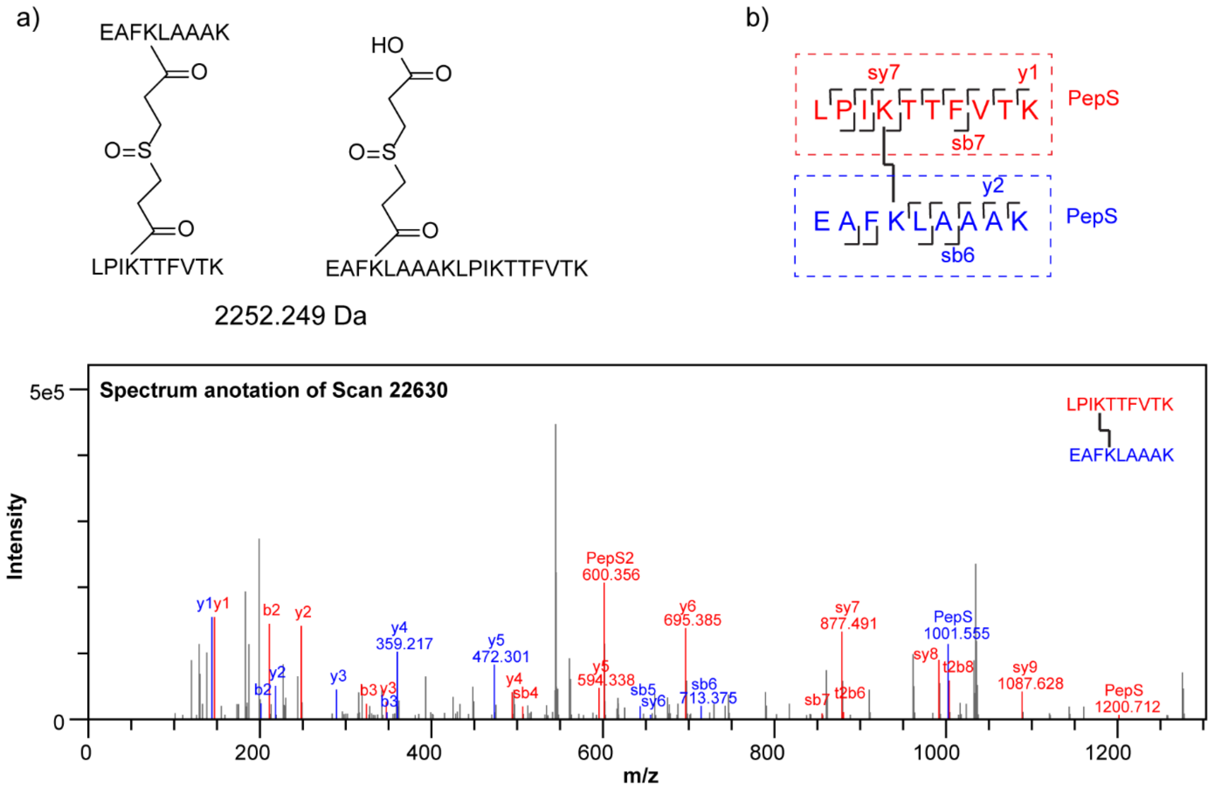

Spectrum anotation of Scan 21668
EAKDLVESAPAALK-3
VAVIKAVR-5
PepL
958.563
PepS
908.581
y5
498.317
y8
785.428
y9
884.497
y7
656.386
y10
997.581
y11
1112.608
y1
b2
b3
y3
sb3
y1
y2
y2
y3
sb4
sy4
y6
ly4
sb5
sy5
sb6
ly5
lb6
ly6
sy7
PepS
1494.793
Intensity
1e5
0
0
500
1000
1500
m/z

c)


Spectrum anotation of Scan 12325
KSDQNVR-1
AKPTQAKGVYIK-7
y6
717.341
y5
630.309
PepS2
449.723
y4
y2
y1
y1
b2
y2
sb2
y3
sb3
y3
PepL2
474.714
y5
578.343
PepS
899.446
PepL
949.429
sy7
831.485
sy8
sy9
1060.592
sy10
1157.644
Intensity
5e4
0
0
200
400
600
800
1000
1200
m/z


d)
Spectrum anotation of Scan 16236
PGIVIGKK-7
ELAKASVSR-4
Intensity
1e6
0
m/z
0
200
400
600
800
y5
518.281
y4
447.244
b2
154.074
b3
267.158
y1
146.106
y1
b2
y2
b4
366.227
sy3
PepS2
432.272
b5
sy4
b6
536.332
sy5
597.385
sy6
700.387
sy6
710.469
sy7
771.424
sy7
767.491
ly7
821.407
PepS
864.543
e)
Spectrum anotation of Scan 30425
RAELEAKLAEVLAAANAR-7
NFLVPQGKAVPATK-8
Intensity
2e5
0
m/z
0
500
1000
1500
y1
146.106
y1
174.112
b2
261.111
y2
y2
b3
415.243
y4
b3
y3
y4
430.229
b4
y5
501.266
y5
y6
572.303
y6
b5
b6
y7
685.387
y8
sy8
sb7
lb7
y9
sb8
y10
sy10
1049.587
sb10
1164.614
lb10
f)
Spectrum anotation of Scan 26736
EAPLAIELDHDKVMNMQAK-12
AANKFPAIIYGGK-13
Intensity
5e4
0
m/z
0
500
1000
1500
y1
b2
y2
b3
y3
b4
y4
b5
b5
531.281
y5
b6
b7
699.370
y6
y7
820.394
b8
812.454
b9
925.539
b10
1088.602
t2y10
1353.664

SI Figure S-5. Identification of intra-protein crosslinks composed of consecutive sequences. (a) Intra-crosslinks composed of consecutive sequences (left) have the same precursor mass as the dead-end missed-cleavage product modified with hydrolyzed crosslinker (right). (b) MetaMorpheusXL assigns a crosslink composed of consecutive sequences as an intra-crosslink only if the matched fragment ions could differentiate it from the dead-end crosslink. From the ribosome data, an intra-crosslink composed of consecutive sequences ‘EAFKLAAAK’ and ‘LPIKTTFVTK’ of protein P0ADY7 are shown as an example here. The spectral matches containing indicative fragment ions (*e.g.*, the y2 ion of ‘EAFKLAAAK’) support that the pair is an intra-crosslink instead of dead-end missed-cleavage product.



SI Table-1. Parameters used in this work for searches of crosslinked data with MetaMorpheusXL, XLinkX 2.0 and Kojak 1.5.

MetaMorpheusXL parameters

Crosslinker type: The crosslinker molecule used in the sample; can be user defined.

Search top matches: if selected, MetaMorpheusXL will only consider N top-scoring peptides for peptide pairing.

Search top Num: used together with 'Search top matches'; this defines the N top peptide candidates.

Trim MS/MS peaks: only match the most intense peaks in an MS2 spectrum. Used together with 'Top N peaks' (i.e., N most intense peaks) and 'Minimum ratio' (peaks must be this intense compared to the base peak).

Minimum Score allowed: the lowest peptide score after the 'first pass' that will be considered.

parameters	Cleavable	Non-cleavable
Precursor Mass tolerance	10 ppm	10 ppm
Crosslinker type	DSSO	DSS
Search top matches	-	✓
Search top Num	-	300
Use Provided Precursor	✓	✓
Deconvolute Precursor	✓	✓
Trim MS1 Peaks	-	-
Trim MS/MS Peaks	✓	✓
Top N Peaks	200	500
Minimum ratio	0.01	0.005
Generate decoy proteins	✓	✓
Max missed cleavages	2	2
protease	trypsin	trypsin
Initiator methionine	Variable	Variable
Max modification isoforms	4096	4096
Min peptide length	5	5
Product mass tolerance	20 ppm	20 ppm
Ions to search	B ions, Y ions	B ions, Y ions
Minimum Score allowed	5	2
Fixed modification	Carbamidomethyl of C	Carbamidomethyl of C

Variable modification	Oxidation of M	Oxidation of M
Localize all modification	✓	✓
Output for Percolator	-	✓
Output for Crosslink	✓	✓

Parameters used in XlinkX2.0

XlinkX 2.0 was used as a node in Thermo Scientific Proteome Discoverer 2.2. Parameters used in XlinkX 2.0 are listed below.

XlinkX 2.0 Detection
Acquisition strategy: MS2
Crosslink Modification: DSSO / + 158.004 Da (K)
Minimum S/N: 1.5
Enable protein N-terminus linkage: false
Xlinkx Filter
Select: Crosslinks
Xlinkx Search
Retain FASTA file indexes: True
Enzyme Name: Trypsin(full)
Maximum Missed Cleavages: 2
Maximum Peptides Considered: 10
Minimum Peptide Length: 5
Maximum Number Modifications: 3
Minimum Peptide Mass: 300
Maximum Peptide Mass: 7000
Precursor Mass Tolerance :10 ppm
FTMS Fragment Mass Tolerance: 20 ppm
Static Modification: Carbamidomethyl / + 57.021 Da (C)
Dynamic Modification: Oxidation /+ 15.995 Da (M)
FDR threshold: 0.01
FDR strategy: Percolator

3.3 Parameter used in Kojak1.5

percolator_version = 3.0
enrichment = 0
instrument = 0
MS1_centroid = 1
MS2_centroid = 1
MS1_resolution = 100000
MS2_resolution = 7500
cross_link = nK   nK 138.0680742 DSS
mono_link = nK 156.0786

```
fixed_modification = C 57.02146
fixed_modification_protC = 0
fixed_modification_protN = 0
modification = M 15.9949
modification_protC = 0
modification_protN = 0
diff_mods_on_xl = 0
max_mods_per_peptide = 2
mono_links_on_xl = 0
enzyme = [KR]|{P}
fragment_bin_offset = 0.0
fragment_bin_size = 0.03
ion_series_A = 0
ion_series_B = 1
ion_series_C = 0
ion_series_X = 0
ion_series_Y = 1
ion_series_Z = 0
decoy_filter = DECOY
isotope_error = 1
max_miscleavages = 2
max_peptide_mass = 8000.0
min_peptide_mass = 500.0
max_spectrum_peaks = 0
ppm_tolerance_pre = 10.0
prefer_precursor_pred = 2
spectrum_processing = 0
top_count = 300
truncate_prot_names = 0
turbo_button = 1
```

MetaMorpheusXL User Manual

1. Download the current version of MetaMorpheus from https://github.com/smith-chem-wisc/MetaMorpheus/releases. MetaMorpheusInstaller.msi is suggested for Windows users.
2. Double-click the .msi file to install MetaMorpheus. Open MetaMorpheus after installation.
3. Click the ‘New XL Task’. This will open a window to set the parameters for a new crosslink search. Parameters are described in this document, below. After choosing your parameters, click ‘Add the XLSearch Task’.

Crosslink Search panel:

- ‘Crosslink Precursor mass tolerance’: Sets precursor mass tolerance, in Daltons (Da) or parts per million (ppm).
- ‘Crosslinker Type’: choose the crosslinker used in your sample. If ‘UserDefined’ is chosen, additional crosslinker information must be specified.
- ‘Search Top matches’: this option can help speed up searches. This option defines the number of peptides from the open search to pair.

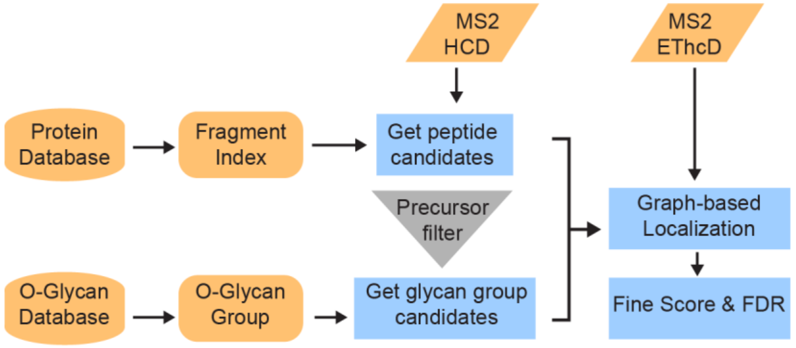
Search Parameters panel:

- ‘Ions to search’: Ion types should be specified to match the fragmentation method (e.g., b and y ions for HCD data).
- When searching a whole proteome database, selecting ‘Search Top matches’ is recommended, along with setting the number of database partitions to ~4-8.

4. Drag your database and spectra files into MetaMorpheus and click ‘Run All tasks’. The results will be in the same folder as the data files.
5. If you have any problems, support is available by reading the wiki (Help -> Open Wiki page), opening an issue on GitHub (Help -> Submit an issue on GitHub), or by emailing the MetaMorpheus development team at mm_support@chem.wisc.edu .

Chapter 4

O-Pair Search with MetaMorpheus for O-glycopeptide Characterization



Adapted from Lu. L.*, Riley, N.M.*, Shortreed, M.R., Bertozzi, C.R. & Smith, L.M. (2020). O-Pair Search with MetaMorpheus for O-glycopeptide Characterization. Nature Methods (Accepted In Principle)

Abstract

We report O-Pair Search, a new approach to identify O-glycopeptides and localize O-glycosites. Using paired collision- and electron-based dissociation spectra, O-Pair Search identifies O-glycopeptides using an ion-indexed open modification search and localizes O-glycosites using graph theory and probability-based localization. O-Pair Search reduces search times more than 2,000-fold compared to current O-glycopeptide processing software, while defining O-glycosite localization confidence levels and generating more O-glycopeptide identifications. O-Pair Search is freely available: https://github.com/smith-chem-wisc/MetaMorpheus.

Main Text

Mass spectrometry (MS) is the gold standard for interrogating the glycoproteome, enabling the localization of glycans to specific glycosites.[1–3] Recent applications of electron-driven dissociation methods have shown promise in localizing modified O-glycosites even in multiply glycosylated peptides[4]. Yet, standard approaches for interpreting tandem MS spectra are ill-suited for the heterogeneity of O-glycopeptides. Perhaps the most challenging problem for O-glycopeptide analysis is mucin-type O-glycosylation, which is abundant on many extracellular and secreted proteins and is a crucial mediator of immune function, microbiome interaction, and biophysical forces imposed on cells, among others[5]. Mucin-type O-glycans are linked to serine and threonine residues through an initiating N-acetylgalactosamine (GalNAc) sugar, which can be further elaborated into four major core structures (cores 1-4) or remain truncated as terminal GalNAc (Tn) and sialyl-Tn antigens[6]. These O-glycosites occur most frequently in long serine/threonine rich sequences (**Supplementary Fig. 1**), such as PTS mucin tandem repeat domains, which exist with macroheterogeneity defined by the occurrence of O-glycosylation and with microheterogeneity defined by a large number of potential O-glycans[7–11]. The number of serine and threonine residues present in glycopeptides derived from mucin-type O-glycoproteins, combined with the consideration of dozens of potential O-glycans at each site, leads to a combinatorial explosion when generating databases of theoretical O-glycopeptides to consider for each tandem MS/MS spectrum (**Supplementary Note 1**).

Current O-glycoproteomic analysis pipelines are unable to search for multiply O-glycosylated peptides within reasonable time frames even for simple mixtures of O-glycoproteins, much less for proteome-scale experiments. Recent efforts to combat this search time issue have forgone site localization for the more expedient option of identifying only the total glycan mass on a peptide backbone[12]. While effective at lowering time costs, this sacrifices valuable information about site-specific modifications – which is often the goal of intact glycopeptide analysis in the first place. Such an approach also fails to report the number and composition of individual glycans for multiply glycosylated O-glycopeptides, where multiple smaller oligosaccharides may represent the same mass as a larger single glycan or a combination of different oligosaccharides. Open modification searches and combinations of peptide database searching with de novo glycan sequencing have also recently been reported, but neither address the time issues that challenge analysis of highly modified O-glycopeptides.[13,14] Moreover, electron-driven dissociation methods are required to localize O-glycosites[8–11,15], yet current software tools fail to capitalize on combinations of collision-based and electron-based fragmentation spectra that are acquired for the same precursor ion. This is coupled with a general lack of ability to confidently localize glycosites within multiply glycosylated O-glycopeptides.

Here, we describe the O-Pair Search strategy implemented in the MetaMorpheus platform[16] to provide a pipeline for rapid identification of O-glycopeptides and subsequent localization of O-glycosites using paired collision- and electron-based dissociation spectra collected for the same precursor ion (**Fig. 1a**). O-Pair Search first uses an ion-indexed open search[17] of higher energy collisional dissociation (HCD) spectra to rapidly identify combinations of peptide sequences and total O-glycan masses, which are generated through combinations of entries in an O-glycan database. Graph-theoretical localization[18–21] then defines site-specific O-glycan localizations using ions present in EThcD spectra (electron transfer dissociation with HCD supplemental activation) (**Fig. 1b**). Peptide backbone fragments (b/y-type ions) rarely retain glycan mass during HCD fragmentation, making them good candidates for an ion-indexed search, while retention of intact glycans on c/z dot-fragments in EThcD spectra enable confident localization,15 as exemplified in the paired HCD-EThcD spectra for the quadruply glycosylated peptide in **Fig. 1c**. Localization is followed by localization probability calculations using an extension of the phosphoRS[22] algorithm used for phosphosite localization (max score of 1), in addition to scoring of fine scoring (which includes calculation of Y-type ions) and false discovery rate calculations performed separately for O-glycosylated and non-modified peptides.

We also introduce here the concept of Localization Levels, which is the culmination of the O-Pair Search (**Fig. 1d**). Inspired by early adoption of class levels for phosphopeptide localizations[23] and more recently for proteoforms[24], we developed this classification system to more accurately describe the quality and confidence of glycopeptide and glycosite identifications. Level 1 glycopeptide identifications indicate that all glycans identified in the total glycan mass modification are localized to specific serine and threonine residues with a localization probability > 0.75. Glycopeptides with glycosite assignments with localization probabilities < 0.75 are assigned as Level 1b, even though they are still identified as localized by the graph theory approach. Level 1b assignments also occur when a glycosite is assigned without the presence of sufficient spectral evidence (e.g., fragments cannot explain a glycosite, but the sequence contains only one serine or threonine). We currently borrow the 0.75 cutoff from phosphopeptide precedents[23], empirical determination of localization cutoffs will likely need to be determined in future work using libraries of synthetic glycopeptide standards, as has been done with phosphopeptides[25]. That said, such libraries are currently difficult to generate. Level 2 assignments occur when at least one glycosite is assigned a glycan based on spectral evidence, but not all glycans can be assigned unambiguously. Level 3 identifications represent a confident match of glycopeptide and total glycan mass, but no glycosites can be assigned unambiguously. Indeed, Level 3 glycopeptides (such as those reported in HCD-only methods by default[12]) are still useful to note the presence of glycosylated residues somewhere in a given sequence. Overall, our classification system provides a straightforward approach to qualify glycoproteomic datasets without having to exclude confident identifications that have no site-specific information.

In addition to Localization Level assignments, O-Pair Search also reports matched peptide and glycan fragment ion series and their intensities for each of the paired spectra, the presence of N-glycosylation sequons to identify potentially confounding assignments, and localization probabilities for all sites, both localized and not. As with all glycopeptide-centric workflows, O-Pair Search reports compositions with limited ability to comment on topology of glycans. That said, oxonium ions can help distinguish between certain glycans, such as the ratio of 138.055 m/z and 144.0655 m/z to differentiate HexNAc residues as GlcNAc or GalNAc.[26] O-Pair Search reports this ratio by default to aid with interpretation. For example, 95.7% of identifications with localized H1N1 glycans have a 138/144 ratio less than 2, suggesting GalNAc-Gal rather than Man-GlcNAc for H1N1 identity. Furthermore, during the review process for this manuscript, another approach using fragment ion-indexed open searching of glycopeptide spectra (but without localization as discussed here) was reported from Nesvizhskii and co-workers[27], the group who developed the highly efficient ion-indexed strategy. Comparisons between MSFragger-Glyco and O-Pair Search are discussed in **Supplementary Note 2**.

We first compared O-Pair Search to Byonic, the most commonly used O-glycopeptide identification software[28]. Byonic, which uses a look-up peaks approach to speed up search times relative to traditional database searching[29], can also search HCD and EThcD spectra, although it is agnostic of paired spectra originating from the same precursor. To benchmark performance, we used a recently published dataset[15] of O-glycopeptides from mucin glycoproteins using a combination of trypsin and the mucin-specific protease StcE, which cleaves only in glycosylated mucin domains[30]. This data originates from sequential digestion of four recombinant mucin standards (CD43, MUC16, PSGL-1, and Gp1ba), using StcE to cleave mucin domains followed by N-glycan removal with PNGaseF and tryptic digestion. We initially searched a file with HCD and EThcD paired spectra from this dataset, using a protein sequence file of the four mucins (to minimize Byonic search times) and a glycan database of 12 common O-glycans presented in **Fig. 1a12**. The number of compositions considered in each search is not simply 12, however, but instead is a combination of possible compositions determined by the number of glycans allowed per glycopeptide. When four glycans are allowed per peptide, this actually represents 439 different mass offset values, i.e., the number of unique masses present in 1819 different glycan combinations (**Supplementary Note 1**). O-Pair Search identified more localized (**Fig. 2a**) and total (**Fig. 2b**) O-glycopeptide spectral matches (GlycoPSMs) than Byonic when allowing either 2 or 3 glycans per peptide (**Supplementary Fig. 2** and **Fig. 2**, respectively). This holds true even when relaxing the scoring thresholds used to obtain confident Byonic identifications (**Supplementary Fig. 3**). Note, all O-Pair Search identifications represent two spectra from an HCD-EThcD spectral pair. Conversely, Byonic is agnostic to paired scans, meaning identifications can come from HCD and EThcD spectra that were collected for the same precursor (pair) or from spectra identified separately from their paired counterpart.

Importantly, O-Pair Search dramatically decreased search times, with ~45-fold and ~2,100-fold faster searches than Byonic when considering 2 or 3 glycans per peptide, respectively (**Fig. 2c**). O-Pair Search required approximately 30 seconds to complete a search considering 4 glycans per peptide, while the Byonic search was terminated after the search failed to complete in over 33,000 minutes (~3.5 weeks). Improvements in search speed are accompanied by ~2-3-fold increases in the number of localized glycosites identified. In addition to more than doubling the number of total identified spectra, O-Pair Search identified the majority of spectra that Byonic returned as GlycoPSMs for both HCD (**Fig. 2d**) and EThcD (**Fig. 2e**) scans, and the overwhelming majority (~95%) of the shared identified scans mapped to the same glycopeptide (**Fig. 2f**). These searches were completed using a FASTA file containing sequences only for the four mucin standards, which highlights the impracticality of O-glycopeptide searches in Byonic for complex mixtures. Moreover, O-Pair Search performed localization calculations and reported Localization Levels within the reported search time while Byonic spectra had to be further processed after the search to obtain localization information (see Methods). **Supplementary Fig. 4** compares microheterogeneity seen at localized glycosites between O-Pair Search and Byonic searches.

The ability to rapidly search O-glycopeptide data allowed us to vary the number of O-glycans to consider per peptide for easy evaluation of optimal search conditions. **Fig. 2g** shows that search times remain less than a minute when considering 5 glycans per peptide, while up to 8 glycans can be considered per peptide in searches requiring less than 20 minutes. Allowing for more glycans per peptide does not change the spectral assignments to various glycopeptides (**Supplementary Fig. 5**), indicating the robustness of O-Pair Search identifications. The number of non-modified identifications remained similarly constant (**Supplementary Fig. 6**). Similarly, different glycan databases can be searched within reasonable timeframes (**Supplementary Fig. 7**).

Evaluating retention time rules further supports O-Pair Search identifications, where glycopeptide identifications containing 0, 1, and 2 sialic acids on the same peptide backbone have predictable elution time shifts (**Supplementary Fig. 8**)[31,32]. O-Pair Search localization of different glycosites also enabled visualization of chromatographically resolved glycopeptide positional isomers (**Supplementary Fig. 9**). That said, glycan isomers (i.e., same composition, different connectivity) remain a challenge that currently requires manual interpretation.[9,10] Interestingly, processing of this published dataset to evaluate the best fragmentation conditions for O-glycopeptides generated the same overall conclusions as the previously reported Byonic searches, although the differences between different supplemental activation energies for EThcD appear more subtle than before (**Supplementary Fig. 10**). Sequences with up to five localized glycosites were identified, but the majority of Level 1 identifications had 1, 2, or 3 localized glycosites (**Supplementary Fig. 11**). Note, these searches were completed using 16 cores, but similar performance can also be achieved on most standard computing systems using fewer cores (**Supplementary Fig. 12**). Overall, this method enabled characterization of dozens of glycosites on each glycoprotein in the mixture (**Supplementary Fig. 13**). It is important to note that O-Pair Search can process HCD and stepped collision energy HCD spectra collected in both product-dependent and standard acquisition modes, too (**Supplementary Fig. 14**). The vast majority of these identifications are Level 3, however, because collision-based fragmentation largely does not support glycosite localization.

We also evaluated O-Pair Search search times and false discovery rates using several entrapment protein databases with varying complexity (**Fig. 2h**). A description of the databases used for benchmarking is provided in **Supplemental Note 2**; briefly, databases were designed to represent different proteome backgrounds not present in the sample (true negatives), with the four mucin standard target sequences (true positives) appended. Entrapment backgrounds ranged from 20 canonical human mucins to the entire E. coli and yeast proteomes. Search times for the mucin, FBS, cell surface glycoprotein, and E. coli entrapment databases (all with < 1,000 entries) remained under ~20 minutes when using 16 cores, while the yeast entrapment proteome took ~3.4 hours. Still, this is approximately one fourth the time Byonic required for a far less complex search (**Fig. 2c**). Sensitivity, as measured by the number of O-glycopeptide identifications, varied with the entrapment backgrounds, which was also evident for non-modified peptide identifications (**Supplementary Fig. 6**). This highlights the known issue of proper database size selection in glycoproteomics[33], which can be more thoroughly explored for O-glycoproteomics now that O-Pair Search enables reasonable search times. Importantly, Level 1 peptides were the least affected, supporting their high confidence assignments. O-Pair Search maintained acceptable false discovery rates (0-3%) even when challenged with these entrapment databases (**Fig. 2i**), performing well compared to Byonic (**Supplemental Fig 15**) and to previous reports34,35. Explanations for the few false hits that were reported include 1) candidates with similar sequence components such as the yeast peptide TNNFFLPSEDESGPVQSSVK (an observed false hit) and the CD43 peptide GASGPQVSSVK, 2) errors in monoisotopic mass assignment due to incorrectly assigned nitrogen deamidation (a known problem in glycoproteomics), or 3) inflated precursor candidate list sizes with larger databases that exclude true peptide candidates from consideration. Future work will continue to investigate methods to improve the sensitivity and precision of the search, such as enhanced indexing scoring, appropriate candidate list sizes, an additional fine scoring step to look for additional fragment types (e.g., [c-1]●, [z+1], w-type ions used by Byonic and Protein Prospector), or use of re-scoring algorithms, e.g., Percolator[36]. We also searched using proteomes from expression systems of these recombinant mucins (CHO and NS0 cells) (**Supplemental Fig. 16**), which are prohibitively large databases for analogous Byonic searches.

Finally, we applied O-Pair Search to a large dataset of urinary O-glycopeptides, which has been analyzed in a number of studies[8–10,37]. The raw data for this dataset represents glycopeptides purified from urine from four donors using affinity chromatography with wheat germ agglutinin and is available through the MassIVE repository (MSV000083070).8 Pap et al. provide identifications from Protein Prospector and Byonic for EThcD scans from Fraction 1 (the "shoulder fraction", three raw data files available) and Fraction 2 (the "GlcNAc fraction", two raw data files available). In that dataset, searching was somewhat restricted, presumably due to complexity issues discussed above: only EThcD scans filtered for the presence of sialic acid oxonium ions were searched, with a 4-glycan database and 2 variable modifications considered per peptide. We searched Fraction 1 with O-Pair Search using the entire human proteome database (~20,300 entries) with 2 glycans considered per peptide from the 12 common O-glycan database used above, and we compared the results to the reported identifications for the other two search engines (**Fig. 2j**). Because this dataset had the potential to harbor N-glycopeptides as well, we filtered out all identifications that included an N-sequon from our O-Pair Search results. Even so, O-Pair Search nearly doubled the total number of GlycoPSMs from either search engine. Of the 382 spectra identified by both Protein Prospector and Byonic, O-Pair Search identified ~90% of them (342 spectra) while providing an additional 506 GlycoPSMs not reported by either. Of the total 1,287 spectra identified as GlycoPSMs, O-Pair Search identified ~85% of them (1,098 spectra). The original study reported a predominance of sialylated glycopeptides, which is recapitulated by O-Pair Search with >97.5% of GlycoPSMs (1,071 of 1,098) containing a sialic acid. When comparing identifications from the 342 scans identified in all three search algorithms, all return the same glycopeptide sequence. Protein Prospector reports a Site Localization In Peptide (SLIP) score[38] for modification sites that we used to convert identifications to our Localization Level scheme (**Fig. 2k**). O-Pair Search reports more Level 1 and 1b O-glycopeptide identifications than the total number of Protein Prospector GlycoPSMs, and the proportion of localized and partially localized identifications (Levels 1-2) is more favorable with O-Pair Search. Similar trends hold for Fraction 2 (**Supplementary Fig. 17**).

We expanded our analysis of this dataset to explore the use of a larger glycan database (32 glycans vs 12) and the effect of searching with more glycans allowed per peptide (5 vs 2). **Fig. 2l** compares results from these different search parameters for Fraction 1, Fraction 2, and all 10 files available for download from the urinary O-glycoproteome dataset. In Fraction 1, The larger O-glycan database boosted identifications for Fraction 1, but lowered identifications in Fraction 2 and the entire dataset as a whole. This indicates that Fraction 1 likely harbored glycopeptides with more diverse glycans while the majority of the dataset did not. Also, according to the original study, Fraction 2 contained more multiply modified O-glycopeptides, which may produce less efficient peptide backbone fragments sufficient for reliable identification. Conversely, considering more glycans per peptide provided slight benefits in all cases. By requiring only a few hours to perform a whole proteome-search with a variety of glycopeptide possibilities, O-Pair Search provides a flexible platform to explore O-glycoproteomics data. When considering only Level 1 and 1b GlycoPSMs, our results represent 447 unique O-glycopeptides with localized O-glycosites, and O-Pair Search identified 354 localized O-glycosites in total when allowing 5 glycans per peptide from the 12-glycan database.

In all, we show that O-Pair Search can reduce O-glycopeptide search times by >2000x over the most widely used commercial glycopeptide search tool, Byonic. Additionally, O-Pair Search identifies more O-glycopeptides than Byonic and provides O-glycosite localizations using graph theory and localization probabilities. O-Pair Search also introduces a novel classification scheme to unify data reporting across the glycoproteomic community. These Localization Levels are automatically calculated by O-Pair Search to indicate if all (Level 1), at least one (Level 2), or none (Level 3) of the O-glycosites are confidently localized. We further demonstrate the utility of O-Pair Search by searching a large published dataset of urinary O-glycopeptides, significantly increasing the number of glycopeptides identified and providing site-specific localization for >350 O-glycosites. We also note that O-Pair Search allows user-specified glycan databases to enable unbiased searches of a variety of O-glycosylation classes, including O-fucose, O-mannose, and O-glucose. That said, these O-glycans often lack monosaccharides that generate the most commonly used oxonium ions for product-dependent methods, so more method development may be needed to optimize data acquisition. A report published while this wok was under review also described the classification of mucin-type O-glycans using B- and Y-type; we will seek to incorporate this into our workflow.[39] Current limitations include the reliance on HCD to generate good peptide fragmentation for the open-search step. Others have shown that starting with EThcD data may be a viable option[37], although this also brings several inherent challenges. This may be alleviated in our approach by indexing both HCD and EThcD spectra for open searching. Even so, true peptides may still rank low among all the peptide candidates. Scoring refinement could improve all open-search approaches (including ours), especially in complex datasets where many precursor candidates must be considered. Also, it remains difficult for any glycoproteomics software to identify glycan isomers (i.e., same composition, different connectivity), a challenge not addressed here. Perhaps better automation of this could be achieved as more studies are published with O-glycopeptides modified with defined glycan structures. Regardless, O-Pair Search can process both product-dependent and standard acquisition methods with a variety of O-glycan databases, making it a flexible tool for a variety of O-glycoproteomics applications.

Conclusions

In all, we show that O-Pair Search can reduce O-glycopeptide search times by >2000x over the most widely used commercial glycopeptide search tool, Byonic. Additionally, O-Pair Search identifies more O-glycopeptides than Byonic and provides O-glycosite localizations using graph theory and localization probabilities. O-Pair Search also introduces a novel classification scheme to unify data reporting across the glycoproteomic community. These Localization Levels are automatically calculated by O-Pair Search to indicate if all (Level 1), at least one (Level 2), or none (Level 3) of the O-glycosites are confidently localized. We further demonstrate the utility of O-Pair Search by searching a large published dataset of urinary O-glycopeptides, significantly increasing the number of glycopeptides identified and providing site-specific localization for >350 O-glycosites. We also note that O-Pair Search allows user-specified glycan databases to enable unbiased searches of a variety of O-glycosylation classes, including O-fucose, O-mannose, and O-glucose. That said, these O-glycans often lack monosaccharides that generate the most commonly used oxonium ions for product-dependent methods, so more method development may be needed to optimize data acquisition. A report published while this wok was under review also described the classification of mucin-type O-glycans using B- and Y-type; we will seek to incorporate this into our workflow.[39] Current limitations include the reliance on HCD to generate good peptide fragmentation for the open-search step. Others have shown that starting with EThcD data may be a viable option[37], although this also brings several inherent challenges. This may be alleviated in our approach by indexing both HCD and EThcD spectra for open searching. Even so, true peptides may still rank low among all the peptide candidates. Scoring refinement could improve all open-search approaches (including ours), especially in complex datasets where many precursor candidates must be considered. Also, it remains difficult for any glycoproteomics software to identify glycan isomers (i.e., same composition, different connectivity), a challenge not addressed here. Perhaps better automation of this could be achieved as more studies are published with O-glycopeptides modified with defined glycan structures. Regardless, O-Pair Search can process both product-dependent and standard acquisition methods with a variety of O-glycan databases, making it a flexible tool for a variety of O-glycoproteomics applications.

METHODS

O-Pair Search Algorithm

O-Pair Search has been implemented in MetaMorpheus[16], an open-source search software useful for a variety of different applications including: bottom-up, top-down, PTM discovery, crosslink analysis and label free quantification. O-Pair is optimally designed for identifying O-glycopeptides from tethered collision- and electron-based dissociation spectra collected from the same precursor ion. However, it is also capable of identifying O-glycopeptides from spectra obtained using other fragmentation schemes and modalities. Before the beginning of the open search, MetaMorpheus tracks precursors available from the data and also calculates precursors for potential coisolated peptides.[16] O-Pair Search occurs in three stages: (**Fig. 1a**) 1) identification of peptide candidates using an ion-indexed open search; 2) localization of O-glycosites with a graph-based localization algorithm; and 3) calculation of site-specific localization probabilities. Upon completion of these stages, the O-glycopeptide localization levels (**Fig. 1d**) are determined and reported along with the false discovery rates (FDR), which are presently estimated using the target-decoy strategy.

1. Ion-indexed open search. O-Pair Search uses ion-indexed open search[17] to quickly identify peptide candidates for each spectrum. O-glycosylation is a labile modification and O-glycopeptides under collision-based dissociation in mass spectrometry generate peptide backbone fragment ions rarely retaining the glycans. Thus, even though an O-glycopeptide can be modified with multiple O-glycans, an O-glycopeptide HCD spectrum could be searched to determine the amino acid backbone without considering the O-glycans.

In an ion-indexed open search, a lookup table is created that includes a complete set of theoretical target and decoy fragment masses from the entire protein database, each labeled with the peptide from which it is derived. A collection of all peptides with fragments matching any peak in a given MS2 spectrum is assembled. The peptide candidates are then chosen from those peptides with the most matching fragments.

For each peptide candidate retained from the open search, the mass difference between the unmodified peptide backbone and the experimental precursor mass is computed. The mass difference is hypothesized to be the sum of all glycan masses on the peptide. We refer to the collection of glycans on a given peptides as the glycan group: mass of glycan group = precursor mass - peptide mass. All glycan groups whose mass equals the mass difference within the specified mass tolerance are considered as glycan group candidates for glycosite localization.

2. Graph-based localization. The graph algorithm is specially optimized for O-glycosite localization. A directed acyclic graph is constructed to represent all possible O-glycan modified forms of a peptide candidate and each of its corresponding glycan group candidates. If a peptide candidate corresponds to several different glycan group candidates within the mass tolerance limitation, several graphs are constructed.

The graph is constructed from left to right, beginning with a ‘Start’ node at the N-terminal side of the peptide and ending with an ‘End’ node at the C-terminal side. Nodes, vertically aligned, are added to the graph for each corresponding serine or threonine because these amino acids are the only two allowed for O-glycosite occupancy. One vertical node designates the site as unoccupied and is labeled with ‘N’. Vertical nodes are then added, one for each potential glycan at the current position. These are labelled ‘A’, ‘B’ and so on. Additional vertical nodes are added representing combinations of glycans that may have occurred for the portion of the peptide represented by that vertical column of the graph. Combination nodes are labelled, for example. ‘A+B’. These nodes and labels are repeated at each serine and threonine. Next, adjacent nodes are connected by edges representing the accumulation of glycans across the peptide backbone. Nodes that are not possible given the constraints of the total peptide mass, which stipulate the number and kinds of glycans on the peptide remain disconnected. This process culminates in a graph representing all possible glycopeptides, where each individual continuous path from Start to End represents one unique glycopeptide.

Next, we associate theoretical fragment ions with each node. Here we need to make clear which amino acids and glycans from the peptide are included. Beginning at the N-terminus, the node represents the peptide up to AND INCLUDING the amino acid listed for the node. Beginning at the C-terminus, the node represents the peptide up to BUT NOT INCLUDING the amino acid listed for the node. The two portions of a peptide associated with a node are complementary to each other and do not cross over. Each node has associated with it all possible theoretical peptide fragment masses whose accumulated mass can be uniquely attributed to the glycopeptide segment containing the amino acids up to that point. The MetaMorpheus score for the entire peptide is the count of matching fragments from all nodes in the path plus the fraction of spectrum intensity attributable to the matched fragments. The glycopeptide with the highest MetaMorpheus score can be extracted with dynamic programming and is designated as the match and reported in the results.

We provide the hypothetical example illustrated in **Fig.1b** to aid understanding of the graph theoretical model. The example O-glycopeptide contains 8 O-glycosites. The glycan group consists of two glycans ‘A’ and ‘B’. Either of the two glycans can occupy any one of the eight positions subject to the following requirements: a maximum of two glycans can be on the peptide, only one glycan is allowed per position; and each glycan can appear only once on a given peptide.

For this example, there are 56 total (**Supplementary Table 2**) different modified forms in the graph. The weight of nodes vertically aligned is determined by the number of associated theoretical fragment ions. In the example, the nodes associated with amino acid S9 can be matched to theoretical fragments c9, c10, c11, z9, z10, z11. The path highlighted in orange represents that the peptide is modified on S9 with glycan A and S12 with glycan B.

3. Site-Specific Localization Probability. We use an iterative method to track the localization scores from all the potential paths of the graph to calculate site specific localization probability of a glycoPSM. These scores are integrated with a random event-based localization method similar to a method described previously in PhosphoRS[22]. The integer part of the localization is the MetaMorpheus score, k, which is the total number of matched peaks. This is applied to a cumulative binomial distribution for calculating probability P as follows:

$$\sum_{k}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

In the formula, n is the number of theoretical fragment ions; p is the probability of randomly matching a single theoretical fragment ion given specified tolerances.

One significant difference from PhosphoRS is that the extracted peak depth is not optimized to achieve maximal differentiation. Finally, localization level is assigned by considering the ambiguity of paths, the matched fragment ions corresponding to each localized O-glycosite and the site-specific probabilities.

Data Analysis

All searches were performed on a PC running Windows 10 Education (version 1909), with two 2.20 GHz Intel Xeon Silver 4114 CPU processors with 64 GB of installed RAM. Up to 40 virtual processors were available to use for searching. Generally, 16 cores were used per search, but variations were used as described in the text. An O-glycan database of 12 common O-glycans was used for all searches[12], except for the 32-glycan database used for the urinary O-glycopeptide dataset as described in **Fig. 2l**, which was compiled using literature sources.9,40 Both glycan databases are provided as supplementary data. Data from these analyses are available in the Supplementary Information. A FASTA database of the four standard mucins used in the literature data (CD43, MUC16, PSGL-1, and Gp1ba) were used for all searches unless otherwise noted, and known signaling peptide sequences were removed from the FASTA entries.

Byonic Searching.

The standalone Byonic[28] environment (v 3.7.4, Protein Metrics) was used for all searches of the mucin O-glycopeptide dataset[15], where the maximum allowed cores is 16. O-glycan modification from the 12 O-glycan database was set to common2, common3, or common4, as indicated in the text (meaning they could occur 2, 3, or 4 times, respectively, on the same glycopeptide). The total common max value was set to match the value used for O-glycans, and the total rare max was set to 1. Other modifications were: carbamidomethyl at cysteine (+57,021644, fixed), oxidation at methionine (+15.994915, common2), and deamidation at asparagine (+0.984016, rare1). A FASTA file of the four mucin standards was used as the protein database, with reverse sequences appended as decoys by Byonic. See **Supplementary Note 2** for more discussion about databases. Cleavage specificity was set as fully semi-specific for C-terminal to R and K residues (i.e., semi-tryptic) with two missed cleavages allowed. Precursor mass tolerance was set to 10 ppm with fragment mass tolerance(s) set to 20 ppm. Fragmentation was set to HCD & EThcD for appropriate raw files, and protein FDR was set to 1%. Byonic results were processed as described in ref 15. Briefly, following each search, peptide spectral match (PSM) lists were exported as .csv files from the Byonic viewer using all columns. Filtering Byonic search results is necessary to retain only high-quality identifications and minimize false positives35; here, filtering metrics included a Byonic score greater than or equal to 200, a logProb value greater than or equal to 2, and peptide length greater than 4 residues. Localization was calculated using fragments present in identified spectra as reported in reference 15. The relaxed filtering metrics (**Supplementary Fig. 3**) used a score filter of 50 or higher and a required logProb value greater than or equal to 1.

O-Pair Search.

O-Pair Search was performed in MetaMorpheus (0.0.307), which is available at https://github.com/smith-chem-wisc/MetaMorpheus. O-Pair Search is designed to be used with high-resolution data41. The "Glyco Search" option was selected, where the O-glycopeptide search feature was enabled and the Oglycan.gdb glycan database was selected, representing the same 12 common O-glycan database used above. The "Keep top N candidates" feature was set to 50, and Data Type was set as HCD with Child Scan Dissociation set as EThcD. The "Maximum OGlycan Allowed" setting was varied as discussed in the text, where this number represents both the maximum number of O-glycan modifications that could occur on a glycopeptide candidate and the number of times each O-glycan could occur per peptide. For the majority of searches following the results obtained in Fig. 2g, the Maximum Oglycan Allowed" was set to 5 unless otherwise noted. Under Search Parameters, both "Use Provided Precursor" and "Deconvolute Precursors" were checked. Peak trimming was not enabled and Top N peaks and minimum ratio were set to 1000 and 0.01, respectively. In-Silico Digestion Parameters were set to generate decoy proteins using reversed sequences, and the initiator methionine feature was set to "Variable". The maximum modification isoforms allowed was 1024, and the minimum and maximum peptide length values were set to 5 and 60 respectively. The protease was set to semi-trypsin with 2 missed cleavages allowed, unless otherwise noted (**Supplementary Fig. 4**). The number of database partitions was set to 1 unless noted below. Precursor and product mass tolerances were 10 and 20 ppm, respectively, and the minimum score allowed was 3. The maximum number of threads, i.e., cores, was varied as described in the text, with 16 cores being the default used in this study unless otherwise noted. Modifications were set as Carbamidomethyl on C as fixed, and Oxidation on M and Deamidation on N as variable.

O-Pair Search produces two separate PSM files, one for non-glycopeptides and one for glycopeptides. The numbers of non-glycopeptide identifications were calculated by filtering the single_psm file to include only target PSMs (T) with q-values less than 0.01. The same target and q-value filterings were used for O-glycopeptide identifications in the glyco_psm file. Localization Level assignments were calculated using the provided outputs following target and q-value filtering, and all were confirmed manually for data represented in **Fig. 2a-f**. The UpSet plot in **Supplementary Fig. 5** was made using https://asntech.shinyapps.io/intervene/[42].

Entrapment databases used for **Fig. 2h** and **2i** were compiled from several different sources. The canonical mucin database (20 entries) was compiled using annotated mucins available at http://www.medkem.gu.se/mucinbiology/databases[43]. The FBS database (86 entries) was generated from data provided from Shin et al[44]. The database of CD markers, i.e., cluster of differentiation markers known to be cell surface molecules, was downloaded from the Human Protein Atlas (https://www.proteinatlas.org/)[45]. The E. coli, yeast, and mouse proteome databases were retrieved from the Uniprot Consortium[46]. The CHO secretome was downloaded from Park et al.[47] Sequences for the four mucin standards in the mixture that was analyzed were appended to each. See **Supplementary Note 3** for more discussion about the databases used. For searches performed with each of these databases, the Number of Database Partitions was set to 16, and 16 cores were also used for each search. The false discovery rate was calculated after filtering for target hits and q-value < 0.01 in the glyco_psms file, by taking the ratio of the total number of GlycoPSMs that did not originate from the four mucin standard proteins (false positives) to the total number of GlycoPSMs. This was performed when filtering based on Localization Levels as indicated in the text.

Analysis of Urinary O-glycopeptide Dataset.

Raw data is available for download from MassIVE (identifier MSV000083070) as provided in ref 9, and processed data for part of this dataset (Fraction 1 and Fraction 2) is available in ref 8. As described in the Supplemental Material in ref 31, raw files 170919_11.raw, 170921_06.raw, and 170922_04.raw correspond to Fraction 1. Raw files 170919_08.raw and 170921_03.raw are the only two files available for download from MassIVE that are from Fraction 2. We processed those sets of three and two files as Fraction 1 and Fraction 2, respectively, and then processed all ten files available for download from MassIVE, as indicated in **Fig. 2l**. Identifications from Protein Prospector and Byonic provided in the supplemental material from ref 8 were used from all three search conditions provided (described in detail in ref 8), with duplicate identifications between the searches removed. To convert Protein Prospector identifications to our Localization Levels scheme, all identifications containing "@" but not "|" were classified as Level 1 or 1b, because "@" indicates a modification assigned at a specific residue while "|" indicates an ambiguous assignment. Level 2 identifications were then added by included GlycoPSMs that included an "@", whether or not other characters indicating ambiguity were present because "@" meant at least one modification was localized.

ACKNOWLEDGMENTS

We greatly appreciate discussions with Zach Rolfs, Robert J. Millikin and other Smith group members to enhance software analysis speed and address challenges in implementing ideas. This work was supported by National Institute of Health (NIH) Grant R35 GM126914 awarded to L.M.S. and R01 GM59907 awarded to C.R.B, as well as with support from the Howard Hughes Medical Institute. N.M.R. was funded through an NIH Predoctoral to Postdoctoral Transition Award (Grant K00 CA21245403).

DATA AVAILABILITY

The data used in this manuscript are available through the Proteome-Xchange Consortium via the PRIDE partner repository48 with the dataset identifier PXD017646 (ref 15) and via MassIVE (http://massive.ucsd.edu) with identifier MSV000083070 (ref 9). Processed data using Byonic and Protein Prospector for the urinary O-glycopeptide data set was downloaded from ref 8.

CODE AVAILABILITY

O-Pair Search is available in MetaMorpheus (0.0.307), which is open-source and freely available at https://github.com/smith-chem-wisc/MetaMorpheus under a permissive license. All source code was written in Microsoft C# with .NET CORE 3.1 using Visual Studio.

REFERENCES

(1) Abrahams, J. L.; Taherzadeh, G.; Jarvas, G.; Guttman, A.; Zhou, Y.; Campbell, M. P. Recent advances in glycoinformatic platforms for glycomics and glycoproteomics. *Current Opinion in Structural Biology*. Elsevier Ltd June 2020, pp 56–69.

(2) You, X.; Qin, H.; Ye, M. Recent advances in methods for the analysis of protein o-glycosylation at proteome level. *Journal of Separation Science*. January 2018, pp 248–261.

(3) Suttapitugsakul, S.; Sun, F.; Wu, R. Recent Advances in Glycoproteomic Analysis by Mass Spectrometry. *Analytical Chemistry*. American Chemical Society 2020, pp 267–291.

(4) Riley, N. M.; Coon, J. J. The Role of Electron Transfer Dissociation in Modern Proteomics. *Analytical Chemistry*. American Chemical Society January 2018, pp 40–64.

(5) Reily, C.; Stewart, T. J.; Renfrow, M. B.; Novak, J. Glycosylation in health and disease. *Nature Reviews Nephrology*. Nature Publishing Group June 2019, pp 346–366.

(6) Brockhausen, I.; Stanley, P. Chapter 10 O-GalNAc Glycans. *Essentials Glycobiol.* **2017**, *1*, 1–9.

(7) Darula, Z.; Medzihradszky, K. F. Analysis of mammalian O-glycopeptides - We have made a good start, but there is a long way to go. *Molecular and Cellular Proteomics*. American Society for Biochemistry and Molecular Biology Inc. January 2018, pp 2–17.

(8) Pap, A.; Klement, E.; Hunyadi-Gulyas, E.; Darula, Z.; Medzihradszky, K. F. Status Report on the High-Throughput Characterization of Complex Intact O-Glycopeptide Mixtures. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (6), 1210–1220.

(9) Darula, Z.; Pap, Á.; Medzihradszky, K. F. Extended Sialylated O-Glycan Repertoire of Human Urinary Glycoproteins Discovered and Characterized Using Electron-Transfer/Higher-Energy Collision Dissociation. *J. Proteome Res.* **2019**, *18* (1), 280–291.

(10) Pap, A.; Tasnadi, E.; Medzihradszky, K. F.; Darula, Z. Novel O -linked sialoglycan structures in human urinary glycoproteins. *Mol. Omi.* **2020**, *16* (2), 156–164.

(11) Khoo, K. H. Advances toward mapping the full extent of protein site-specific O-GalNAc glycosylation that better reflects underlying glycomic complexity. *Current Opinion in Structural Biology*. Elsevier Ltd June 2019, pp 146–154.

(12) Mao, J.; You, X.; Qin, H.; Wang, C.; Wang, L.; Ye, M. a new searching strategy for the identification of O-linked glycopeptides. *Anal. Chem.* **2019**, *91* (6), 3852–3859.

(13) Ahmad Izaham, A. R.; Scott, N. E. Open database searching enables the identification and comparison of glycoproteomes without defining glycan compositions prior to searching. *bioRxiv* **2020**, 2020.04.21.052845.

(14) Huang, J.; Jiang, B.; Zhao, H.; Wu, M.; Kong, S.; Liu, M.; Yang, P.; Cao, W. Development of a Computational Tool for Automated Interpretation of Intact O - Glycopeptide Tandem Mass Spectra from Single Proteins. *Anal. Chem.* **2020**.

(15) Riley, N. M.; Malaker, S. A.; Driessen, M.; Bertozzi, C. R. Optimal Dissociation Methods Differ for N- and O-glycopeptides. *J. Proteome Res.* **2020**.

(16) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17* (5), 1844–1851.

(17) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14* (5), 513–520.

(18) Liu, X.; Hengel, S.; Wu, S.; Tolić, N.; Pasa-Tolić, L.; Pevzner, P. A. Identification of ultramodified proteins using top-down tandem mass spectra. *J. Proteome Res.* **2013**, *12* (12), 5830–5838.

(19) Frank, A. M.; Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L.; Pevzner, P. A. Interpreting top-down mass spectra using spectral alignment. *Anal. Chem.* **2008**, *80* (7), 2499–2505.

(20) Pevzner, P. A.; Dančík, V.; Tang, C. L. Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.* **2001**, *7* (6), 777–787.

(21) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; et al. Informed-Proteomics: Open-source software package for top-down proteomics. *Nat. Methods* **2017**, *14* (9), 909–914.

(22) Taus, T.; Köcher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **2011**, *10* (12), 5354–5362.

(23) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **2006**, *127* (3), 635–648.

(24) Smith, L. M.; Thomas, P. M.; Shortreed, M. R.; Schaffer, L. V.; Fellers, R. T.; LeDuc, R. D.; Tucholski, T.; Ge, Y.; Agar, J. N.; Anderson, L. C.; et al. A five-level classification system for proteoform identifications. *Nature Methods*. Nature Publishing Group October 2019, pp 939–940.

(25) Marx, H.; Lemeer, S.; Schliep, J. E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A. J. R.; Kuster, B. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* **2013**, *31* (6), 557–564.

(26) Halim, A.; Westerlind, U.; Pett, C.; Schorlemer, M.; Rüetschi, U.; Brinkmalm, G.; Sihlbom, C.; Lengqvist, J.; Larson, G.; Nilsson, J. Assignment of Saccharide Identities through Analysis of Oxonium Ion Fragmentation Profiles in LC–MS/MS of Glycopeptides. *J. Proteome Res.* **2014**, *13* (12), 6024–6032.

(27) Polasky, D. A.; Yu, F.; Teo, G. C.; Nesvizhskii, A. I. Fast and Comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *bioRxiv* **2020**, 2020.05.18.102665.

(28) Bern, M.; Kil, Y. J.; Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinformatics* **2012**, *Chapter 13*, Unit13.20.

(29) Bern, M.; Cai, Y.; Goldberg, D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **2007**, *79* (4), 1393–1400.

(30) Malaker, S. A.; Pedram, K.; Ferracane, M. J.; Bensing, B. A.; Krishnan, V.; Pett, C.; Yu, J.; Woods, E. C.; Kramer, J. R.; Westerlind, U.; et al. The mucin-selective protease StcE enables molecular and functional analysis of human cancer-associated mucins. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (15), 7278–7287.

(31) Choo, M. S.; Wan, C.; Rudd, P. M.; Nguyen-Khuong, T. GlycopeptideGraphMS: improved glycopeptide detection and identification by exploiting graph theoretical patterns in mass and retention time. *Anal. Chem.* **2019**, *91* (11), 7236–7244.

(32) Klein, J.; Zaia, J. Relative Retention Time Estimation Improves N-Glycopeptide Identifications by LC–MS/MS. *J. Proteome Res.* **2020**, *19* (5), 2113–2121.

(33) Khatri, K.; Klein, J. A.; Zaia, J. Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. *Anal. Bioanal. Chem.* **2017**, *409* (2), 607–618.

(34) Liu, M. Q.; Zeng, W. F.; Fang, P.; Cao, W. Q.; Liu, C.; Yan, G. Q.; Zhang, Y.; Peng, C.; Wu, J. Q.; Zhang, X. J.; et al. PGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nat. Commun.* **2017**, *8*, 438.

(35) Lee, L. Y.; Moh, E. S. X.; Parker, B. L.; Bern, M.; Packer, N. H.; Thaysen-Andersen, M. Toward Automated *N* -Glycopeptide Identification in Glycoproteomics. *J. Proteome Res.* **2016**, *15* (10), 3904–3915.

(36) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (11), 1719–1727.

(37) Chalkley, R. J.; Medzihradszky, K. F.; Darula, Z.; Pap, A.; Baker, P. R. The effectiveness of filtering glycopeptide peak list files for Y ions. *Mol. Omi.* **2020**, *16* (2), 147–155.

(38) Baker, P. R.; Trinidad, J. C.; Chalkley, R. J. Modification site localization scoring integrated into a search engine. *Mol. Cell. Proteomics* **2011**, *10* (7).

(39) Park, G. W.; Lee, J.; Lee, H. K.; Shin, J. H.; Kim, J. Y.; Yoo, J. S. Classification of Mucin-Type O-Glycopeptides Using Higher-Energy Collisional Dissociation in Mass Spectrometry. *Anal. Chem.* **2020**.

(40) Xu, G.; Goonatilleke, E.; Wongkham, S.; Lebrilla, C. B. Deep Structural Analysis and Quantitation of O-Linked Glycans on Cell Membrane Reveal High Abundances and Distinct Glycomic Profiles Associated with Cell Type and Stages of Differentiation. *Anal. Chem.* **2020**, *92* (5), 3758–3768.

(41) Wenger, C. D.; Coon, J. J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *J. Proteome Res.* **2013**, *12* (3), 1377–1386.

(42) Khan, A.; Mathelier, A. Intervene: A tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics* **2017**, *18* (1), 287.

(43) Lang, T.; Klasson, S.; Larsson, E.; Johansson, M. E. V.; Hansson, G. C.; Samuelsson, T. Searching the Evolutionary Origin of Epithelial Mucus Protein Components - Mucins and FCGBP. *Mol. Biol. Evol.* **2016**, *33* (8), 1921–1936.

(44) Shin, J.; Kim, G.; Kabir, M. H.; Park, S. J.; Lee, S. T.; Lee, C. Use of Composite Protein Database including Search Result Sequences for Mass Spectrometric Analysis of Cell Secretome. *PLoS One* **2015**, *10* (3), e0121692.

(45) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; et al. Tissue-based map of the human proteome. *Science (80-.).* **2015**, *347* (6220), 1260419–1260419.

(46) Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506–D515.

(47) Park, J. H.; Jin, J. H.; Lim, M. S.; An, H. J.; Kim, J. W.; Lee, G. M. Proteomic analysis of host cell protein dynamics in the culture supernatants of antibody-producing CHO cells. *Sci. Rep.* **2017**, *7* (1), 1–13.

(48) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**, *47* (D1), D442–D450.

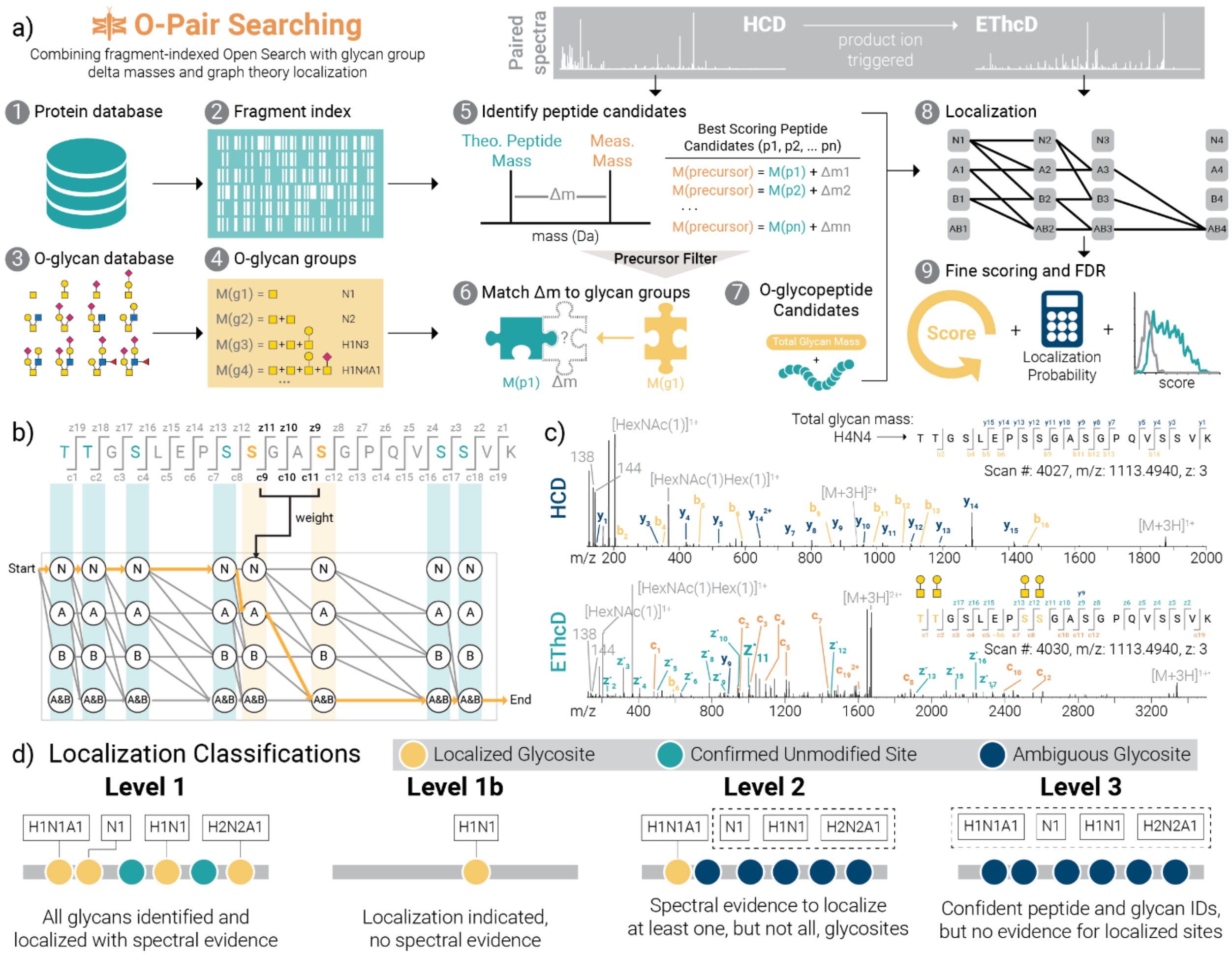
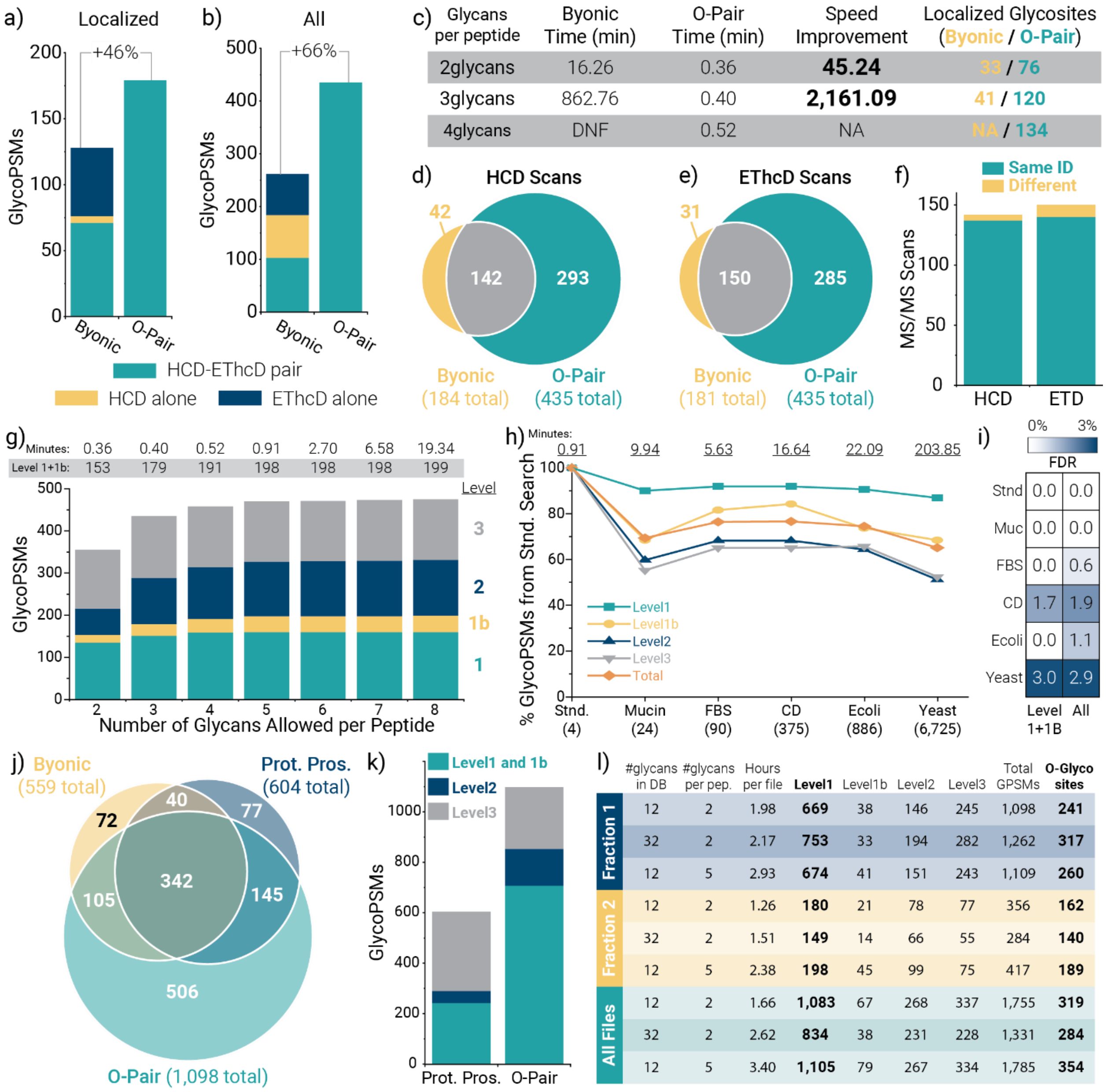

a)
O-Pair Searching
Combining fragment-indexed Open Search with glycan group delta masses and graph theory localization
Paired spectra
HCD
product ion triggered
EThcD
1 Protein database
2 Fragment index
3 O-glycan database
4 O-glycan groups
M(g1) = □ N1
M(g2) = □+□ N2
M(g3) = □+□+□ H1N3
M(g4) = □+□+□+□ H1N4A1
5 Identify peptide candidates
Theo. Peptide Mass
Meas. Mass
Δm
mass (Da)
Best Scoring Peptide Candidates (p1, p2, ... pn)
M(precursor) = M(p1) + Δm1
M(precursor) = M(p2) + Δm2
M(precursor) = M(pn) + Δmn
Precursor Filter
6 Match Δm to glycan groups
M(p1) Δm
M(g1)
7 O-glycopeptide Candidates
Total Glycan Mass
8 Localization
N1 N2 N3 N4
A1 A2 A3 A4
B1 B2 B3 B4
AB1 AB2 AB3 AB4
9 Fine scoring and FDR
Score
Localization Probability
score
b)
T T G S L E P S S G A S G P Q V S S V K
z19 z18 z17 z16 z15 z14 z13 z12 z11 z10 z9 z8 z7 z6 z5 z4 z3 z2 z1
c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19
weight
Start
N A B A&B
End
c)
HCD
[HexNAc(1)]1+
138
144
[HexNAc(1)Hex(1)]1+
[M+3H]2+
[M+3H]1+
Total glycan mass: H4N4
Scan #: 4027, m/z: 1113.4940, z: 3
m/z 200 400 600 800 1000 1200 1400 1600 1800 2000
EThcD
[HexNAc(1)Hex(1)]1+
[HexNAc(1)]1+
[M+3H]2+
[M+3H]1+
Scan #: 4030, m/z: 1113.4940, z: 3
m/z 400 800 1200 1600 2000 2400 2800 3200
d) Localization Classifications
Localized Glycosite
Confirmed Unmodified Site
Ambiguous Glycosite
Level 1
H1N1A1 N1 H1N1 H2N2A1
All glycans identified and localized with spectral evidence
Level 1b
H1N1
Localization indicated, no spectral evidence
Level 2
H1N1A1 N1 H1N1 H2N2A1
Spectral evidence to localize at least one, but not all, glycosites
Level 3
H1N1A1 N1 H1N1 H2N2A1
Confident peptide and glycan IDs, but no evidence for localized sites

Figure 1. O-Pair Search through MetaMorpheus for fast and confident identification of O-glycopeptides. a) The workflow describes processing steps in the O-Pair Search strategy, which generates a fragment ion index [1, 2] and O-glycan groups [3, 4] from user defined protein and O-glycan databases, respectively. Using an ultrafast, fragment-index-enabled open modification search [5] paired with a match of delta masses to aggregate glycan mass combinations [6] enables identification of O-glycopeptide candidates from HCD spectra [7]. Paired EThcD spectra are then used for graph theory-based localization calculations to rapidly assign modification sites for all glycans comprising the O-glycan group [8]. Finally, more detailed re-scoring of spectra, localization probability calculations, and false discovery rate corrections are performed before returning identifications to the user [9]. **b)** A demonstration of graph theory-based localization using a hypothetical example of an O-glycopeptide TTGSLEPSSGASGPQVSSVK from human mucin-type O-glycoprotein CD43 (leukosialin), which has 8

potential O-glycosites. Here we consider how graph theory determines O-glycosites using c/zdot fragments present in EThcD spectra when two glycans (termed A and B for the sake of demonstration) are presented as modifications. **c)** An example of paired HCD and EThcD spectra for quadruply-O-glycosylated TTGSLEPSSGASGPQVSSVK, showing a Level 1 identification where all calculated glycan mass shifts can be confidently localized to discrete residues. Note, no fragments in the HCD spectrum retain any glycan masses. Rather, the thorough peptide backbone fragmentation without glycan retention shows how the sequence was confidently retrieved with a defined mass shift matching a combination of O-glycans. The subsequent EThcD spectrum then enables localization of all 4 O-glycosites (gold) even with the presence of 4 other unmodified potential sites. D) O-Pair Search defines levels of localization for each GlycoPSM. A Level 1 assignment indicates that all glycans can be unambiguously localized to single S or T residues using spectral evidence, while Level 1b also indicates localization in instances when spectral evidence is lacking (e.g., only one possible modification site). Level 2 localizations have at least one glycan, but not all, localized to a single S or T. Level 3 GlycoPSMs include the remaining pool of identifications, where peptide sequence and glycan aggregate mass are confidently assigned, but no individual glycan can be localized to a specific residue. Note, “H”, “N”, and “A” represent hexose, HexNAc, and Neu5Ac, respectively.



Glycans per peptide	Byonic Time (min)	O-Pair Time (min)	Speed Improvement	Localized Glycosites (Byonic / O-Pair)
2glycans	16.26	0.36	**45.24**	33 / 76
3glycans	862.76	0.40	**2,161.09**	41 / 120
4glycans	DNF	0.52	NA	NA / 134

	#glycans in DB	#glycans per pep.	Hours per file	**Level1**	Level1b	Level2	Level3	Total GPSMs	**O-Glyco sites**
Fraction 1	12	2	1.98	**669**	38	146	245	1,098	**241**
	32	2	2.17	**753**	33	194	282	1,262	**317**
	12	5	2.93	**674**	41	151	243	1,109	**260**
Fraction 2	12	2	1.26	**180**	21	78	77	356	**162**
	32	2	1.51	**149**	14	66	55	284	**140**
	12	5	2.38	**198**	45	99	75	417	**189**
All Files	12	2	1.66	**1,083**	67	268	337	1,755	**319**
	32	2	2.62	**834**	38	231	228	1,331	**284**
	12	5	3.40	**1,105**	79	267	334	1,785	**354**

Figure 2. Performance of O-Pair Search for O-glycopeptide characterization. Comparing the number of **a)** localized and **b)** total glycopeptide spectral matches (GlycoPSMs) returned from Byonic and from O-Pair Search for HCD-pd-EThcD data collected from StcE digestions of four recombinant mucin standards. Note, only Level 1 and 1b identifications are considered for the localized O-Pair Search data, and 3 glycans per peptide were allowed for both searches. Byonic identifications are grouped into HCD-EThcD pairs (where paired scans identified the same O-glycopeptide), HCD alone, and EThcD alone. The latter two cases are where an identification came only from an HCD scan or EThcD scan, but the other spectrum in

the pair did not return a hit. O-Pair Search improves the number of localized and total identifications by 46% and 66% over Byonic, respectively. **c)** The table compares the search times required for Byonic and O-Pair Search when considering 2, 3, and 4 glycans per peptide. Note, the 4 glycans per peptide for Byonic was canceled after approximately 33,000 minutes of search time (~3.5 weeks) because it had not advanced in reported search progress for over one week. The number of localized glycosites identified by the searches is also provided for comparison, which correspond to the GlycoPSMs in panel a and Supplemental Fig. 3. In addition to more than doubling the number of total identified spectra, O-Pair Search identified the majority of scans that Byonic returned as GlycoPSMs for both **d)** HCD and **e)** EThcD scans, and **f)** the overwhelming majority (~95%) of the shared identified scans mapped to the same glycopeptide. **g)** O-Pair Search enabled consideration of more glycans per peptide while keeping search times reasonable. **h)** O-Pair Search also allowed the use of several different protein database backgrounds much larger in size without untenable search time increases. Here, "Total" indicates all identifications, i.e., the sum of all Localization Level identifications. **i)** Use of entrapment databases with proteins not present in the sample did not inflate false discovery rates above approximately 1-3%. **j)** O-Pair Search was used to process files from a published urinary O-glycopeptide study that previously reported Protein Prospector (Prot. Pros.) and Byonic results. O-Pair Search nearly doubled the total number of GlycoPSMs from either search engine, identifying ~90% of spectra shared by the two search algorithms while providing an additional 506 GlycoPSMs not reported by either. **k)** Protein Prospector reports localized glycosites, which we converted into our Localization Level system and compared with O-Pair results. **l)** Results from several O-Pair searches of Fraction 1 (three files), Fraction 2 (two files), and all ten files available from the urinary O-glycopeptide study.

SUPPLEMENTARY INFORMATION

This Supplementary Information includes Supplementary Figs. 1-18, Supplementary Notes 1-4, and Supplementary Tables 1-3:

Supplementary Fig. 1: Potential glycosites per theoretical peptide digested from standard mucin proteins in this study

Supplementary Fig. 2: Comparing Byonic and O-Pair Search when allowing 2 glycans per peptide

Supplementary Fig. 3: Comparing Byonic and O-Pair Search when relaxing Byonic filtering metrics for a 3-glycans-per-peptide search

Supplementary Fig. 4. Comparing glycosite heterogeneity between O-Pair Search and Byonic

Supplementary Fig. 5: Overlap of identifications when allowing for more glycans per peptide

Supplementary Fig. 6: Non-modified peptide identifications

Supplementary Fig. 7: Searching with various glycan databases

Supplementary Fig. 8: Elution times correlate for related glycoforms of the same peptide sequence

Supplementary Fig. 9: Visualizing eluting isoforms of localized glycopeptides

Supplementary Fig. 10: Re-evaluating ETD and EThcD fragmentation data using O-Pair Search

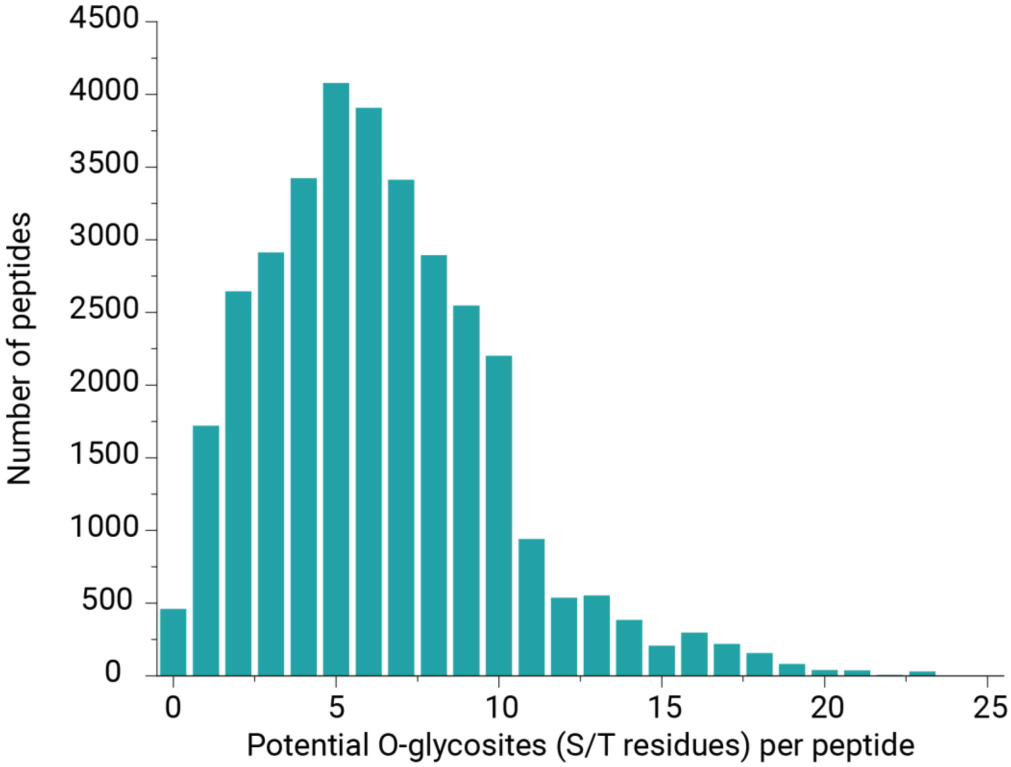
Supplementary Fig. 11: Number of localized glycosites per peptide

Supplementary Fig. 12: Search speed benefits with O-Pair Search remain even with fewer cores

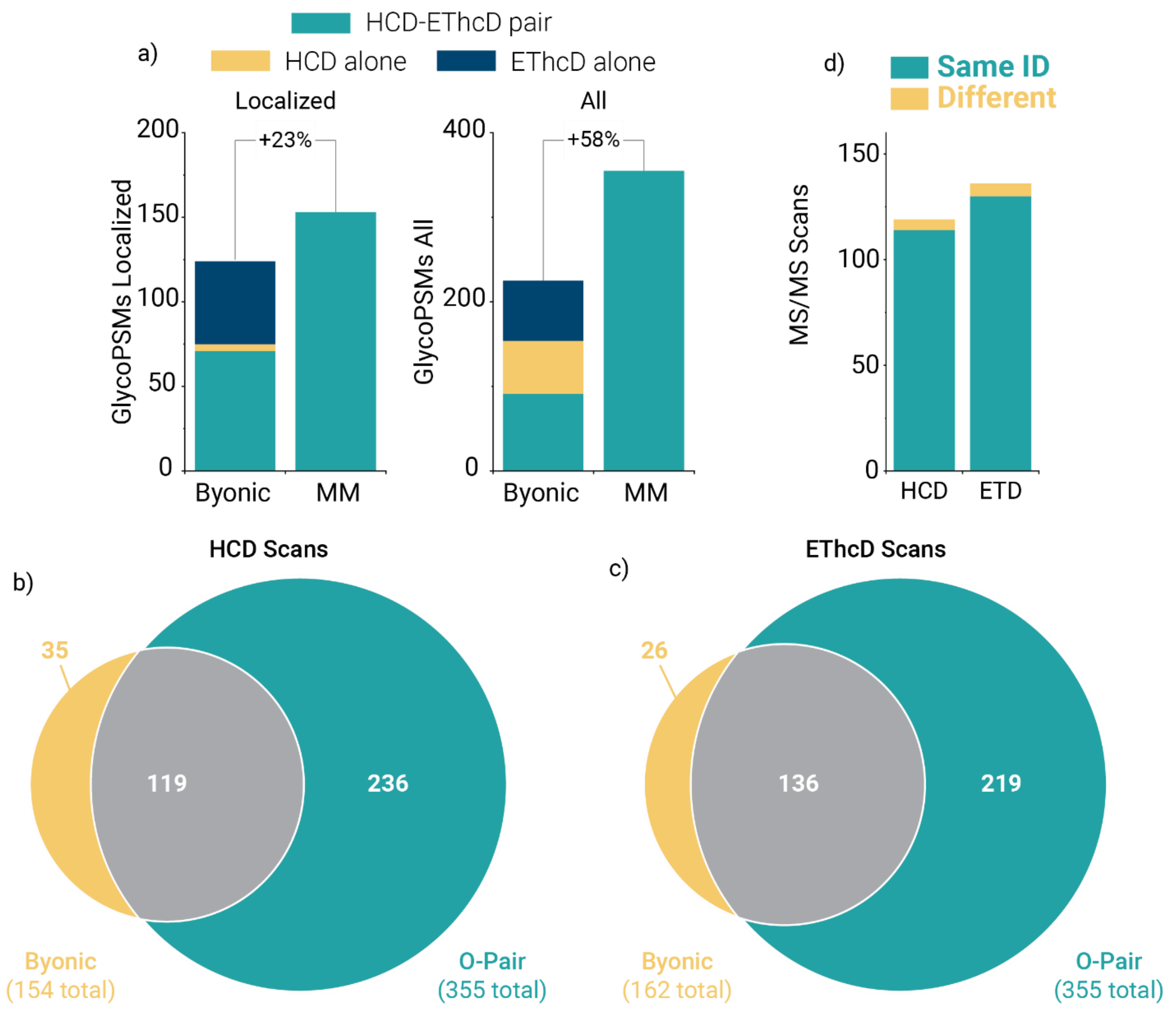
Supplementary Fig. 13: Identification of O-glycosites in four standard mucins

Supplementary Fig. 14: Processing HCD and stepped collision energy HCD data with O-Pair Search

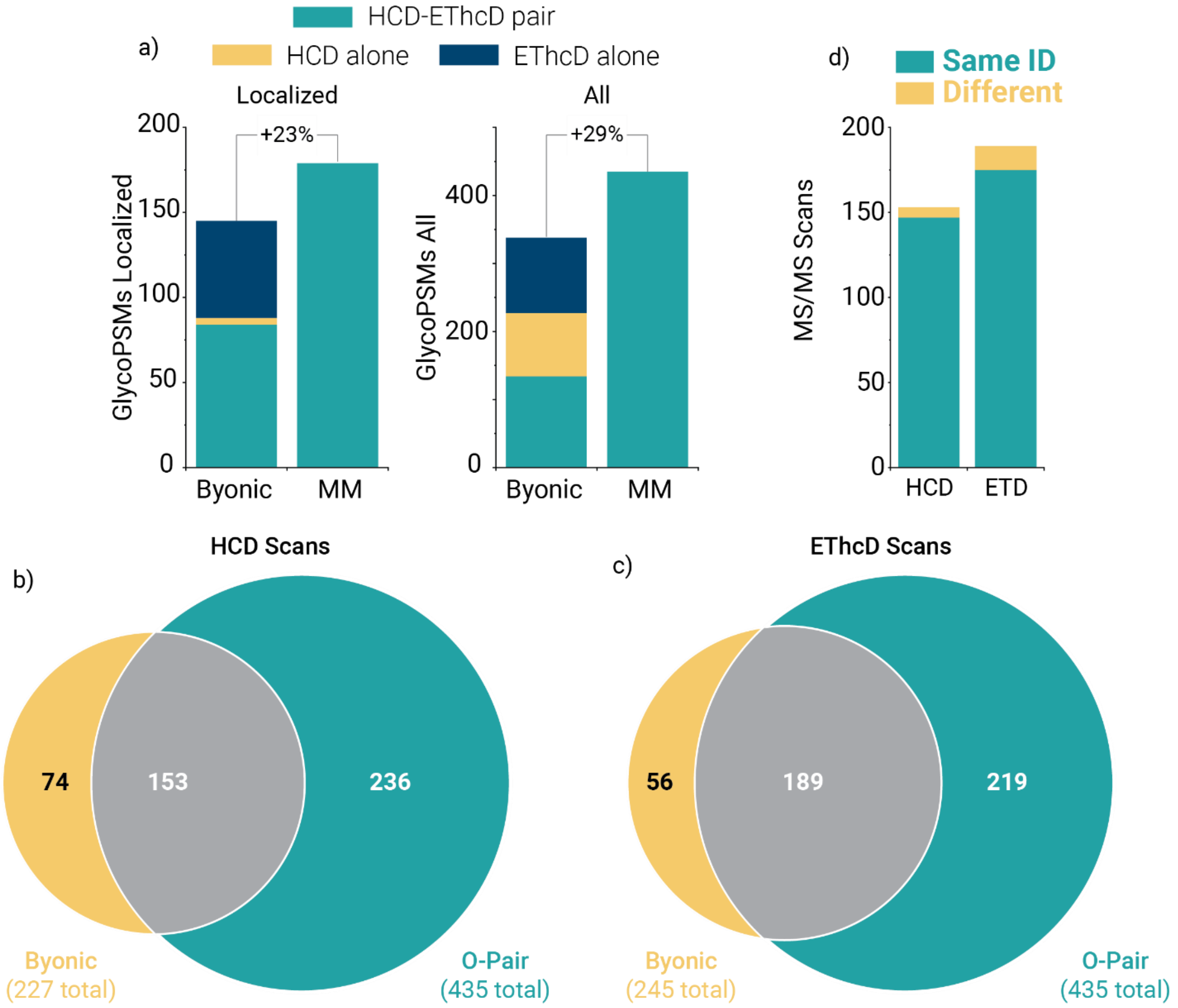
Supplementary Fig. 15: Entrapment FDR with Byonic and O-Pair Search



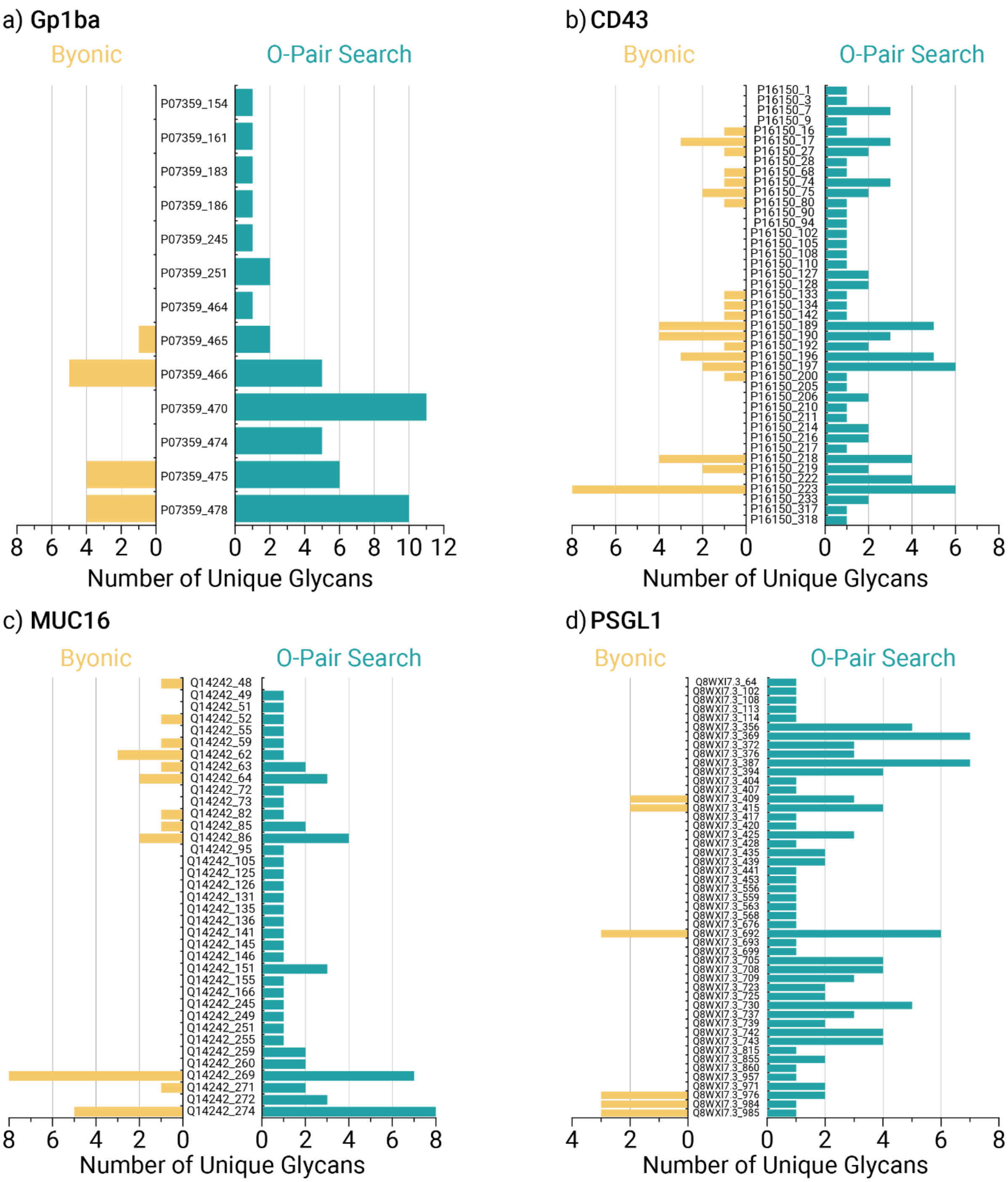
Supplementary Figure 1. **Potential glycosites per theoretical peptide digested from standard mucin proteins in this study.** On average, peptides derived from mucins can have approximately 6 serine/threonine residues.



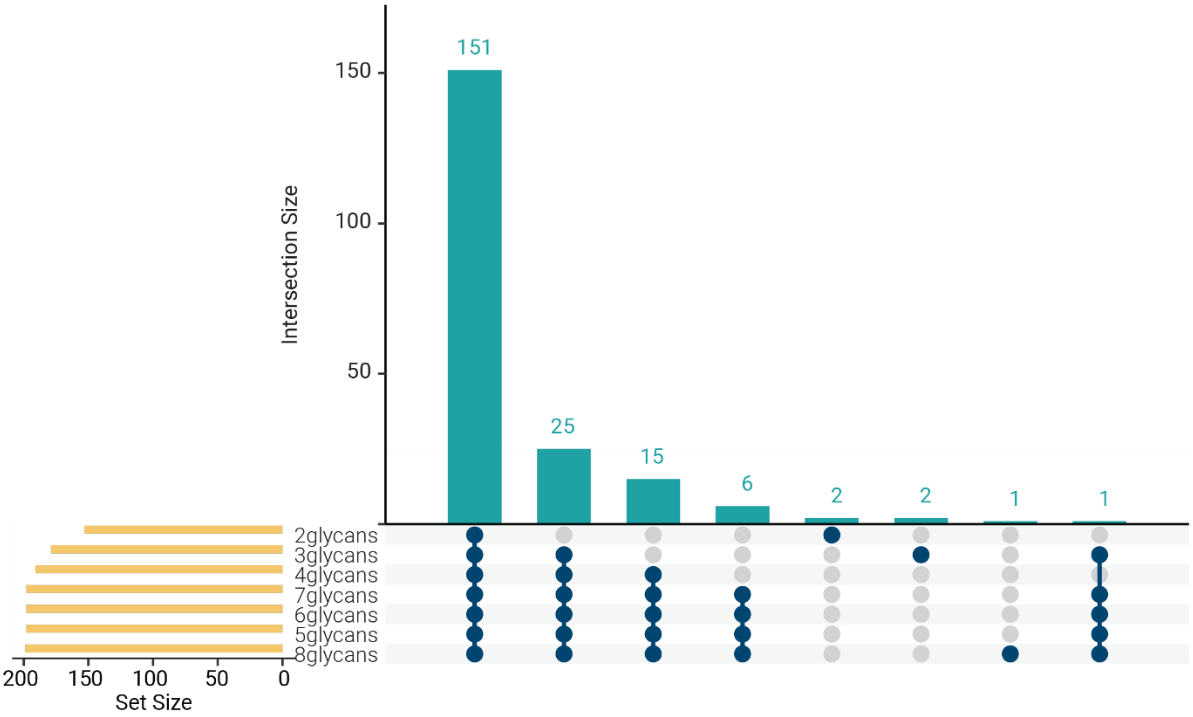
Supplementary Figure 2. Comparing Byonic and O-Pair Search when allowing 2 glycans per peptide. a) O-Pair searching identifies more localized and total (all) GlycoPSMs when allowing for 2 glycans per peptide from a 12-glycan database. Byonic identifications are grouped into HCD-EThcD pairs (where paired scans identified the same O-glycopeptide), HCD alone, and EThcD alone. The latter two cases are where an identification came only from an HCD scan or EThcD scan, but the other spectrum in the pair did not return a hit. O-Pair Search identified the majority of scans that Byonic returned as GlycoPSMs for both **b)** HCD and **c)** EThcD scans, and **d)** the overwhelming majority (~95%) of the shared identified scans mapped to the same glycopeptide.



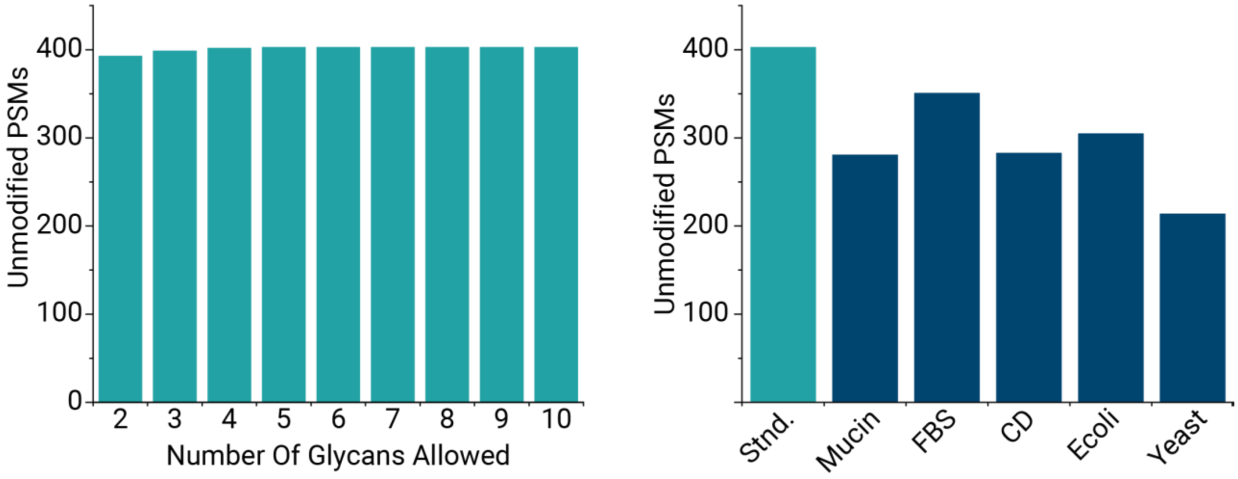
Supplementary Figure 3. Comparing Byonic and O-Pair Search when relaxing Byonic filtering metrics for a 3-glycans-per-peptide search. **a)** O-Pair searching identifies more localized and total (all) GlycoPSMs when allowing for 3 glycans per peptide from a 12-glycan database. Byonic identifications were filtered for scores > 50 and logProb values > 1 (compared to 200 and 2, respectively, as presented in the main text **Fig. 2**). are grouped into HCD-EThcD pairs (where paired scans identified the same O-glycopeptide), HCD alone, and EThcD alone. The latter two cases are where an identification came only from an HCD scan or EThcD scan, but the other spectrum in the pair did not return a hit. O-Pair Search identified the majority of scans that Byonic returned as GlycoPSMs for both **b)** HCD and **c)** EThcD scans, and **d)** the overwhelming majority (~95%) of the shared identified scans mapped to the same glycopeptide. The majority of additional HCD identifications from the relaxed Byonic filtering are not those that overlap with O-Pair Search identifications, while the overlapping identifications increased more than non-overlapping identifications for EThcD scans. This further supports previous findings that EThcD scans are underscored relative to HCD scans in Byonic, with HCD scans having a greater likelihood of being assigned to less confident identifications. This observation also further highlights the benefits of using HCD and EThcD spectral pairs when assigning identifications, as O-Pair Search does.



Supplementary Figure 4. Comparing glycosite heterogeneity between O-Pair Search and Byonic. Localized glycosites from GlycoPSMs from a) Gpb1a, b) CD43, c) MUC16, and d) PSGL-1 are shown as reported by Byonic and O-Pair Search, where 3 glycans were allowed per glycosite (Fig. 2a). The number of unique glycan compositions at each site is reported for Byonic (yellow) and O-Pair Search (teal). Note, the difference in glycosites reported here and data in Supplementary Fig. 13 is the difference between searching with 3 and 5 glycans per peptide allowed with O-Pair Search.

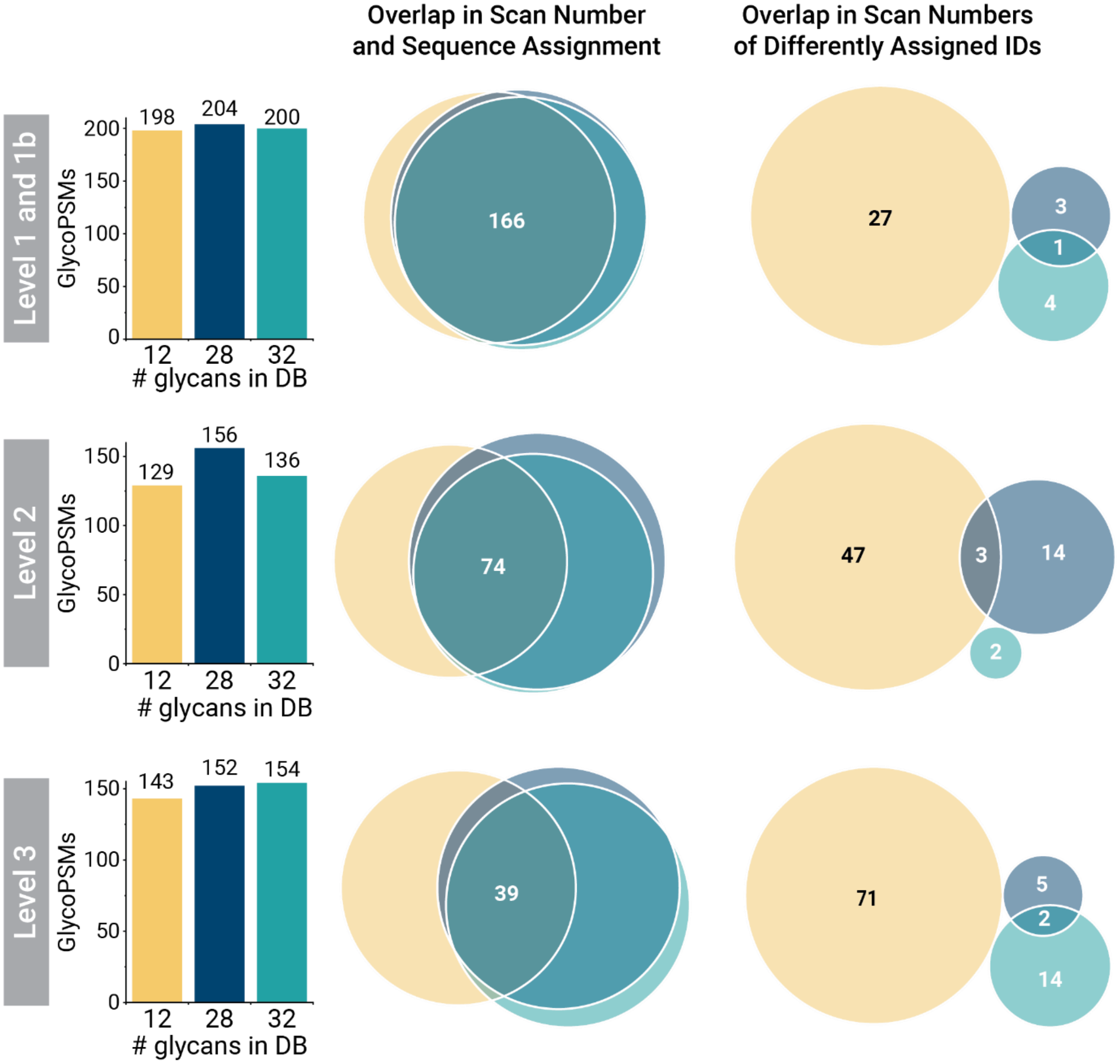


Supplementary Figure 5. Overlap of identifications when allowing for more glycans per peptide. The UpSet plot, which functions as a multi-dimensional Venn diagram depicting intersection size between different group as the top bar graph, shows that the core number of identifications stays the same between searches that allow for more glycans per peptide, with additional identifications remaining in the intersection of searches with progressively more glycans allowed. The total identifications count (Level1 and 1b) for each search are shown as “set size” in the bar graph in the bottom left corner.

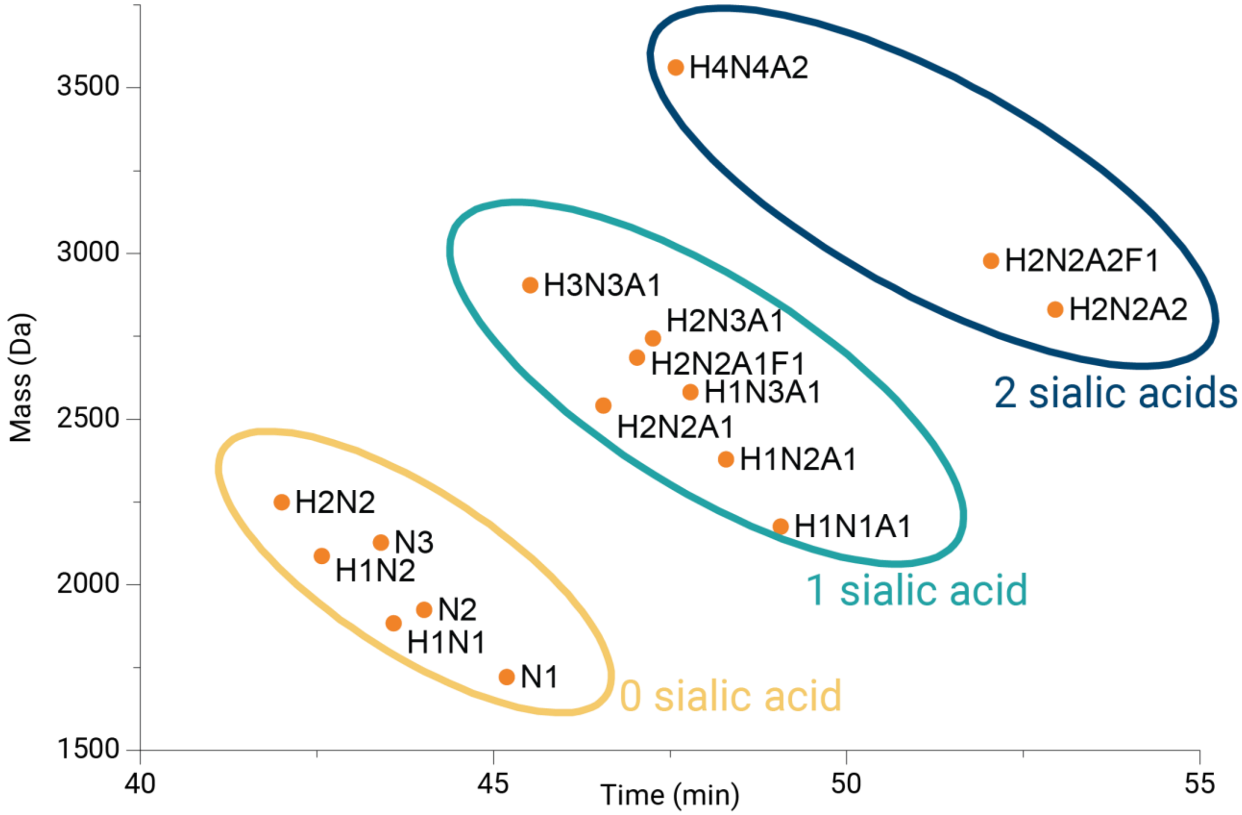


Supplementary Figure 6. Non-modified peptide identifications. O-Pair Search also returns identifications for non-modified peptides, i.e., standard peptide spectral matches (PSMs) that include oxidized methionine variable modifications. The number of non-modified PSMs are shown for a) the searches for the variable number of O-glycans allowed per peptide and b) the searches with six entrapment databases (with the database of 4 standard mucins provided as a reference in aqua). The size/complexity of the entrapment database caused fewer identifications for non-modified peptides, similar to O-glycopeptide identification (**Fig. 2h**).

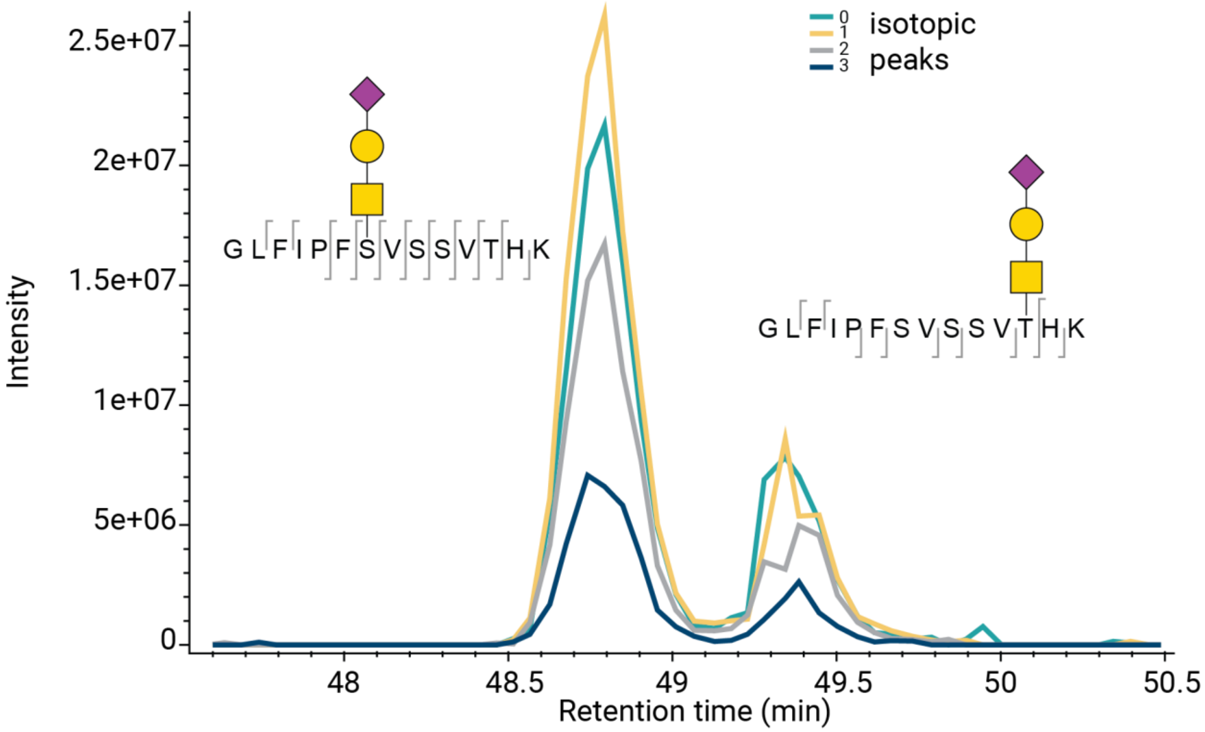
	12 glycan DB	28 glycan DB	32 glycan DB
Total GlycoPSMs	470	512	490
Search Time (min)	0.93	27.28	51.49



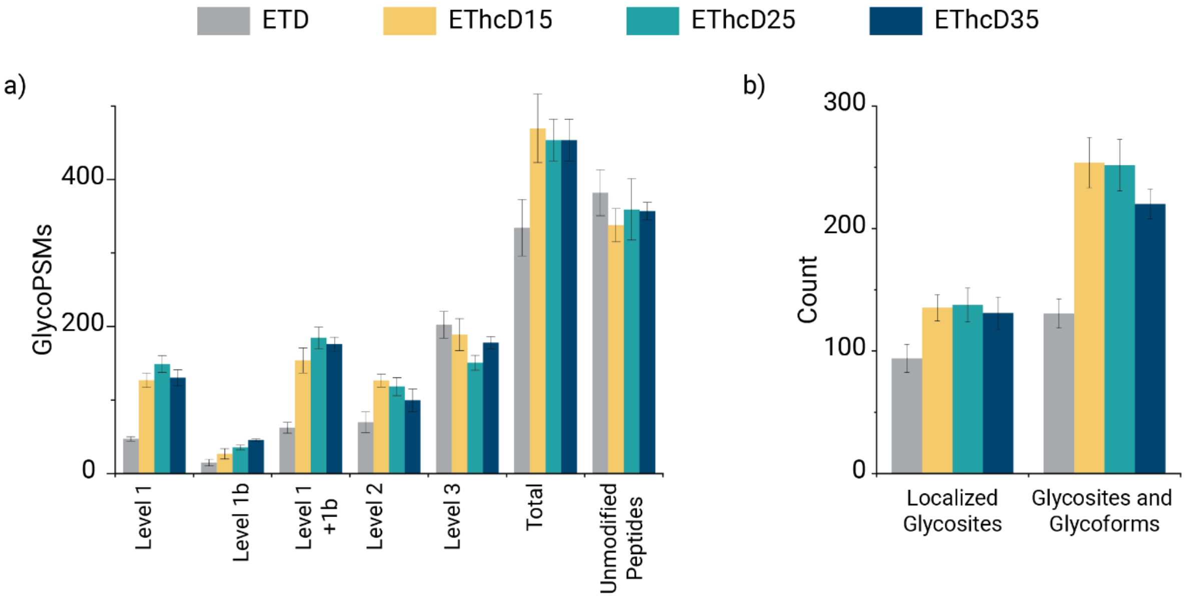
Supplementary Figure 7. Searching with various glycan databases. Identifications were made with O-Pair Search using glycan databases (DBs) of 12 (yellow), 28 (navy), and 32 (teal) glycans and protein database with the 4 mucin sequences. The number of total glycoPSMs and the search times for each search are provided in the table at the top. Bar graphs compare numbers of identifications from Level 1 and 1b, Level 2, and Level 3 classifications. Venn diagrams on the left for each level indicate how many identifications in the three different searches had the same glycopeptide assigned to the same spectral pair. Spectral pairs that returned identifications unique to one condition were also compared by looking at overlap in their spectrum numbers to determine how many spectra are assigned different identifications under glycan database search conditions (Venn diagrams on the right). Here, an overlap of 3 between conditions shows that three spectral pairs were identified as different glycopeptides between the two searches. Note, Venn diagrams on the right are not to-scale with those on the left. These data demonstrate that larger glycan databases can still be searched in reasonable time frames with O-Pair Search and return similar number of identifications. Interestingly, Level 1, 1b, and Level 2 identifications were highest with the 28-glycan database, but Level 3 identifications increased with glycan database size. This indicates that the "right" glycan database size exists for a given dataset and that identifications do not simply increase with the more glycans considered, except perhaps for Level 3 identifications where peptide and glycan mass combinations could increase as the combinatorial space inflates. The high degree of overlap in Level 1 identifications between search conditions confirms their high quality. Indeed, the one spectral pair that was assigned to different sequences for the 28-glycan and 32-glycan searches showed very similar identifications: TLTLNFTISNLQYSPDMGKGS (28 glycan) and TLTLNFTISNLQYSPDMGKG (32 glycan). As expected, overlap gets worse amongst all searches as the identifications get less confident (Level 2 and Level 3). The lack of overlap between the 12-glycan database and the other two searches suggests that the combinatorial space considered for a glycan search can alter the sequences assigned. Even so, the Venn diagrams to the right show that the lack of overlap is not because the same spectra are assigned different identifications (a relatively rare occurrence), but rather that different pools of spectra are identified when considering different glycan combinatorial space. The ability to search O-glycopeptide data with large glycan databases like this has been largely unexplored due to the computational complexities addressed in this manuscript. O-Pair Search now makes these comparisons more feasible and will enable new investigations into relationships between protein database, glycan database, and spectral scoring metrics.



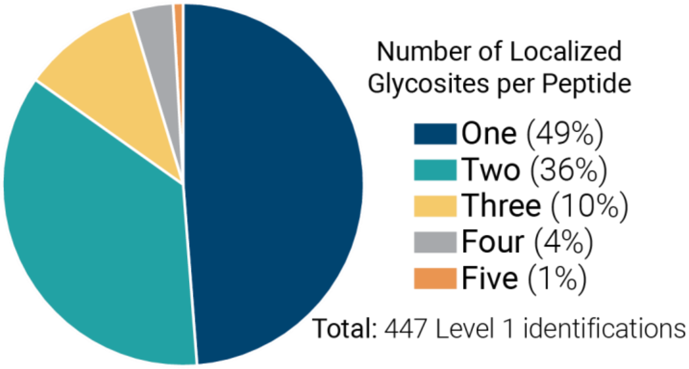
Supplementary Figure 8. Elution times correlate for related glycoforms of the same peptide sequence. The peptide "GLFIPFSVSSVTHK" from PSGL-1 was identified multiple times with different glycan compositions. The retention times cluster with the feature of glycans. The glycopeptides are grouped by the number of sialic acids. Glycans that contain more sialic acids elute later (between different groups); glycans that contain more neutral sugars tend to elute earlier (within one group).



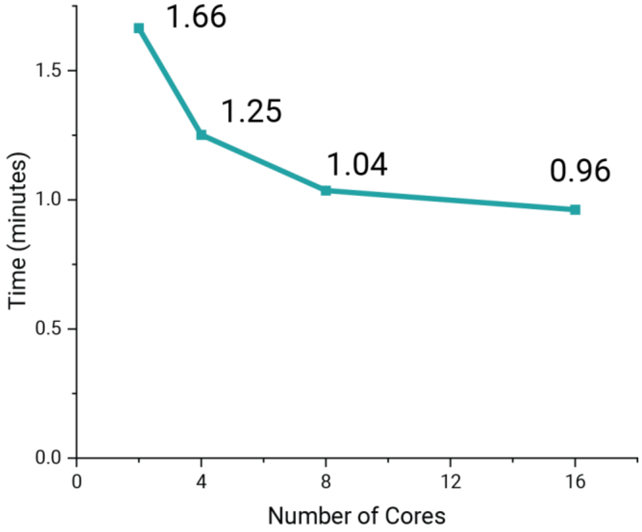
Supplementary Figure 9. Visualizing eluting isoforms of localized glycopeptides. Two glycopeptides, both with the aggregate mass of "GLFIPFSVSSVTHK-(H1N1A1)", are positional isoforms identified with O-Pair Search. The presence of two different elution profiles supports the presence of the two different glycoforms localized by O-Pair Search. The extracted ion current for the mass is shown for the four most abundant isotopic peaks for each glycoform.



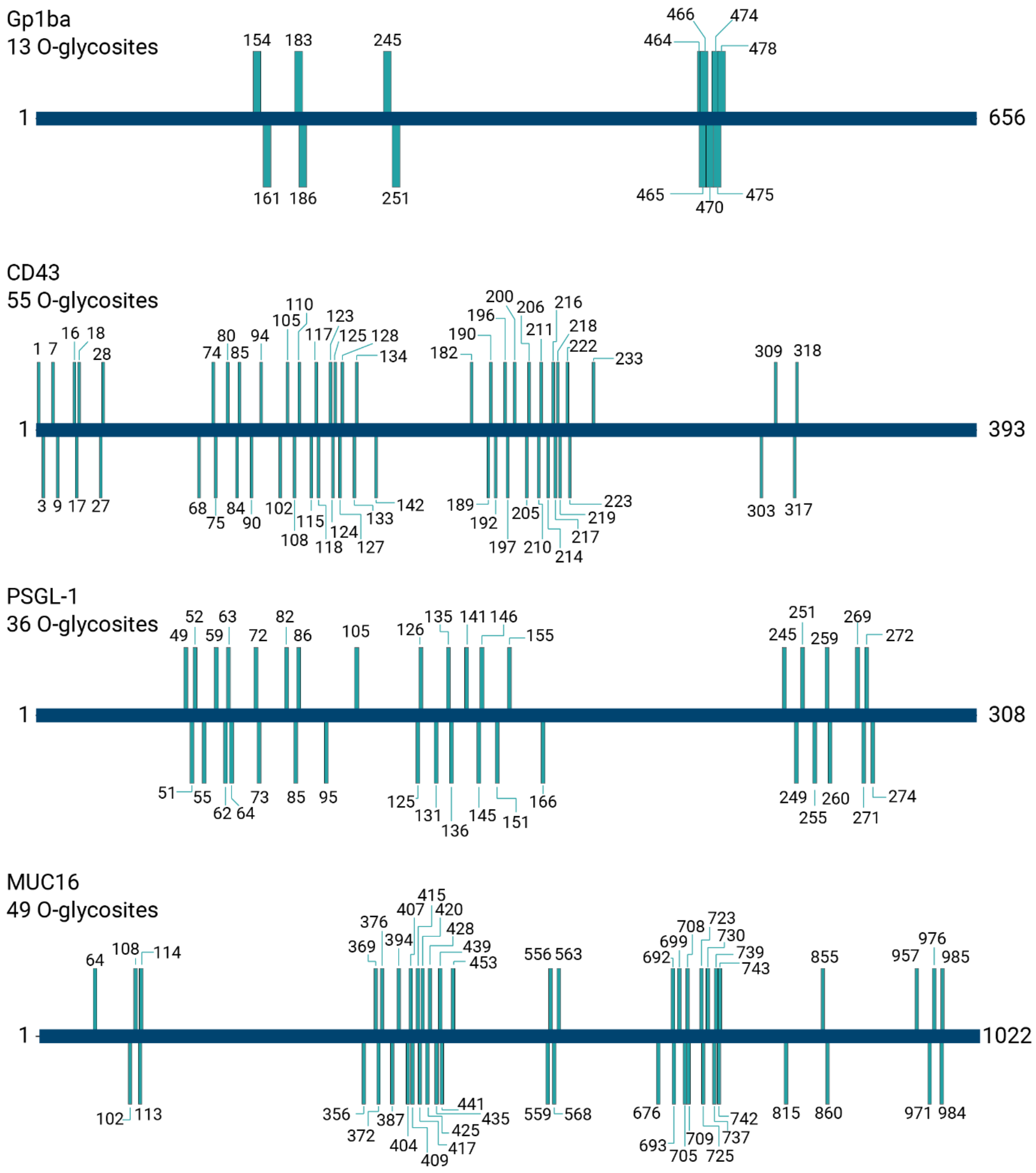
Supplementary Figure 10. **Re-evaluating ETD and EThcD fragmentation data using O-Pair Search.** Comparisons of electron-driven dissociation methods investigated in Main Text ref 11 are shown using O-Pair Search with Localization Level information. GlycoPSMs are shown in **a)**, including the combined Level 1 and 1b identifications, the total number of identifications, and the number of non-modified peptides detected per method. Unsurprisingly, ETD has a much higher proportion of Level 3 identifications relative to Level 1 than the EThcD methods, because non-dissociative electron transfer prevents the generation of peptide backbone fragments that are needed to localize glycosites. Even though EThcD15 generates the most total GlycoPSMs, EThcD25 has the advantage for fully localized (Level 1) identifications. **b)** The advantage in localized GlycoPSMs afforded by EThcD25 provides a slight gain in the number of localized glycosites, although the difference is minimal compared to previous analyses with Byonic. The mid-range supplemental activation energies (EThcD15 and EThcD25) provide the most localized glycosites and the most glycoforms, i.e., the combination of glycosites and the different glycans observed on them.



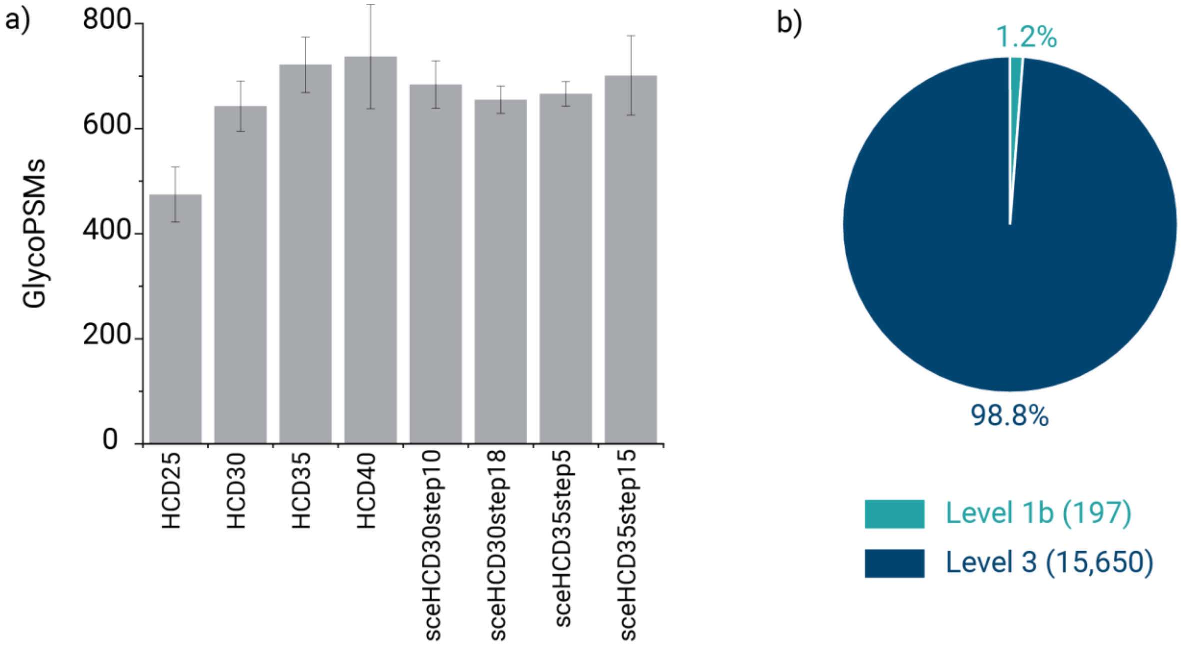
Supplementary Figure 11. Number of localized glycosites per peptide. Identifications from the three EThcD25 datasets were pooled and filtered for Level 1 identifications only. This pie graph shows the distribution of the 447 total Level 1 identifications that had 1-5 localized glycosites.



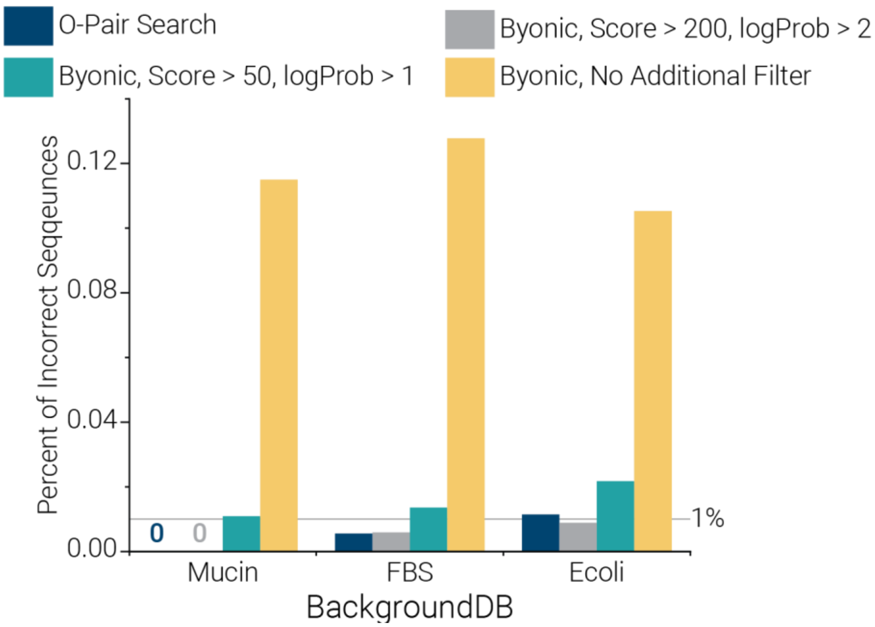
Supplementary Figure 12. Search speed benefits with O-Pair Search remain even with fewer cores. Here, the time needed to complete a full analysis of a raw file analogous to that shown in Fig. 2c and Fig. 2g is shown for 2, 4, 8, and 16 cores when allowing for 5 glycans per peptide with a glycan database of 12 glycans. O-Pair Search speed translates to any computer used for searching, not just hyperthreaded computers.



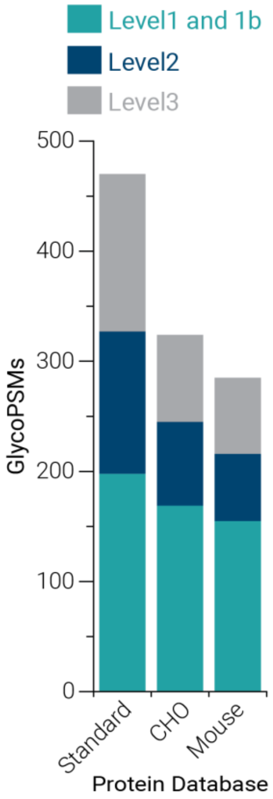
Supplementary Figure 13. **Identification of O-glycosites in the four standard mucins**. Searching with 5 glycans allowed from a 12 O-glycan database identifies 153 O-glycosites, with each protein extensively modified. Here, blue horizontal bars represent the protein sequences with N- and C-terminal residues numerically labeled, and green vertical bars indicate O-glycosites (residue numbers provided) localized with O-Pair Search. For comparison (but not depicted in the figure), a Byonic search with 3 glycans allowed identified 96 total O-glycosites with evidence for 41 of them as localized: 4 localized of 7 identified O-glycosites in Gp1ba, 19 localized of 37 identified O-glycosites in CD43, 12 localized of 17 identified O-glycosites in PSGL-1, and 6 localized of 35 identified O-glycosites in MUC16, Note, signal peptides were removed from the protein sequences prior to searching.



Supplementary Figure 14. **Processing HCD and stepped collision energy HCD (sceHCD) data with O-Pair Search.** Even though optimal searching conditions use paired HCD and EThcD scans to facilitate glycosite localization, HCD and sceHCD can also be searched. Here, searches were performed on published data from the same digestion of mucin standard proteins used in the main text, but run with scouting HCD scans followed by triggered HCD or sceHCD scans of with different collision energy settings[1]. **a)** O-Pair Search assigns glycoPSMs from an open search but is unable to use graph theory for glycosite localization because peptide backbone fragments largely fail to retain the glycan mass under HCD fragmentation. Values are averages of triplicate injections (3 separate files), and error bars show one standard deviation. **b)** Approximately 1% of identifications are classified as Level 1b because only one plausible glycoform exists, but the vast majority are Level 3 identifications, meaning no glycosite information can be discerned. Regardless, this makes the rapid searching provided by O-Pair Search a viable option for HCD-only studies, as well. Although not shown here, O-Pair Search can also process traditional data acquisition regimes, i.e., data-dependent acquisition without the scouting HCD scans.



Supplementary Figure 15. **Entrapment FDR with Byonic and O-Pair Search.** Three of the entrapment background databases used to test O-Pair Search were also used to test Byonic. These were selected because they could be completed with reasonable search times with Byonic when allowing for 2 glycans per peptide. Three levels of Byonic filtering are shown: no filtering, which would be the equivalent of Byonic's standard FDR calculation (yellow), filters to keep only glycoPSMs with Byonic Scores > 50 and logProb values > 1 (teal), and filters to keep glycoPSMs with Byonic Scores > 200 and logProb values > 2 (gray). This comparison shows how necessary post-Byonic-search filtering is to assure quality glycoPSMs. O-Pair Search data (navy) is the same as reported in Fig. 2i (All) and is the result of filtering glycoPSM q-values to 0.01.



CHO background O-glycopeptides

Leukosialin	135
P-selectin glycoprotein ligand 1	88
Mucin-16 segment	51
Platelet glycoprotein Ib alpha chain	47
ATP-dependent RNA helicase DDX42	1
Beta-klotho	1
Ankyrin-2	1

CHO background non-modified peptides

Mucin-16 segment	185
Platelet glycoprotein Ib alpha chain	90
P-selectin glycoprotein ligand 1	22
Leukosialin	10
Serine protease HTRA1	3
Lactadherin (Fragment)	1
Endoplasmic reticulum chaperone BiP	23
Procollagen-lysine,2-oxoglutarate 5-dioxygenase 1	2
Endoplasmin	1
Lactadherin	4
POM121-like protein 2	1
Melanoma inhibitory activity protein 3	1
Aldehyde oxidase	1
Nidogen-1	1
Collagen alpha-2(V) chain	1
Leucine-rich repeat and fibronectin type-III domain-containing protein 2	1
Uncharacterized protein	1
HTRA2	1
40S ribosomal protein S15a	1
Alpha-mannosidase	1
CTGF	1

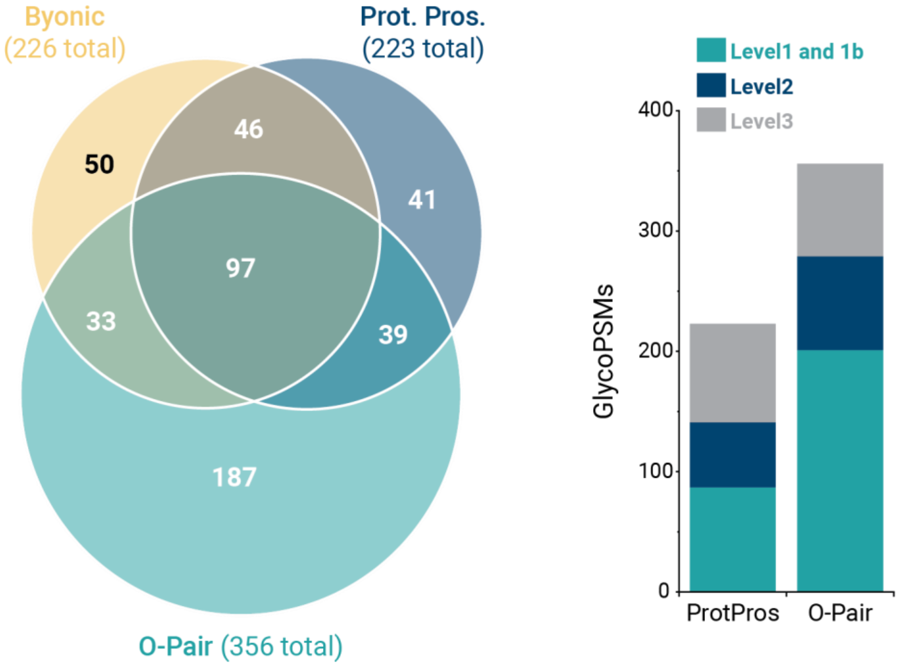
Mouse background O-glycopeptides

Leukosialin	126
P-selectin glycoprotein ligand 1	77
Mucin-16 segment	37
Platelet glycoprotein Ib alpha chain	38
Ig gamma-1 chain C region, membrane-bound form	4
Histone acetyltransferase KAT6A	1
Versican core protein	1
Disks large homolog 3	1

Mouse background non-modified peptides

Mucin-16 segment	89
Platelet glycoprotein Ib alpha chain	25
P-selectin glycoprotein ligand 1	3
Leukosialin	5
Serine protease HTRA1	3
Endoplasmic reticulum chaperone BiP	13
Endoplasmin	1
Progranulin	1
Heat shock-related 70 kDa protein 2	1

Supplementary Figure 16. **O-Pair Searches using expression system background proteomes.** Both CHO cells and NS0 murine cells were used as expression systems for the recombinant mucins used in the dataset searched here. These searches were analogous to approach used with background proteomes used for entrapment FDR calculations: the four mucin standard sequences were appended either a CHO secretome protein database (2,370 entries)[2] and the entire mouse proteome (17,030 entires). Five glycans per peptide were allowed from a 12-glycan database. The searches took ~57 minutes and ~463 minutes for CHO and mouse backgrounds, respectively. The bar graph on the left shows the number of identifications delinated by Localization Level from the standard protein file (4 recombinant mucins) and the two searches with expression system backgrounds. Lists show the proteins identified in the two searches as O-glycopeptides and non-modified peptides. Identifications are shown in black or gray for either the recombinant mucins or proteins from the background proteomes, respectively.



Supplementary Figure 17. **Comparing O-Pair Search with Byonic and Protein Prospector for Fraction 2 of the urinary O-glycopeptide dataset.** O-Pair Search has a high degree of overlap in identified spectra with the other two search algorithms while returning a greater number of unique identifications (left). We converted Protein Prospector to our Localization Level classification and compared to O-Pair Search (right).

SUPPLEMENTARY NOTE 1

Computational complexity of searching one MS2 experimental spectrum	localization	non localization
Traditional narrow search	$\mathrm{O}\,(n * \sum_{i=0}^{m} \frac{(s)!}{(s-i)!*i!} * g^{i})$	$\mathrm{O}\,(n * \sum_{i=0}^{m} \frac{((g-1) + i)!}{(g-1)!*i!})$
ion-indexed open search	brute force localization $\mathrm{O}\,(x + k * r * \frac{s!}{(s-m)!})$	$\mathrm{O}\,(x + k)$
	graph-based localization $\mathrm{O}\,(x + k * r * \ d)$	

Supplementary Table 1. Computation complexity analysis

n is the average number of theoretical peptides within the specified precursor mass tolerance.
m is the maximum number of glycan allowed on a single peptide.
s is the number of S/T amino acids within a single peptide.
g is the number of glycan types.
k is the count of top peptide candidates. In general, *k* is always much smaller than *n*.
r is the number of glycan groups with the same mass. (Supplementary Table 2.)
x is the number of peaks in an MS2 experimental spectrum, which is used to represent the constant ion indexing time and can be ignored in general.
d is the depth of the graph. $d = \sum_{i=0}^{m} \frac{(m)!}{(m-i)!*i!} = \ 2^{m}$.

Maximum number of glycan allowed per peptide	Total Combinations	Number of unique mass	Average number of glycan groups with same mass (*r*)
2	90	60	1.5
3	454	184	2.5
4	1819	439	4.1
5	6187	895	6.9
6	18563	1637	11.3
7	50387	2765	18.2
8	125969	4394	28.7

Supplementary Table 2. Average number of glycan groups of the 12 glycan database.

Computational complexity analysis

We observed a great reduction in search time for O-Pair Search compared against the commercial software program Byonic for similarly configured O-glycopeptide searches. This prompted us to perform an analysis of the computational complexity of the O-Pair Search algorithm and other related methods to determine the source of this time savings.

To simplify the analysis, we treat each theoretical-experimental spectrum comparison as one-unit of computation to calculate the associated computational complexity.

1) Traditional database search (narrow)

The traditional database search for O-glycopeptide identification requires the construction of a complete O-glycopeptide database prior to the matching process. The O-glycopeptide database is built using O-glycans as variable modifications. In the variable modification scenario, each peptide yields a combinatorial number of glycopeptides based the number of modifiable sites within the peptide and the number of different glycans that can occupy those sites1. The computational complexity (Supplementary Table 1) of traditional glycopeptide search is:

$$\mathrm{O}\left(n * \sum_{i=0}^{m} \binom{s}{i} g^{i} \right)$$

2) Ion-Indexed Open Search with localization

The O-Pair Search algorithm, described in detail in the Methods Section, consists of an ion-indexed open search to ascertain the amino acid sequence of the peptide backbone and a graph-theoretical approach to determining the positions of glycans adorning the peptide. We analyze below the theoretical complexity of each stage separately.

2.1) Ion Indexing

Ion-indexed open search strategies have been used for bottom-up peptide identification and post-translational modification discovery[2]. Ion-indexed open searches are particularly advantageous for the latter because there is no need to select theoretical peptides based on the experimental precursor mass, which would have required construction of all theoretically modified forms. Ion indexes are generally constructed from theoretical fragments of the unmodified peptide backbone for each peptide in the database. In the case of O-pair Search, the use of an ion-indexed open search for the first stage means that there is no need to consider all the possible different glycopeptide forms that could be constructed for each peptide. A theoretical peptide database containing all possible glycans would be massive. Peptide candidates only are identified in the ion-indexed open search during the first stage. Possible glycans are considered in the next. The ion indexing process in MetaMorpheus has been heavily optimized for high speed. The computational complexity (Supplementary Table 1) for ion indexing is related to the number of peaks in the MS2 spectrum, which can be considered as a small constant:

$$\mathrm{O}\ (x)$$

2.2) Brute force localization

MetaMorpheus O-pair Search does not perform a brute force localization. However, we provide here an analysis of the brute force localization approach as a point of reference for the reader to understand it in relation to the computational complexity of the graph-based localization described immediately after. Glycan group candidates are determined after obtaining peptide candidates from the ion-indexed open search. The brute force approach to localization consists of matching each glycan group to each different peptide candidate (combinatorial), including all possible localization isomers. This process yields $\frac{s!}{(s-m)!}$ glycopeptides for consideration. The overall computational complexity (Supplementary Table 1) for ion-indexed open search with brute force localization is:

$$\mathrm{O}\ (x + k * r * \frac{s!}{(s-m)!})$$

2.3) Graph-based Localization

The graph-based localization was described in detail in the methods section. Here we provide additional details to compute the computational complexity. There are two sets of theoretical fragment ions associated with glycopeptides in a graph. One set of theoretical fragments are shared between multiple glycopeptides and the other set of fragments is unique to a single glycopeptide. The shared fragments are predominant. The graph-based localization saves time in constructing all possible glycopeptide forms and simultaneously enables the matching of fragments without redundancy, taking advantage of the knowledge of shared fragments.

The computational complexity is proportional to the depth $d = 2^m$ of the graph, which is a function of the number of possible glycan modifications m. For O-glycopeptides in general, m is not a large number. The overall computational complexity (Supplementary Table 1) for ion-indexed open search with graph-based localization is:

O ($x + k * r * d$)

3) Comparison

To compare the computational complexity, we ignore the values n, x and k and focus on analyzing how the numbers of glycan types and glycosites influence the computational complexity.

For example, using g = 12 glycans, s = 8 glycosites and m = 2 glycans (as in Fig 1b), the computational complexity (Supplementary Table 1) is as follows:

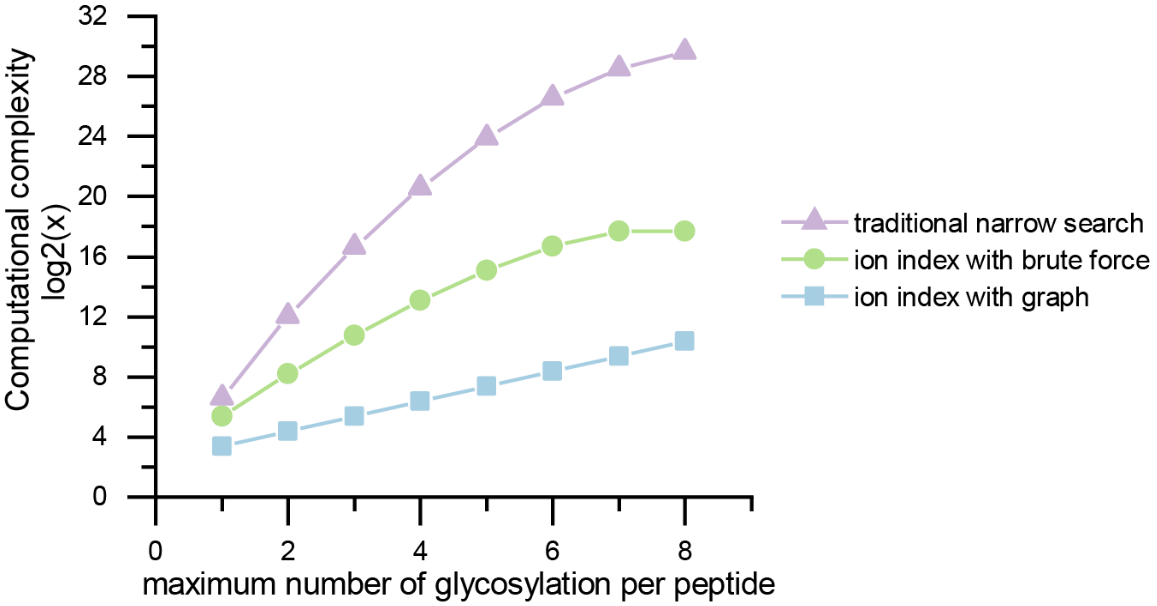
in traditional search $\sum_{i=0}^{m} \frac{(s)!}{(s-i)!*i!} * g^i = 4291$

in ion-indexed open search with brute force localization $r * \frac{s!}{(s-m)!} = 1.5 * 56 = 84$

in ion-indexed open search with graph-based localization $r * m^2 = 1.5 * 4 = 6$

Next, we look at how these numbers evolve as a function of m. We set g = 12 and s = 8 and find that the computational complexity changes dramatically for traditional narrow search and brute force methods (Supplementary Figure 14, Supplementary Table 2).

Using the ion-indexed open search and graph-based localization, O-Pair Search avoids the ‘combinatorial explosion’ that happens in for traditional searches of O-glycopeptide characterization. O-Pair Search in this way could use a large sized glycan database and expands the potential for O-glycopeptide characterization.



Supplementary Figure 12. Comparing the computational complexity in traditional narrow search, ion-indexed open search with brute force localization and graph-based localization. Note, the y-axis is on a log2 scale.

m	graph	brute force	traditional
1	8	32	97
2	16	224	4129
3	32	1344	100897
4	64	6720	1552417
5	128	26880	1.55E+07
6	256	80640	9.91E+07

7	512	161280	3.86E+08
8	1024	161280	8.16E+08

Supplementary Table 3. Comparing the computational complexity in traditional narrow search, ion-indexed open search with brute force localization and graph-based localization. The number of glycan types is set to g = 12, the number of glycosites is set to s = 8, the number of glycan groups with the same mass r is about 4, and m is the maximum number of glycosylation allowed on a single peptide.

SUPPLEMENTARY NOTE 2

During the review process for this manuscript, another approach using fragment ion indexed open searching of glycopeptide spectra was reported from Nesvizhskii and co-workers, the group who developed the highly efficient ion-indexed strategy.[5] The MSFragger-Glyco manuscript re-analyzes data from Zhang et al., which used the O-glycoprotease OpeRATOR to generate O-glycopeptides and exclusively relies on HCD fragmentation.[6] Without electron-based dissociation, localization of the glycosite via spectral evidence is largely impossible, which is one of the reasons we elected not to analyze this dataset. Regardless, MSFragger-Glyco reported impressive boosts in identifications over the published dataset using a database of 110 glycan compositions. MSFragger-Glyco also reduced search times, as O-Pair Search does.

An important distinction between the two algorithms is the incorporation of O-glycosite localization with O-Pair Search versus no such calculation with MSFragger-Glyco. The most obvious importance here is the ability of O-Pair Search data to provide site-specific glycosylation, arguably the most biologically relevant information. But even more so, assigning glycan structures to specific sites is much more computationally challenging than only assigning a total glycan aggregate mass to a peptide sequence. This was first addressed in the main script with discussion of the O-Search algorithm from Mao et al,[3] where static glycan aggregate mass combinations are searched with no regard for O-glycosite localization. To be sure, MSFragger-Glyco improves the speed at which this type of search can be done, but it does not extend the concept beyond identification of total glycan mass plus peptide sequence with no glycosite localization. This sacrifices site assignment in addition to the number and composition of individual glycans. This type of search is possible when using O-Pair Search for data that relies exclusively collisional dissociation (**Supplemental Figure 5**), but it falls short of the goals that inspired the development of O-Pair Search in the first place. Instead, O-Pair Search provides the ability to not only identify a glycosylated sequence but also offers site assignment that can identify individual glycans that comprise a larger total glycan aggregate mass.

Thus, the number of glycan compositions considered by O-Pair Search is not merely the size of the O-glycan database used for the search (12 glycans in default search), but is also the combination of all the possible glycan compositions derived from that database. For example, an O-Pair Search that uses 12 glycans with a maximum of 4 glycans allowed per peptide actually evaluates 439 single mass offsets, i.e., the number of unique masses present in 1819 different glycan combinations. See **Supplemental Note 1** above for discussion of how these combinations are generated. Considering this, it is clear that O-Pair Search handles glycan composition lists just as big, if not significantly larger, than what was reported by MSFragger-Glyco,[5] and it maintains the rapid performance expected of ion-indexed searches. Furthermore, a key component to this performance is the graph theoretical localization performed by O-Pair Search. Brute force localization approaches come with a significant time cost (Supplemental Note 1), so the incorporation of graph theory-based O-glycosite localization into the total processing time reported represents a distinct advantage for O-Pair Search.

Finally, an important aspect of the MSFragger-Glyco work is extension of glyco-searches to include other PTMs, i.e., phosphorylation crosstalk analysis. The crosstalk examples from MSFragger-Glyco will likely inspire future work of glycosylation PTM crosstalk, which is commendable in and of itself.

SUPPLEMENTARY NOTE 3

Several entrapment databases were constructed to evaluate O-Pair Search performance, both for the time it takes to complete searches with progressively larger databases and for false discovery rate calculations. The goal was to create background proteomes that not only varied in size and complexity, but also established different degrees of relevance to the four mucin standard proteins that were actually present in the sample: CD43, PSGL-1, MUC16, and Gp1ba.

First, we appended the four true positive mucins to 20 canonical human mucin sequences[7]. Here, we create a “high mucin” background, where the false positives are very similar to the target proteins in sequence composition and proclivity for O-glycosylation. This also represents a close approximation of real-world scenarios where digested mucin O-glycopeptides would be screened from pools of several mucins with the need to accurately identify which glycopeptides match to which proteins. This entrapment database also represents an algorithmic challenge, where there are a multitude of O-glycopeptide candidates generated with a high number of O-glycosites.

The second entrapment database comprised the four true mucin standard sequences combined with common fetal bovine serum (FBS) proteins[8], many of which are glycosylated. This database represents a “high glycoprotein, mammalian but not human” background, where the majority of glycoproteins in the database have the capacity to be glycosylated in actuality but there should be virtually no peptides mapping to bovine sequences considering the human mucin standard protein source. This also represents a situation many researchers may encounter where FBS proteins are present due to preparation conditions and may need to be accounted for. Computationally, this is moderately more entries than the mucin background above, but with fewer overall mucin-type proteins to produce large quantities of high S/T density glycopeptide candidates.

Human cluster of differentiation (CD) markers, which are known cell surface proteins, were downloaded from the Human Protein Atlas[9] and used for the third entrapment database. This set of proteins represents a “high glycoprotein, human” background, as the majority of CD markers are known to be glycosylated. Several CD markers are classified as mucins, so this database challenges the accuracy of assignment using proteins that could potentially have real O-glycopeptide assignments while also increasing the number of entries ~5-10 fold over the first two databases.

In principle, proteins from these first three entrapment databases could have the type of O-glycosylation also seen on the true positive mucin standards. Choosing entrapment databases of the entire proteomes of E. coli and S. cerevisiae removes this constraint, as neither species should contain the glycans in the O-glycan database or peptides sequences that map to mammalian proteins. They represent “non-glycoprotein, non-species related” backgrounds that should have little in common with the true positive mucin standard proteins. They also represent significant increases in database size to ~900 and ~6700 protein entries, respectively. Thus, these two databases serve to challenge the accuracy of O-Pair Search and the search speed when considering proteome-scale data.

Finally, we searched a subset of the CHO proteome[2] and the entire mouse proteome as these are both expression systems for the recombinant mucins in this study. These data further support the need for appropriately sized databases for glycoproteomics searches[10]. While there were some peptides from contaminating proteins from the expression systems, they were in the significant minority. Thus, inclusion of unnecessarily large background proteomes can hurt sensitivity, as is seen with the decrease in identifications in **Supplementary Fig. 16**. That said, the majority of identification losses were in Level 2 and 3 identifications, which are less confident. Level 1 and

1b identifications showed significantly fewer losses relative to the other Localization Levels, indicating that high confidence identifications are the least affected by database size (as was seen with the entrapment database searches).

Together, these experiments tested the speed and accuracy of O-Pair Search, both when the background looks similar to the known mucin standards and when it looks entirely different to the mucins of interest.

SUPPLEMENTARY NOTE 4

Glycan databases, mucin standard protein sequences, and raw files used for different data throughout the paper are provided here.

1. The 12 O-glycan database:

Glycan	Mass (Da)
(N)	203.0794
(N(H))	365.1322
(N(A))	494.1748
(N(H)(N))	568.2116
(N(H(A)))	656.2276
(N(H)(N(H)))	730.2644
(N(H(A))(N))	859.307
(N(H(A))(A))	947.323
(N(H(A))(N(H)))	1021.3598
(N(H)(N(H(A))(F)))	1167.4177
(N(H(A))(N(H(A))))	1312.4552
(N(H(A))(N(H(A))(F)))	1458.5131

* N: HexNAc, H: Hexose, A: NeuAc, F: Fucose.

2. The 28 O-glycan database:

Glycan	Mass
HexNAc(1)	203.0794
HexNAc(1)Hex(1)	365.1322
HexNAc(2)	406.1588
HexNAc(1)NeuAc(1)	494.1748
HexNAc(1)Hex(1)Fuc(1)	511.1901
HexNAc(1)Hex(2)	527.185
HexNAc(2)Hex(1)	568.2116
HexNAc(1)Hex(1)NeuAc(1)	656.2276

HexNAc(1)Hex(2)Fuc(1)	673.2429
HexNAc(2)Hex(1)Fuc(1)	714.2695
HexNAc(2)Hex(2)	730.2644
HexNAc(1)Hex(1)Fuc(1)NeuAc(1)	802.2855
HexNAc(1)Hex(2)NeuAc(1)	818.2804
HexNAc(2)Hex(1)NeuAc(1)	859.307
HexNAc(2)Hex(1)Fuc(2)	860.3274
HexNAc(2)Hex(2)Fuc(1)	876.3223
HexNAc(1)Hex(1)NeuAc(2)	947.323
HexNAc(1)Hex(2)Fuc(1)NeuAc(1)	964.3383
HexNAc(2)Hex(1)Fuc(1)NeuAc(1)	1005.365
HexNac(2)Hex(2)NeuAc(1)	1021.36
HexNAc(2)Hex(2)Fuc(2)	1022.38
HexNAc(1)Hex(2)NeuAc(2)	1109.376
HexNAc(2)Hex(2)Fuc(1)NeuAc(1)	1167.418
HexNAc(1)Hex(1)NeuAc(3)	1238.418
HexNac(2)Hex(2)NeuAc(2)	1312.455
HexNAc(2)Hex(2)Fuc(2)NeuAc(1)	1313.476
HexNAc(2)Hex(1)NeuAc(3)	1441.498
HexNAc(2)Hex(2)Fuc(1)NeuAc(2)	1458.513

3. The 32 O-glycan database:

Glycan	Mass (Da)
HexNAc(1)	203.0794
HexNAc(1)Hex(1)	365.1322
HexNAc(2)	406.1588
HexNAc(1)NeuAc(1)	494.1748
HexNAc(1)Hex(1)Fuc(1)	511.1901
HexNAc(1)Hex(2)	527.185
HexNAc(2)Hex(1)	568.2116

HexNAc(1)Hex(1)NeuAc(1)	656.2276
HexNAc(1)Hex(2)Fuc(1)	673.2429
HexNAc(2)Hex(1)Fuc(1)	714.2695
HexNAc(2)Hex(2)	730.2644
HexNAc(1)Hex(1)Fuc(1)NeuAc(1)	802.2855
HexNAc(1)Hex(2)NeuAc(1)	818.2804
HexNAc(2)Hex(1)NeuAc(1)	859.307
HexNAc(2)Hex(1)Fuc(2)	860.3274
HexNAc(2)Hex(2)Fuc(1)	876.3223
HexNAc(1)Hex(1)NeuAc(2)	947.323
HexNAc(1)Hex(2)Fuc(1)NeuAc(1)	964.3383
HexNAc(2)Hex(1)Fuc(1)NeuAc(1)	1005.3649
HexNAc(2)Hex(2)NeuAc(1)	1021.3598
HexNAc(2)Hex(2)Fuc(2)	1022.3802
HexNAc(3)Hex(3)	1095.3966
HexNAc(1)Hex(2)NeuAc(2)	1109.3758
HexNAc(2)Hex(2)Fuc(1)NeuAc(1)	1167.4177
HexNAc(1)Hex(1)NeuAc(3)	1238.4184
HexNAc(3)Hex(3)Fuc(1)	1241.4545
HexNAc(2)Hex(2)NeuAc(2)	1312.4552
HexNAc(2)Hex(2)Fuc(2)NeuAc(1)	1313.4756
HexNAc(3)Hex(3)NeuAc(1)	1386.492
HexNAc(3)Hex(3)Fuc(2)	1387.5124
HexNAc(2)Hex(1)NeuAc(3)	1441.4978
HexNAc(2)Hex(2)Fuc(1)NeuAc(2)	1458.5131

4. The four mucin standard sequences:
>sp|P07359|GP1BA_HUMAN Platelet glycoprotein Ib alpha chain OS=Homo sapiens OX=9606 GN=GP1BA PE=1 SV=2

HPICEVSKVASHLEVNCDKRNLTALPPDLPKDTTILHLSENLLYTFSLATLMPYTRLTQL
NLDRCELTKLQVDGTLPVLGTLDLSHNQLQSLPLLGQTLPALTVLDVSFNRLTSLPLGAL
RGLGELQELYLKGNELKTLPPGLLTPTPKLEKLSLANNNLTELPAGLLNGLENLDTLLLQ
ENSLYTIPKGFFGSHLLPFAFLHGNPWLCNCEILYFRRWLQDNAENVYVWKQGVDVKA
MTSNVASVQCDNSDKFPVYKYPGKGCPTLGDEGDTDLYDYYPEEDTEGDKVRATRTV
VKFPTKAHTTPWGLFYSWSTASLDSQMPSSLHPTQESTKEQTTFPPRWTPNFTLHMESIT
FSKTPKSTTEPTPSPTTSEPVPEPAPNMTTLEPTPSPTTPEPTSEPAPSPTTPEPTSEPAPSPT
TPEPTSEPAPSPTTPEPTPIPTIATSPTILVSATSLITPKSTFLTTTKPVSLLESTKKTIPELDQP
PKLRGVLQGHLESSRNDPFLHPDFCCLLPLGFYVLGLFWLLFASVVLILLLSWVGHVKP
QALDSGQGAALTTATQTTHLELQRGRQVTVPRAWLLFLRGSLPTFRSSLFLWVRPNGR
VGPLVAGRRPSALSQGRGQDLLSTVSIRYSGHSL

>sp|P16150|LEUK_HUMAN Leukosialin OS=Homo sapiens GN=SPN PE=1 SV=1
STTAVQTPTSGEPLVSTSEPLSSKMYTTSITSDPKADSTGDQTSALPPSTSINEGSPLWTSI
GASTGSPLPEPTTYQEVSIKMSSVPQETPHATSHPAVPITANSLGSHTVTGGTITTNSPETS
SRTSGAPVTTAASSLETSRGTSGPPLTMATVSLETSKGTSGPPVTMATDSLETSTGTTGPP
VTMTTGSLEPSSGASGPQVSSVKLSTMMSPTTSTNASTVPFRNPDENSRGMLPVAVLVA
LLAVIVLVALLLLWRRRQKRRTGALVLSRGGKRNGVVDAWAGPAQVPEEGAVTVTVG
GSGGDKGSGFPDGEGSSRRPTLTTFFGRRKSRQGSLAMEELKSGSGPSLKGEEEPLVASE
DGAVDAPAPDEPEGGDGAAP

>sp|Q8WXI7.3|MUC16_HUMAN Mucin-16 segment OS=Homo sapiens OX=9606 GN=MUC16 PE=1 SV=3
MPLFKNTSVSSLYSGCRLTLLRPEKDGAATRVDAVCTHRPDPKSPGLDRERLYWKLSQL
THGITELGPYTLDRHSLYVNGFTHQSSMTTTRTPDTSTMHLATSRTPASLSGPTTASPLL
VLFTINFTITNLRYEENMHHPGSRKFNTTERVLQGLLRPVFKNTSVGPLYSGCRLTLLRP
KKDGAATKVDAICTYRPDPKSPGLDREQLYWELSQLTHSITELGPYTLDRDSLYVNGFT
QRSSVPTTSIPGTPTVDLGTSGTPVSKPGPSAASPLLVLFTLNFTITNLRYEENMQHPGSR
KFNTTERVLQGLLRSLFKSTSVGPLYSGCRLTLLRPEKDGTATGVDAICTHHPDPKSPRL
DREQLYWELSQLTHNITELGPYALDNDSLFVNGFTHRSSVSTTSTPGTPTVYLGASKTPA
SIFGPSAASHLLILFTLNFTITNLRYEENMWPGSRKFNTTERVLQGLLRPLFKNTSVGPLY
SGCRLTLLRPEKDGEATGVDAICTHRPDPTGPGLDREQLYLELSQLTHSITELGPYTLDR
DSLYVNGFTHRSSVPTTSTGVVSEEPFTLNFTINNLRYMADMGQPGSLKFNITDNVMKH
LLSPLFQRSSLGARYTGCRVIALRSVKNGAETRVDLLCTYLQPLSGPGLPIKQVFHELSQ
QTHGITRLGPYSLDKDSLYLNGYNEPGPDEPPTTPKPATTFLPPLSEATTAMGYHLKTLT
LNFTISNLQYSPDMGKGSATFNSTEGVLQHLLRPLFQKSSMGPFYLGCQLISLRPEKDGA
ATGVDTTCTYHPDPVGPGLDIQQLYWELSQLTHGVTQLGFYVLDRDSLFINGYAPQNLS
IRGEYQINFHIVNWNLSNPDPTSSEYITLLRDIQDKVTTLYKGSQLHDTFRFCLVTNLTMD
SVLVTVKALFSSNLDPSLVEQVFLDKTLNASFHWLGSTYQLVDIHVTEMESSVYQPTSSS
STQHFYLNFTITNLPYSQDKAQPGTTNYQHHHHHH

>sp|Q14242|SELPL_HUMAN P-selectin glycoprotein ligand 1 OS=Homo sapiens GN=SELPLG PE=1 SV=1
QLWDTWADEAEKALGPLLARDRRQATEYEYLDYDFLPETEPPEMLRNSTDTTPLTGPG
TPESTTVEPAARRSTGLDAGGAVTELTTELANMGNLSTDSAAMEIQTTQPAATEAQTTQ
PVPTEAQTTPLAATEAQTTRLTATEAQTTPLAATEAQTTPPAATEAQTTQPTGLEAQTTA

PAAMEAQTTAPAAMEAQTTPPAAMEAQTTQTTAMEAQTTAPEATEAQTTQPTATEAQT
TPLAAMEALSTEPSATEALSMEPTTKRGLFIPFSVSSVTHKGIPMAASNLSVNYPVGAPD
HISVK

5. Data files used for this study

A.	**Data file used for entrapment searches**	
	2019_09_16_StcEmix_35trig_EThcD25_rep1.raw	PXD017646
B.	**Data file used for different glycan numbers and glycan databases**	
	2019_09_16_StcEmix_35trig_EThcD25_rep1.raw	PXD017646
C.	**Data files used for fragmentation analysis and other searches**	
	2019_09_16_StcEmix_35trig_ETD_rep1.raw 2019_09_16_StcEmix_35trig_ETD_rep2.raw 2019_09_16_StcEmix_35trig_ETD_rep3.raw 2019_09_16_StcEmix_35trig_EThcD15_rep1.raw 2019_09_16_StcEmix_35trig_EthcD15_rep2.raw 2019_09_16_StcEmix_35trig_EthcD15_rep3.raw 2019_09_16_StcEmix_35trig_EthcD25_rep1.raw 2019_09_16_StcEmix_35trig_EthcD25_rep2.raw 2019_09_16_StcEmix_35trig_EthcD25_rep3.raw 2019_09_16_StcEmix_35trig_EthcD35_rep1.raw 2019_09_16_StcEmix_35trig_EthcD35_rep2.raw 2019_09_16_StcEmix_35trig_EthcD35_rep3.raw	PXD017646
D.	**Data files used for HCD/stepped HCD data analysis**	PXD017646
	2019_09_16_StcEmix_35trig_HCD25.raw 2019_09_16_StcEmix_35trig_HCD25_rep2.raw 2019_09_16_StcEmix_35trig_HCD25_rep3.raw 2019_09_16_StcEmix_35trig_HCD30.raw 2019_09_16_StcEmix_35trig_HCD30_rep2.raw 2019_09_16_StcEmix_35trig_HCD30_rep3.raw 2019_09_16_StcEmix_35trig_HCD35.raw 2019_09_16_StcEmix_35trig_HCD35_rep2.raw 2019_09_16_StcEmix_35trig_HCD35_rep3.raw	

	2019_09_16_StcEmix_35trig_HCD40.raw 2019_09_16_StcEmix_35trig_HCD40_rep2.raw 2019_09_16_StcEmix_35trig_HCD40_rep3.raw 2019_09_16_StcEmix_35trig_sceHCD30step10.raw 2019_09_16_StcEmix_35trig_sceHCD30step10_rep2.raw 2019_09_16_StcEmix_35trig_sceHCD30step10_rep3.raw 2019_09_16_StcEmix_35trig_sceHCD30step18.raw 2019_09_16_StcEmix_35trig_sceHCD30step18_rep2.raw 2019_09_16_StcEmix_35trig_sceHCD30step18_rep3.raw 2019_09_16_StcEmix_35trig_sceHCD30step15.raw 2019_09_16_StcEmix_35trig_sceHCD30step15_rep2.raw 2019_09_16_StcEmix_35trig_sceHCD30step15_rep3.raw 2019_09_16_StcEmix_35trig_sceHCD30step5.raw 2019_09_16_StcEmix_35trig_sceHCD30step5_rep2.raw 2019_09_16_StcEmix_35trig_sceHCD30step5_rep3.raw	
E.	**Data files used for urinary O-glycopeptides fraction 1 searches**	
	170919_11.raw 170921_06.raw 170922_04.raw	MSV000083070
F.	**Data files used for urinary O-glycopeptides fraction 2 searches**	
	170919_08.raw 170922_01.raw	MSV000083070
G.	**other urinary data files used**	
	171025_06.raw 171027_06.raw 171027_05.raw 180417_07.raw 180417_05.raw	MSV000083070

REFERENCES

(1) Riley, N. M.; Malaker, S. A.; Driessen, M.; Bertozzi, C. R. Optimal Dissociation Methods Differ for N- and O-Glycopeptides. *J. Proteome Res.* **2020**. https://doi.org/10.1021/acs.jproteome.0c00218.

(2) Park, J. H.; Jin, J. H.; Lim, M. S.; An, H. J.; Kim, J. W.; Lee, G. M. Proteomic Analysis of Host Cell Protein Dynamics in the Culture Supernatants of Antibody-Producing CHO Cells. *Sci. Rep.* **2017**, *7* (1), 1–13. https://doi.org/10.1038/srep44246.

(3) Mao, J.; You, X.; Qin, H.; Wang, C.; Wang, L.; Ye, M. A New Searching Strategy for the Identification of O-Linked Glycopeptides. *Anal. Chem.* **2019**, *91* (6), 3852–3859. https://doi.org/10.1021/acs.analchem.8b04184.

(4) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.

(5) Polasky, D. A.; Yu, F.; Teo, G. C.; Nesvizhskii, A. I. Fast and Comprehensive N- and O-Glycoproteomics Analysis with MSFragger-Glyco. *bioRxiv* **2020**, 2020.05.18.102665. https://doi.org/10.1101/2020.05.18.102665.

(6) Yang, W.; Ao, M.; Hu, Y.; Li, Q. K.; Zhang, H. Mapping the O-glycoproteome Using Site-specific Extraction of O-linked Glycopeptides (EXoO). *Mol. Syst. Biol.* **2018**, *14* (11). https://doi.org/10.15252/msb.20188486.

(7) Lang, T.; Klasson, S.; Larsson, E.; Johansson, M. E. V.; Hansson, G. C.; Samuelsson, T. Searching the Evolutionary Origin of Epithelial Mucus Protein Components - Mucins and FCGBP. *Mol. Biol. Evol.* **2016**, *33* (8), 1921–1936. https://doi.org/10.1093/molbev/msw066.

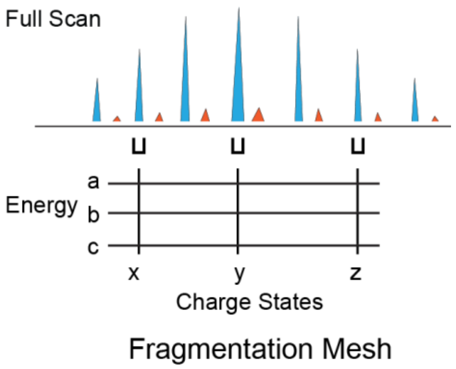
(8) Shin, J.; Kim, G.; Kabir, M. H.; Park, S. J.; Lee, S. T.; Lee, C. Use of Composite Protein Database Including Search Result Sequences for Mass Spectrometric Analysis of Cell Secretome. *PLoS One* **2015**, *10* (3), e0121692. https://doi.org/10.1371/journal.pone.0121692.

(9) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; et al. Tissue-Based Map of the Human Proteome. *Science (80-.).* **2015**, *347* (6220), 1260419–1260419. https://doi.org/10.1126/science.1260419.

(10) Khatri, K.; Klein, J. A.; Zaia, J. Use of an Informed Search Space Maximizes Confidence of Site-Specific Assignment of Glycoprotein Glycosylation. *Anal. Bioanal. Chem.* **2017**, *409* (2), 607–618. https://doi.org/10.1007/s00216-016-9970-5.

Chapter 4

Fragmentation Mesh improves fragmentation efficiency for Top-down proteomics



Adapted from Lei Lu, Mark Scalf, Michael R. Shortreed, Lloyd M. Smith* (2020). Fragmentation Mesh improves fragmentation efficiency for Top-down proteomics. (to be submitted)

ABSTRACT

Top-down proteomics is a major mass spectrometry technology for comprehensive analysis of proteoforms. Different from peptides, proteoforms appear in multiple charge states and isotopic forms in full MS scans. The current data dependent acquisition (DDA) method randomly selects from amongst the abundant precursors (one charge state of a proteoform) for fragmentation. Fragmentation efficiency could differs substantially between different charge states and NCEs (normalized collision energies). In this study we proved that stepped HCD (high collision energy dissociation) and a 'Mesh' strategy could improve fragmentation efficiency and improve identification rate. Mesh fragmentation reduces the possibility of getting low quality fragmentation from single charge state. The new strategy can perform real time deconvolution to keep track of all proteoform charge states. Fragmentation multiple charge states with multiple NCEs is ferformed within an the open source instrument control software program called MetaDrive.

INTRODUCTION

The interest in top-down proteomics has dramatically increased in the recent decade for targeted protein analysis or protein post-translational modification analysis.[1,2] As a useful technology to analyze proteoforms without digestion, top-down proteomics subjects intact proteins to fragmenation and retains protein level information.[3,4]

The technologies for top-down proteomics improved greatly in recent years. Thousands of proteoforms can be identified in large scale studies.[5–7] Limitations still exist for the current fragmentation methods used for top-down proteomics. In top-down proteomics, ESI (electrospray ionization) generates proteoforms with multiple high charge states across the m/z range of a typical full MS spectrum. The data dependent acquisition (DDA) method mainly used for current top-down proteomics is directly adopted from bottom-up proteomics with minor changes of the instrument parameters, ignoring the multiple charge states of proteoforms presented.[1] DDA is generally coupled with HCD (higher energy collisional dissociation), which is one of the most commonly used fragmenation methods in top-down proteomics.[1,6,7]

Previous studies showed that the fragmentation efficiencies are different for the different charge states of one proteoform under the same NCE, or for the same charge states of one proteoform under different NCE.[8,9] DDA randomly select one abundant percursor for each MS/MS scan. Thus it is by luck if one happens to get suitable fragmentation conditions sufficient for proteoform identification. Proteoforms can range from a few kDa to tens of kDa and differnt substantially in composition (e.g. hydrophicity or ionizability). This makes it challenging to select optimal fragmentation conditions that suitably cover this diversity of proteoforms. There is an opportunity to improve the current process of charge state selection and choice of NCE for top-down proteomics.

Advanced precursor ion selection algorithms have been applied for bottom-up and top-down proteomics.[10,11] In 2014, Durbin et al. developed Autopilot to improve data acquistion for top-down proteomics.[11] Autopilot performs intelligent data collection with online mass dectection, real-time searching to guide precursor selection and fragmentation. Their results showed an improvement to the overall fragmentation coverage of many proteins. The authors demonstrated that the intelligent data acquisition could increase unique protein identifications when coupled with advanced instruments. In Autopilot, charge state selection and NCE sequentially adjusted whenever fragmentation yields unconfident identification.

In this study, we demontrate that stepped HCD could greatly improve fragmentation efficiency, with each collision energy provides differential fragmentation of correspondingly labile peptide bonds algong protein sequence.[12,13] We also developed a new fragmentation method called Fragmentation Mesh, which combines several charge states of one proteoform for fragmentation under multiple NCEs. The new method improves fragmentation efficiency and the identification rate for top-down proteomics.

METHODS

Sample Preparation

Ubiquitin from bovine (UniProt Accession P0CG53), Cytochrome C from horse (UniProt Accession L7MRG1), and myoglobin from horse (UniProt Accession P68082) were purchased as standards from Sigma. All samples were resuspended at ∼10 pmol/μL in 49.9:49.9:0.2 acetonitrile/water/formic acid prior to infusions.

Yeast cells were grown to a density of ∼10^6 cells/mL at which time they were washed, pelleted, snap-frozen in liquid nitrogen, and stored at −80 °C until use. Yeast cells were lysed separately, and proteins were reduced and alkylated. Proteins were then precipitated with acetone before being resuspended in separation buffer. The proteins were separated based on molecular weight (MW) using a GELFrEE system (Expedeon),[5] and approximately 400 μg of protein were collected into 11 fractions. Two of the fractions were selected for the top-down analysis. Prior to mass spectrometric analysis, sodium dodecyl sulfate was removed from the fractions via methanol–chloroform precipitation and proteins were reconstituted with 5% acetonitrile (ACN) and 0.2% formic acid in water.

Mass Spectrometry

Standard proteins were infused directly by electrospray into the mass spectrometer for top-down analysis. Full-mass profile scans are performed in the Orbitrap between 375 and 1,500 *m/z* at a resolution of 120,000, followed by MS/MS HCD scans at a resolution of 60,000 and a mass range of 400-2000 *m/z*, where the parent ion selection was controlled by MetaDrive. Four different types of MetaDrive controlled fragmentation schemes were performed separately, including: Top method (a single charge state and a single NCE), Line method (multiple charge states and a single NCE), stepped HCD method (a single charge state and multiple NCE) and Mesh method (multiple charge states and multiple NCEs).

Top-down proteomics analysis of yeast samples (~2 μg protein each injection) were performed using HPLC (NanoAcquity, Waters)-ESI-MS/MS (Q Exactive HF, ThermoFisher Scientific). Four different types of MetaDrive controlled fragmentation schemes were performed separately as described before. For all fragmentation schemes, the full-mass profile scans were performed in the Orbitrap between 375 and 1,500 *m/z* at a resolution of 120,000, followed by MS/MS HCD scans of the top two highest intensity parent ions and 60,000 resolution, with a mass range of 400-2000 *m/z*. Dynamic exclusion was enabled with an exclusion window of 30 s.

Instrument control software program

We developed a new software program, MetaDrive, that performs real-time isotope and deconvolution followed by new methods for precursor selection and fragmentation. The deconvolution algorithm is adopted from MS-Deconv.[14] MetaDrive controls an Orbitrap Q-Exactive through Thermo IAPI (Instrument Application Programming Interface) software.[15] We use MetaDrive to speed up the optimization of protein fragmentation parameters. MetaDrive combines multiple charge states of one proteomform for fragmentation with multiple collisional energies, yielding better sequence coverage and improved identification rate. MetaDrive is written in C# and is publicly available to the community as open-source code at https://github.com/smith-chem-wisc/MetaDrive.

Data analysis

The pTop2 software program (http://pfind.ict.ac.cn/software.html) was used to perform the top-down analysis of the raw files.[16] This software reported the number of matched fragments generated for each spectrum. Most of the default parameter settings were used, except that the max charge was changed to 50, mixture spectra was not checked, and variable modifications including Oxidation[M], dehydro[C] and Acetyl[ProteinN-term] were added. Detailed parameter settings used in pTop2 are described in Supplementary Table 1. A FASTA file of the standard proteins and yeast (2019.09) from UniProt database were used.

RESULTS AND DISCUSSION

Charge state and NCE

Proteoforms appear in multiple charge states in top-down full MS spectra. The current DDA scheme randomly select one charge state each time, not gurrentteeing an optimal fragmentation for the selected proteoform.[1,11] For the specific charge state of one proteoform, an alternate NCE could also lead to different fragmentation efficiency.[8,9]

We first evaluated the fragmentation efficiency for the same proteoform with different charge states and different NCEs. Three standard proteins including ubiquitin, cytochrome C and myoglobin, which generates different charge states distributions, were used for the experiment. (Fig. 1a) To simplify the evaluation, we only considered the number of matched b and y-ions from each MS/MS spectrum identified by pTop, which agrees with the scheme of commonly used software programs.[16,17]

Different charge states of the same proteoform yeiled significantly different number of matched fragment ions. (Fig. 1b) For each proteoform, our results showed that the charge state with the highest intensity didn't always yeild the highest number of matched fragment ions, an observation that agree with previous studies.[9,18] Note, this phenomenon could be unique to HCD. Further evaluation with alternate modes of fragmentation is needed. The best fragmentation for each of the protein is from a low charge state, but not necessarily the lowest charge state for the protein, depending on the NCE. While in the current DDA scheme, the most abundant peaks are more likely to be selected for fragmentation, which is not optimal for fragmentation.

A NCE found to work well for one proteoform frequently does not works well for all others. In our stduy, we showed that NCE 20 could generate more matched fragment ions overall. Still NCE 20 doesn't always generate the best, for example for myoglobin charge state 21. (Fig. 1b)

Fragmentation Mesh

It is inefficient to iterate through several charge states of a proteoform and NCEs for each proteoform to obtain good fragmentation of complex sample. We hypothesized that it might be possible to improve sequence coverage by fragmenting several different charge states together while also emplying multiple NCEs. We implemented a corresponding new data acquistion method in MetaDrive (Fig. 2), which is a software program that can control the QE-HF via IAPI. Each time MetaDrive receive a full MS scan from the instrument, it will perform a real-time deconvolution. MetaDrive could perform four different type of fragmentation schemes based on the deconvolution results, including: Top method (a single charge state and a single NCE), Line method (multiple charge states and a single NCE), Stepped HCD method (a single charge state and multiple NCEs) and Mesh method (multiple charge states and multiple NCEs). Compared with the traditional fragmentation scheme (the Top method), the other three methods increased the probability of good fragmentation.

We compared the fragmentation methods using the three standard proteins. (Fig. 3) The charge state with the highest intensity is selected for each proteoform in the Point method. For the Line and Fragmentation Mesh methods, we enforce MetaDrive to select charge states around the highest intensity one and use NCE 25, which was commonly used in previous studies.[6,19] For the Stepted HCD and Fragmentation Mesh methods, we applied NCE 15-25-35. For all three standard proteins, Stepped HCD and Fragmentation Mesh always generate more matched fragment ions. (Fig. 3) Fragmenting multiple charge states could also improve the fragmentation in general. For ubiqutin and myoglobin, we could observe obvious improvement, especially for myglobin, the number of matched fragment ions doubled for fragmenting charge states 17, 21 and 25 together than just fragmenting charge state 21 under stepped HCD. Except for cytochrome C, there is limited improvement from multiple charge states under one NCE; the matched fragment ions from multiple charge states even decrease under stepped HCD. The process used to select the multiple charge states for each proteoform could also be important, as the best charge state for HCD always is lower than the one with the highest intensity. Further studies for better selection of charge states are needed. The improvement from the multiple NCE is obvious for all three standard proteins. For cytochrome C and myoglobin, use the multiple energies doubled the number of matched fragment ions.

Yeast sample

We applied the four fragmentaion schemes to complex proteoform samples from size-seperated (GELFrEE) yeast cell lysates.[5] Two different fractionations were selected to perform the four fragmentation schemes. All methods follow the top 2 strategy and keep the same mass range for the full MS scan. Based on the results from standard protein experiments, we designed MetaDrive to use stepped HCD with NCE 15-20-25 and to select three lower charge states of each proteoform.

Ultimately, we observed an increase of identification rate in top-down MS due to multiple charge states and multiple collision energies for both fractions. (Fig. 4) The identification rate is calulated as the number of identified spectra divided by the number of total MS2 scans. The usage of multiple energies contributed most of the improvement. (Fig. 4) We observe that Mesh has slightly better identification rate than stepped HCD. However, we didn't find clear evidence that the improvement was due to improved fragmentation efficiency. (S Fig.1) Fragmentation Mesh may not guarantee to get the best fragmentation for a proteoform, but could reduce the possibility of getting low quality fragmentation of certain charge state. It is also possible that the improvement of the Mesh method may be depressed for the following reasons. First, it is possible to select multiple precursors from different proteoforms due to deconvolution error. A deconvolution algorithm with high accuracy is required. Second, selecting multiple charge states of one proteoform could also increase the possibility of coisolation. Stepped HCD could be applied with the vender's control program and it could be used as an alternative method of Fragmentation Mesh for top-down proteomics.

Conclusions

We proved that stepped HCD could improve the fragmentation efficiency and described a novel ‘Mesh’ method for top-down proteomics. The current DDA method randomly selects one charge state of a proteoform but does not often generate a high yield of matched fragment ions. The Mesh fragmentation implemented in MetaDrive could fragment multiple charge states of one proteoform, which could also be applied to other dissociation methods in the future. Our results showed an increased identification and improved sequence coverage of stepped HCD and the Mesh fragmentation method for complicated protein samples.

FIGURES

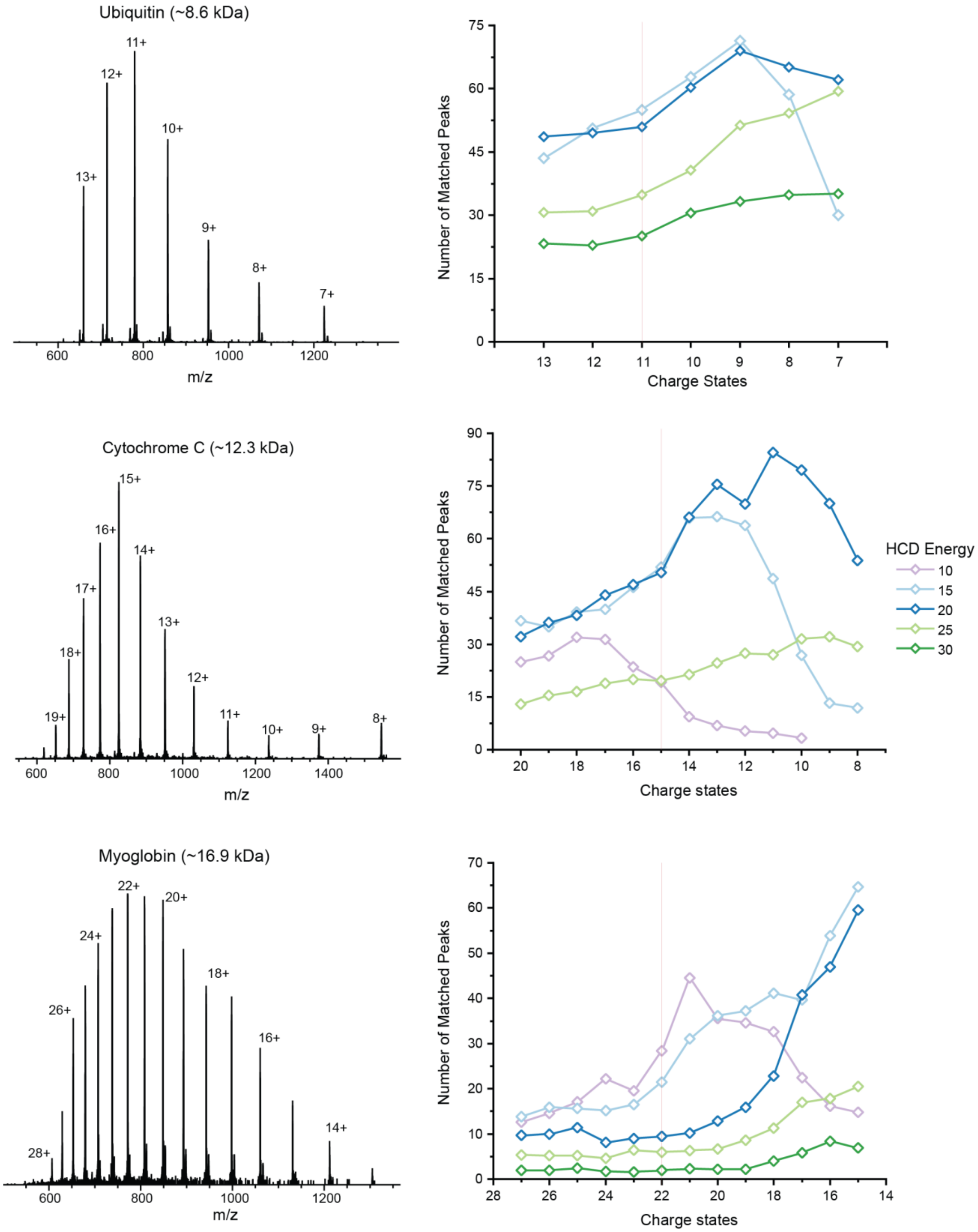
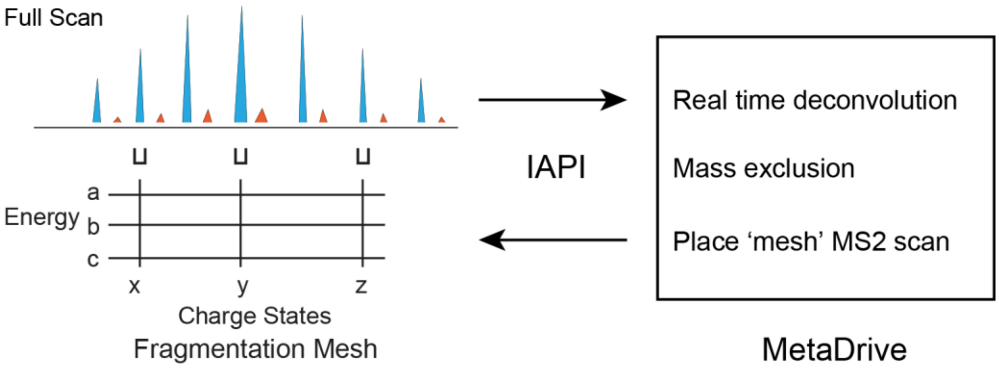


Figure 1. Fragmentation efficiency of different charge states and different NCE. Three

standard proteins (ubiquitin, cytochrome C and myoglobin in sequential) are used to study the effect of charge state and CE on fragmentation efficiency. The number of matched fragment ions changed with charge states and energies. No unique condition works for every standard protein.

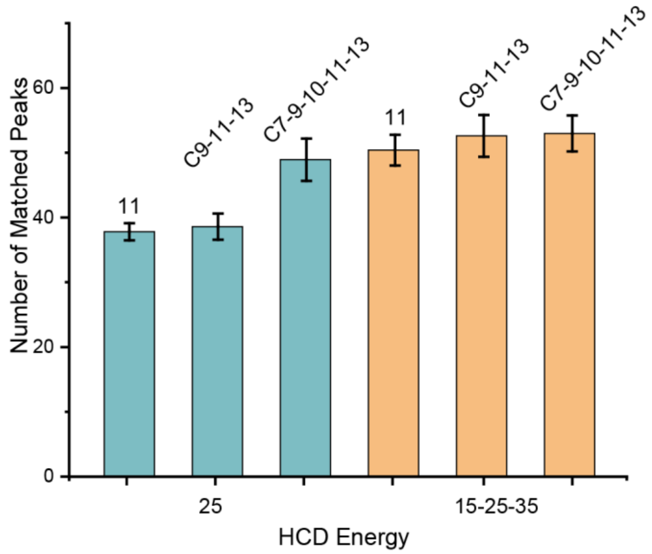


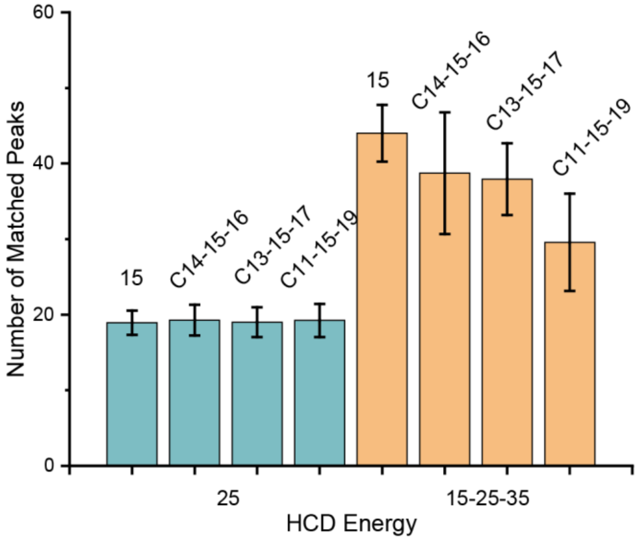
Real time deconvolution

Mass exclusion

Place ‘mesh’ MS2 scan

Figure 2. Workflow of Fragmentation Mesh with MetaDrive. The instrument control software program MetaDrive performed real-time deconvolution of each new full MS spectrum. The deconvolution results contain unknown proteoforms with defined charge states. MetaDrive could choose a proteoform and apply a Fragmentation Mesh. The Fragmentation Mesh contains multiple selected charge states (marked x, y and z) of the proteoform and stepped HCD with different energies (marked a, b and c).


Number of Matched Peaks
0
20
40
60
11
C9-11-13
C7-9-10-11-13
11
C9-11-13
C7-9-10-11-13
25
15-25-35
HCD Energy


Number of Matched Peaks
0
20
40
60
15
C14-15-16
C13-15-17
C11-15-19
15
C14-15-16
C13-15-17
C11-15-19
25
15-25-35
HCD Energy

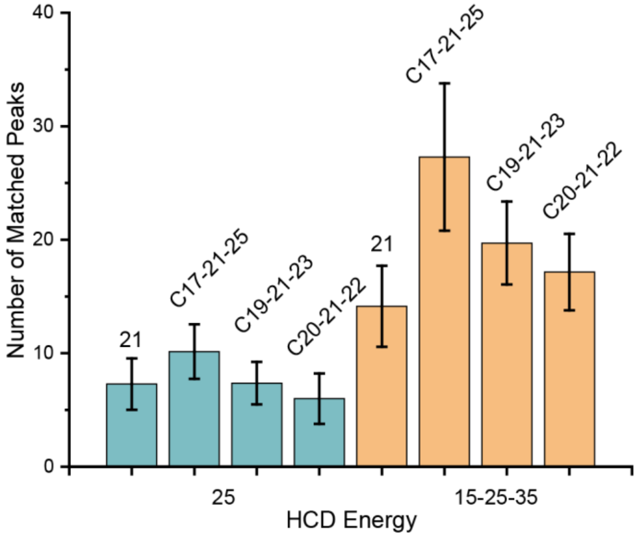

Number of Matched Peaks
0
10
20
30
40
21
C17-21-25
C19-21-23
C20-21-22
21
C17-21-25
C19-21-23
C20-21-22
25
15-25-35
HCD Energy

Figure 3. Fragmentation efficiency between different fragmentation schemes. Three standard proteins (ubiquitin, cytochrome C and myoglobin) are used to test how the different fragmentation schemes improve the number of matched peaks. For each protein, the charge state with the highest intensity is selected for the Point method, and multiple charge states near it are selected for the Line and Fragmentation Mesh methods. Charge states are labeled on each column. Single NCEs are represented with green columns and multiple NCEs are represented with orange columns.

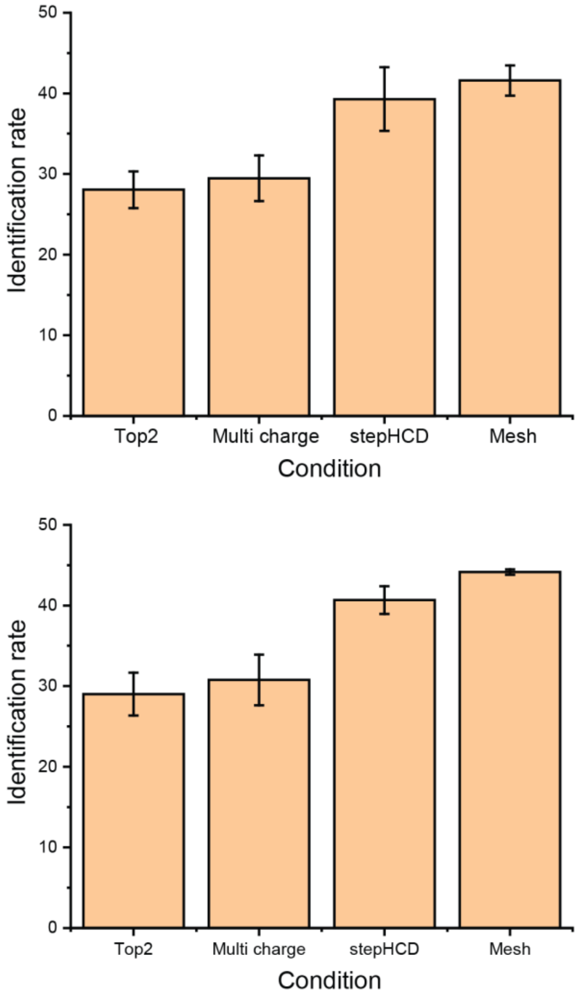


Figure 4. Identification rate of different fragmentation schemes. Two yeast Gel-Free fractions are used to test how the different fragmentation schemes improve identification rate. For each method, three technical replicates were performed.

REFERENCE

(1) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-down Proteomics: Ready for Prime Time? *Anal. Chem.* **2017**, *90* (1), 110–127.

(2) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499–519.

(3) Smith, L. M.; Kelleher, N. L.; Linial, M.; Goodlett, D.; Langridge-Smith, P.; Goo, Y. A.; Safford, G.; Bonilla, L.; Kruppa, G.; Zubarev, R. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10* (3), 186–187.

(4) Smith, L. M.; Kelleher, N. L. Proteoforms as the next Proteomics Currency. *Science (80-.).* **2018**, *359* (6380), 1106–1107.

(5) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M. Mapping Intact Protein Isoforms in Discovery Mode Using Top-down Proteomics. *Nature* **2011**, *480* (7376), 254–258.

(6) Durbin, K. R.; Fornelli, L.; Fellers, R. T.; Doubleday, P. F.; Narita, M.; Kelleher, N. L. Quantitation and Identification of Thousands of Human Proteoforms below 30 KDa. *J. Proteome Res.* **2016**, *15* (3), 976–982.

(7) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. Large-Scale Top-down Proteomics of the Human Proteome: Membrane Proteins, Mitochondria, and Senescence. *Mol. Cell. Proteomics* **2013**, *12* (12), 3465–3473.

(8) Brunner, A. M.; Lössl, P.; Liu, F.; Huguet, R.; Mullen, C.; Yamashita, M.; Zabrouskov, V.; Makarov, A.; Altelaar, A. F. M.; Heck, A. J. R. Benchmarking Multiple Fragmentation Methods on an Orbitrap Fusion for Top-down Phospho-Proteoform Characterization. *Anal. Chem.* **2015**, *87* (8), 4152–4158.

(9) Riley, N. M.; Westphall, M. S.; Coon, J. J. Activated Ion-Electron Transfer Dissociation Enables Comprehensive Top-down Protein Fragmentation. *J. Proteome Res.* **2017**, *16* (7), 2653–2659.

(10) Kreimer, S.; Belov, M. E.; Danielson, W. F.; Levitsky, L. I.; Gorshkov, M. V; Karger, B. L.; Ivanov, A. R. Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-up Proteomic Profiling. *J. Proteome Res.* **2016**, *15* (10), 3563–3573.

(11) Durbin, K. R.; Fellers, R. T.; Ntai, I.; Kelleher, N. L.; Compton, P. D. Autopilot: An Online Data Acquisition Control System for the Enhanced High-Throughput Characterization of Intact Proteins. *Anal. Chem.* **2014**, *86* (3), 1485–1492.

(12) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and Localized Protons: A Framework for Understanding Peptide Dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399–1406.

(13) Liu, M.-Q.; Zeng, W.-F.; Fang, P.; Cao, W.-Q.; Liu, C.; Yan, G.-Q.; Zhang, Y.; Peng, C.; Wu, J.-Q.; Zhang, X.-J. PGlyco 2.0 Enables Precision N-Glycoproteomics with Comprehensive Quality Control and One-Step Mass Spectrometry for Intact Glycopeptide Identification. *Nat. Commun.* **2017**, *8* (1), 1–14.

(14) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins: A Combinatorial Approach. *Mol. Cell. Proteomics* **2010**, *9* (12), 2772–2782.

(15) Meier, F.; Geyer, P. E.; Virreira Winter, S.; Cox, J.; Mann, M. BoxCar Acquisition Method

Enables Single-Shot Proteomics at a Depth of 10,000 Proteins in 100 Minutes. *Nat. Methods* **2018**, *15*, 440–448.

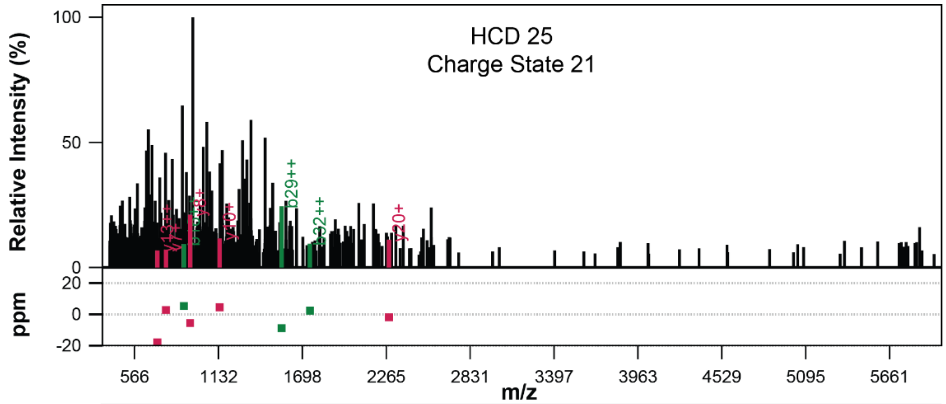
(16) Sun, R.-X.; Luo, L.; Wu, L.; Wang, R.-M.; Zeng, W.-F.; Chi, H.; Liu, C.; He, S.-M. PTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal. Chem.* **2016**, *88* (6), 3082–3090.

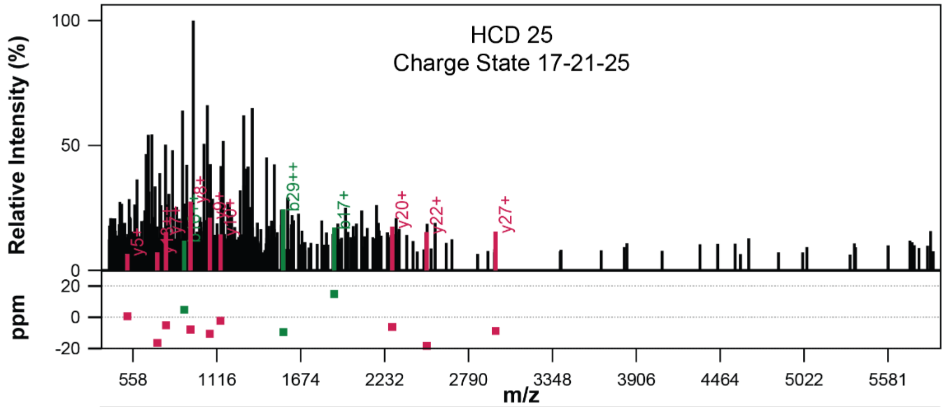
(17) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K. Informed-Proteomics: Open-Source Software Package for Top-down Proteomics. *Nat. Methods* **2017**, *14* (9), 909–914.

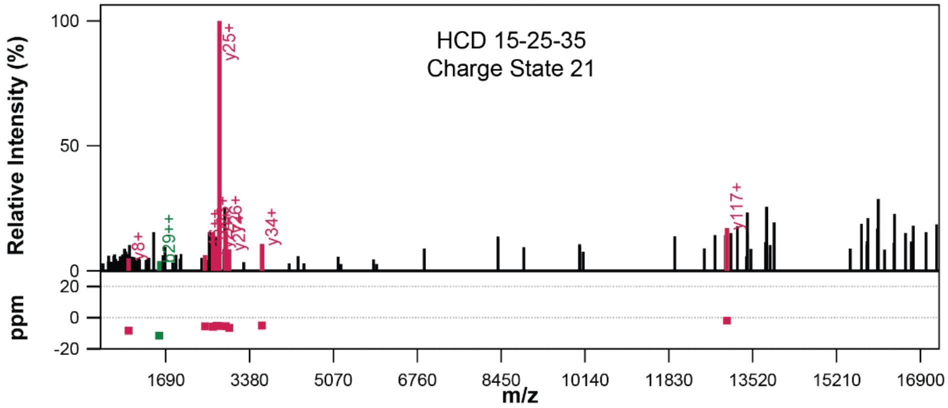
(18) Fabris, D.; Kelly, M.; Murphy, C.; Wu, Z.; Fenselau, C. High-Energy Collision-Induced Dissociation of Multiply Charged Polypeptides Produced by Electrospray. *J. Am. Soc. Mass Spectrom.* **1993**, *4* (8), 652–661.

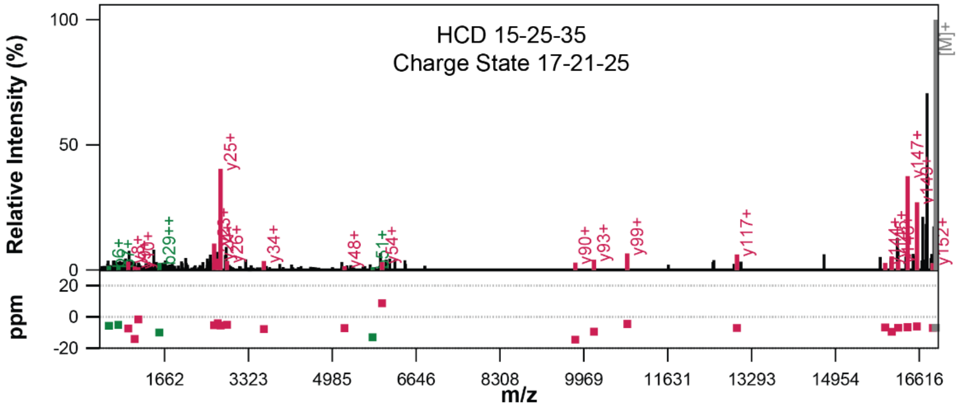
(19) Schaffer, L. V; Shortreed, M. R.; Cesnik, A. J.; Frey, B. L.; Solntsev, S. K.; Scalf, M.; Smith, L. M. Expanding Proteoform Identifications in Top-down Proteomic Analyses by Constructing Proteoform Families. *Anal. Chem.* **2018**, *90* (2), 1325–1333.

SUPPORTING INFORMATION

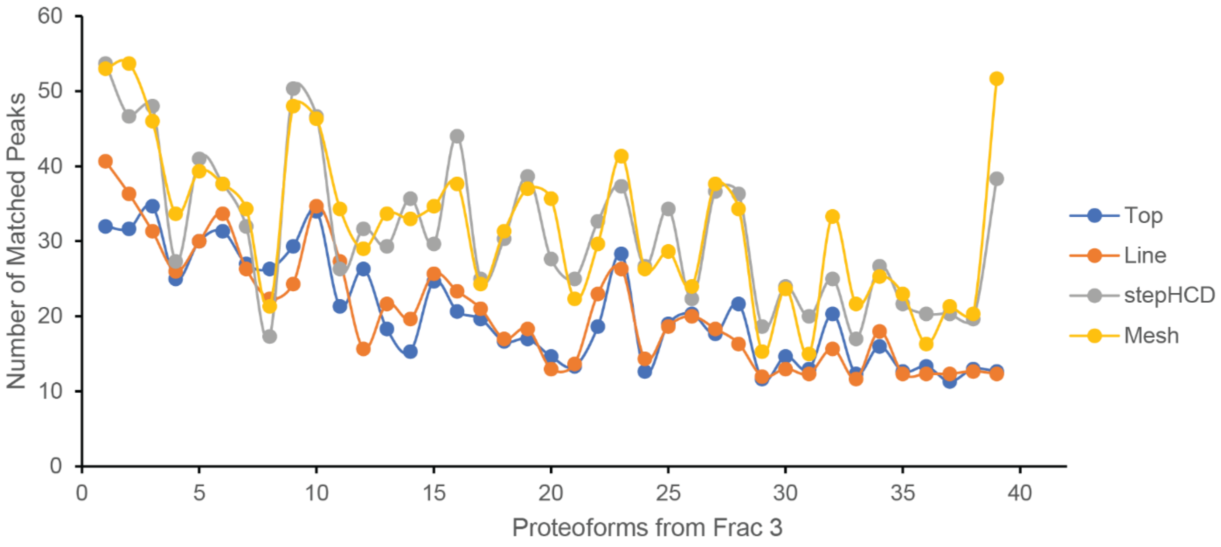


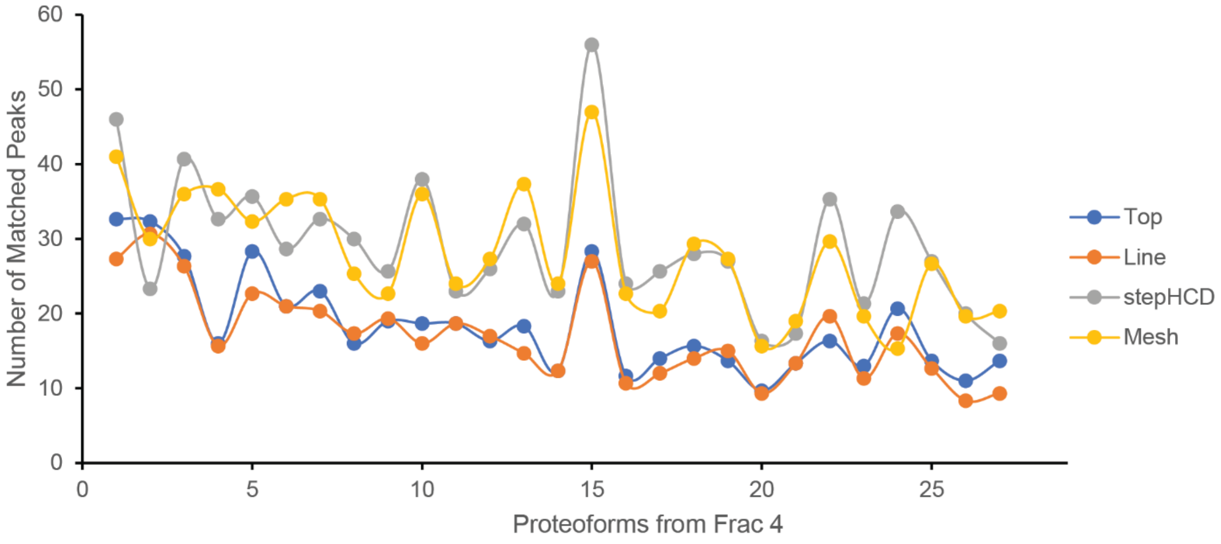






S Figure 1. Comparison spectra of myoglobin from four different fragmentation methods. The spectra are exported from the pTOP2 software program. Compared with the Top method, the Line and Stepped HCD methods increased the number of matched peaks and the Mesh method produced the most matched fragment peaks.





S Figure 2. Comparison of fragmentation efficiency for the four fragmentation methods of shared proteoforms. The proteoforms are identified in all three replicates of the 4 fragmentation methods for the two yeast GEL-Free fractions.The stepped-HCD and the Mesh methods produced more matched peaks. The difference between stepped-HCD and Mesh methods is not obvious.

SI Table 1. Parameters used of pTop2

	pTop2
Isolation Width	15
Mixture Spectra	Unchecked
Maximum Charge	50
M/Z tolerance	20
Maximum Mass	50000
S/N Ratio	1.5
Precursor Tolerance	5.2 Da
Fragment Tolerance	15 ppm
Max Truncated Mass	20000 Da
Search Mode	Tag-Indexed
Second Search	Checked
Max PTM Positions	3
Max Mod Mass	500 Da
Unexpected PTMs	0
Fixed Modification	null
Variable modification	Oxidation[M] Dehydro[C] Acetyl[Protein N-term]
FDR	1%
Separate Filtering	Checked
Quantification	null

Supplementary Note. Real-time Instrument Control Improves Feature Detection with Dynamic BoxCar

Here we present another possible instrument control application. Proteoforms often appear in multiple charge states in intact or top-down MS1 spectra and MS1 spectrum is always dominated by a few proteomforms, which results in low abundant proteoforms be suppressed or the high abundant proteoforms being selected multiple times for fragmentation in top-down. The major downside to this redundancy is that many lower abundance proteoforms will never be detected in MS1 spectra or selected for fragmentation. We created a new open source software program, MetaDrive, that performs online deconvolution of MS1 spectra, which permitting simultaneous determination the m/z ranges being suppressed and proteoforms available for fragmentation. The software then generates extra MS1 spectrum by dynamic selection of the suppressed m/z ranges (Dynamic BoxCar) or controls selection for fragmentation, which substantially increases feature detections and precursor selection efficiency and increases the number of unique proteoform identifications.

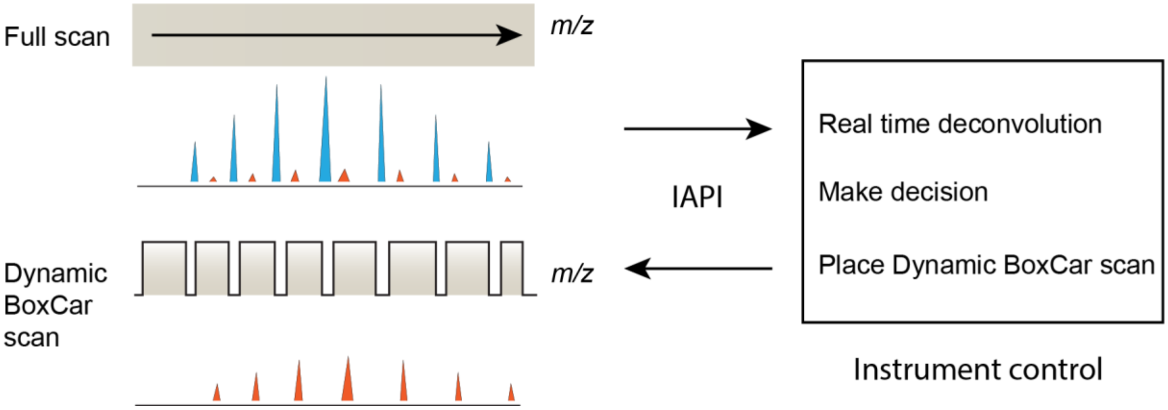


Figure 1. Workflow of Dynamic BoxCar in MetaDrive. MetaDrive performs on-the-fly isotope and charge state deconvolution followed by Dynamic BoxCar or ‘smart’ precursor selection. MetaDrive controls an Orbitrap Q-Exactive through Thermo IAPI (Instrument Application Programming Interface). For intact/top-down MS, the software could perform Dynamic BoxCar to increase the feature detection. For top-down MS, MetaDrive uses a deconvoluted precursor mass, rather than m/z, exclusion methodology where all corresponding charge and isotope states for a given mass are excluded from re-selection, which eliminates the repeated selection of abundant proteoforms. MetaDrive is written by C# and will be made publicly available to the community as open-source code.

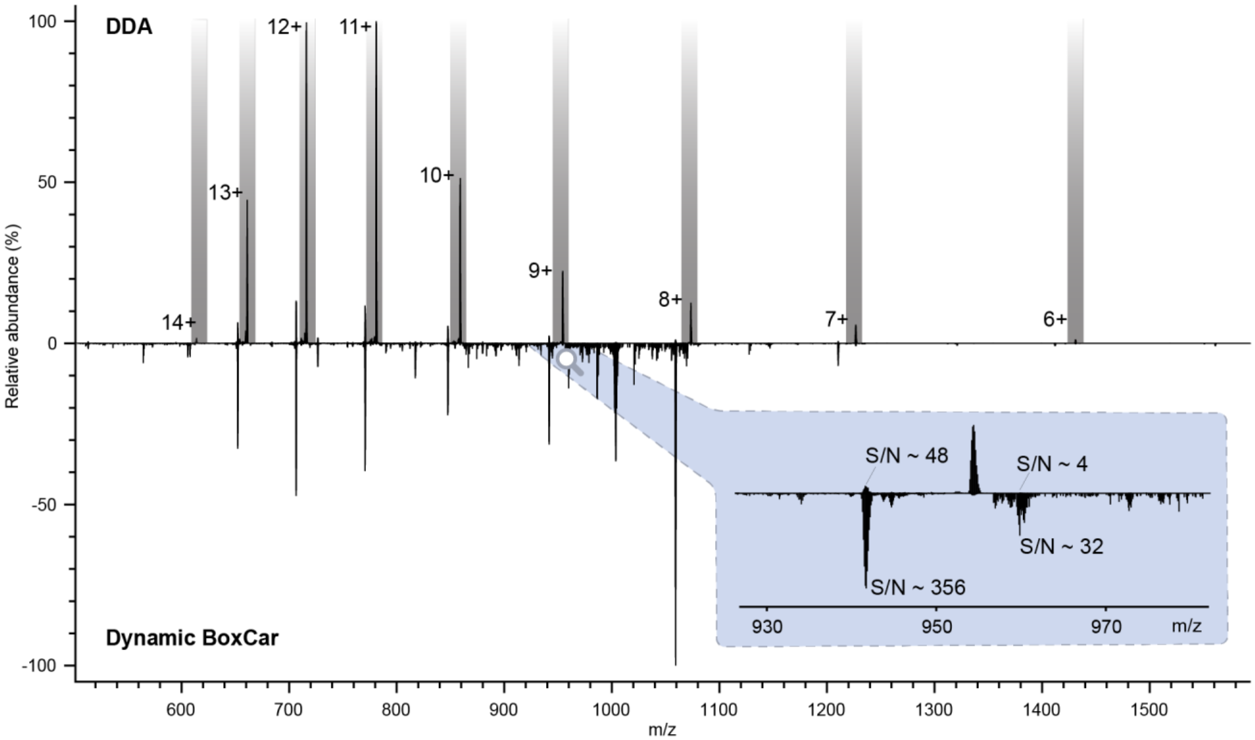


Figure 2. Example shows that blocking the high abundant proteoform with Dynamic BoxCar strategy could improve S/N, and improve the chances to get good signals for low abundant proteoforms.

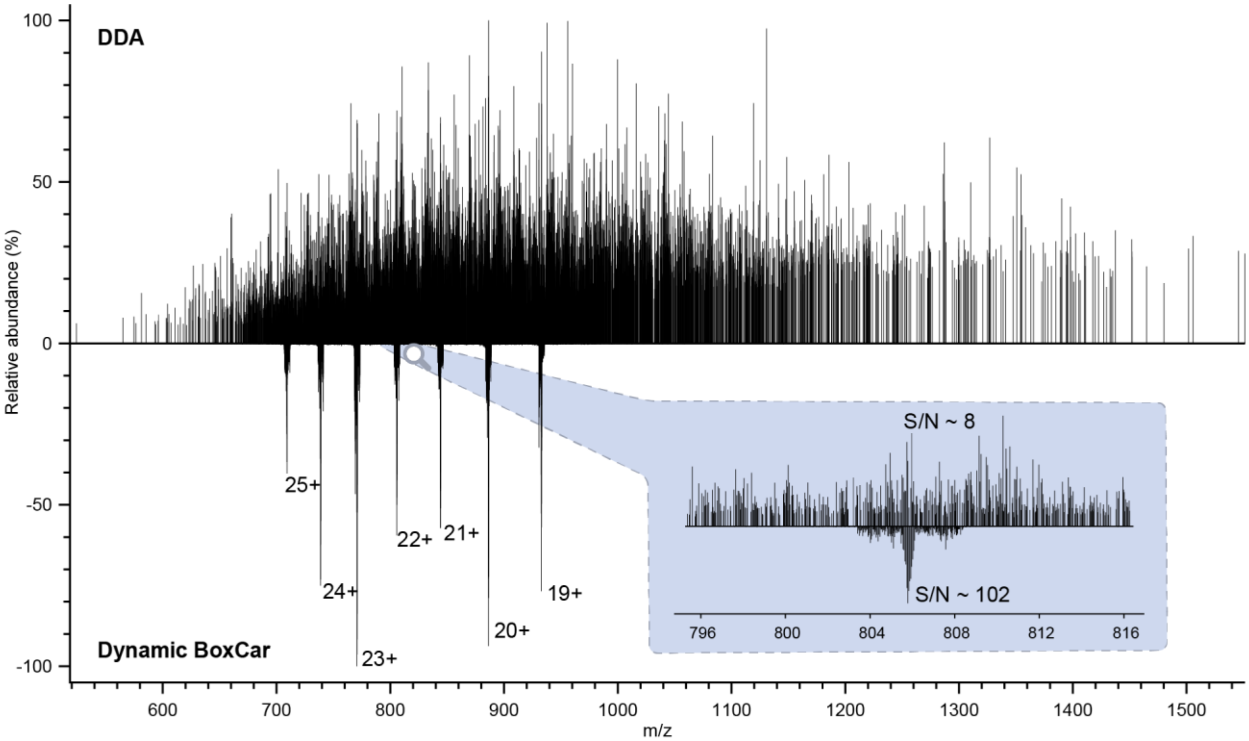


Figure 3. Example shows that proteoforms are still detectable with deconvolution algorithm with Dynamic BoxCar, where the whole spectrum has low S/N.

Chapter 5

Conclusions and Future directions

Conclusions

This dissertation describes the development of data analysis software modules for cross-link proteomics and O-glycoproteomics, which provide robust and efficient tools for the research community[1,2]. Software programs for executing these tasks existed prior to the development of our tools, which used traditional database search algorithms.

The prior algorithms and software for cross-link peptide identification were not able to efficiently analyze some types of proteomics data nor could they accommodate data from both cleavable and non-cleavable cross-links. The research described in this dissertation addressed these limitations by applying an efficient ion-indexed open search strategy[1] (**Chapter 2**). Our software module is faster for non-cleavable cross-links than the commonly used XlinkX[3].

The prior algorithms and software for O-glycopeptide identification were not able to efficiently localize glycosites using a high complexity glycan database. We developed the currently most efficient O-Glycopeptide identification software program O-Pair Search using the ion-indexed open search and introduced the use of graph-based localization for O-glycoproteomics[2,4] (**Chapter 3**). O-Pair search can reduce O-glycopeptide search times by >2000x over the most widely used commercial glycopeptide search tool, Byonic[2,5]. Additionally, O-Pair Search identifies more O-glycopeptides than Byonic and provides O-glycosite localizations using graph theory and localization probabilities. We also introduced a novel classification scheme to unify data reporting across the glycoproteomic community.

The software program modules described in this thesis are widely applicable and freely available through the open-source software program MetaMorpheus[6]. In this chapter, I will discuss some of the challenges and future ideas in cross-linking mass spectrometry proteomics and glycoproteomics data analysis.

Sensitivity for cross-link search

The sensitivity of MetaMorpheus for non-cleavable cross-link search is not high based on our own observations as well as reported by a paper published by another group a year after our publication[7]. The sensitivity is used to represent the false negatives, a 100% sensitivity which means 0 false negatives is the goal for identifications. For non-cleavable cross-link search, the paper reported 70% sensitivity of MetaMorpheus using a stimulated dataset and reported even lower sensitivity for entrapment experiments, where unrelated large-scale databases are incorporated in the search process. We improved the workflow of our search program and the sensitivity increased to about 85% using the stimulated data. The sensitivity is not as high as pLink2, which is another cross-link mass spectrometry proteomics software program reported to have the highest sensitivity (> 95%). It is similar to the popular software program Kojak[8], which is the second best regarding the sensitivity.

A few strategies we could use to improve the sensitivity include: 1. Improve the scoring function. The current Morpheus Score used in MetaMorpheus which mainly counts the number of matched peaks does not take into account peak intensity. In cross-link search, the beta peptide of a cross-linked doublet sometimes does not fragment well and generates only a small number of matched peaks. Due to the easy mismatch of beta peptides, the sensitivity becomes low. 2. Apply machine learning methods to further filter the identifications. The Kojak and pLink2 both utilized machine learning methods to increase identifications. Kojak pipelined with Percolator[9], which is a software program for peptide identification applied semi-supervised machine method, to do so. The pLink2 optimized a support vector machine method. We recently started to test a decision tree method. Note that optimization of the machine learning method requires selection of useful training features. 3. Improve the deconvolution of precursors. Incorrect precursor masses easily result in wrong identifications for cross-link search and better deconvolution functions could yield higher quality precursor masses. The current deconvolution algorithm could be improved by considering peak information from neighbor scans. 4. Recently, peptide spectral prediction[10–12] became available for single peptides or modified peptides with the development of deep learning algorithms. We presume that cross-linked peptides could be similarly predicted and that the resultant cross-linked peptide spectra could be employed to match experimental spectra similarly to how spectral libraries are searched.

Protein structure analysis for cross-link search

Researchers spend a majority of manual efforts to interpret the cross-link identifications for protein structure analysis. Such applications include combining cross-link analysis with CryoEM or X-Ray data to solve labile protein complex structures[13,14]. However, successful use of cross-link proteomics for protein structure analysis is not as common one might expect. One reason is the lack of software support for such integrated analysis. To automate the protein structure analysis process, software program pipelines have been developed to correlate cross-link identifications with protein structure predictions[15]. Such development has been limited to a few labs with experience in both software program development, and in proteomics & protein structure analysis. However, the current software program pipelines require significant effort to master different pieces of programs. Also, there is a lack of interactive software programs to help people alter the protein structure based on cross-links. To make cross-link proteomics more useful for general researchers, it is important to develop protein structure and protein-protein interaction analysis software pipelines with easy access. The candidate software platforms to enable the successful development of such pipeline include: pLink, Kojak, and MetaMorpheus for cross-link proteomics analysis; and Pymol[16] and Rossetta[17] for protein structure analysis.

N-glycoproteomics data analysis

N-glycosylation is an abundant and complex post-translational modification. One of the complexities is from the highly dynamic nature of the glycan structures and compositions, which generates tremendous site-specific heterogeneity or micro-heterogeneity. N-glycan heterogeneity has been reported to affect binding specificities, enzyme activity and functionality. Analysis of intact N-glycopeptides vis mass spectrometry is widely used to preserve site-specific N-glycan heterogeneity information.

Before we developed O-Pair Search for O-glycoproteomics data analysis, we developed a software module for N-glycopeptide identification using an algorithm similar to that of O-Pair Search. N-glycopeptides and O-glycopeptides generate different schema of fragment ions. Specially, N-glycopeptides generally contain one large glycan which generates more sugar-related fragments, while O-glycans are more labile and can occur on multiple sites. The software module is tuned separately for the two types of glycosylation based on the differences between their fragment ions.

Our software module is faster than Byonic[5] regarding N-glycopeptide identification. It is also compatible with data from different fragmentation schemes, which makes it a useful candidate for most researchers. Still there are some limitations of the software module for N-glycopeptide identifications. Compared with pGlyco2[18], our development of N-glycopeptide identification is currently limited by its N-glycan FDR analysis. To address this limitation, we could apply the glycan FDR analysis algorithm used by pGlyco2/Glycresoft[19]. Machine learning methods could also be used for improving N-glycan FDR analysis. A noticeable feature of N-glycopeptides is their highly conserved retention times[19,20]. Researchers have applied retention time analysis for N-glycopeptides to reveal correctness of identification or to infer glyco-peptidoform families, in which some of the glycopeptides are not even fragmented. Such a method could also be applied to our module to further improve the N-glycosylation characterization.

Integrated O-glycopeptide and N-glycopeptide identification

O-glycosylation and N-glycosylation could occur on the same proteins or in the same sample. Some of the current enrichment methods used for glycoproteomics cannot separate the two different types of glycosylated peptides. However, the previous software programs are unable to identify both at the same time, leading to a new category of false identification. Glycans are composed of a variety of different monosaccharides, thus multiple smaller oligosaccharides may have the same mass as a larger single glycan or a combination of different oligosaccharides. In the case where N-glycopeptides coexist with O-glycopeptides, a N-glycopeptide with one large N-glycan is highly likely to be misidentified as an O-glycopeptide with multiple O-glycans in an O-glyco search only software program.

To reduce such possible misidentifications, we integrated our N-glycopeptide and O-glycopeptide software modules to enable the identification of both types at the same time. We tested the integrated software module and found that in three datasets, all of them contained significant amounts of both N-glycopeptides and O-glycopeptides. The three datasets are from either N-glycoproteomics analysis or O-glycoproteomics. Accordingly, we propose that some of the published glycopeptide datasets should be reanalyzed and possible misidentifications should be reported.

The current differentiation between N-glycopeptide and O-glycopeptide identification of the same spectra is based on the Morpheus Score of the matched theoretical fragment ions. The method requires high quality fragmentation of the selected glycopeptides. We propose that machine learning methods such as support vector machines or decision trees could help to differentiate some of the more confusing cases.

Future developments for glycoproteomics data analysis include improving sensitivity using the method described in ‘Sensitivity for cross-link search’ section and quantification analysis using FlashLFQ[21].

REFERENCES

(1) Lu, L.; Millikin, R. J.; Solntsev, S. K.; Rolfs, Z.; Scalf, M.; Shortreed, M. R.; Smith, L. M. Identification of MS-Cleavable and Non-Cleavable Chemically Crosslinked Peptides with MetaMorpheus. *J. Proteome Res. 0* (ja), null.

(2) Lu, L.; Riley, N. M.; Shortreed, M. R.; Bertozzi, C. R.; Smith, L. M. O-Pair Search with MetaMorpheus for O-glycopeptide Characterization. *bioRxiv* **2020**.

(3) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* **2015**, *12* (12), 1179–1184.

(4) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K. Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* **2017**, *14* (9), 909–914.

(5) Bern, M.; Kil, Y. J.; Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinforma.* **2012**, *40* (1), 13–20.

(6) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced global post-translational modification discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17* (5), 1844–1851.

(7) Chen, Z.-L.; Meng, J.-M.; Cao, Y.; Yin, J.-L.; Fang, R.-Q.; Fan, S.-B.; Liu, C.; Zeng, W.-F.; Ding, Y.-H.; Tan, D. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **2019**, *10* (1), 1–12.

(8) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* **2015**, *14* (5), 2190–2198.

(9) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923–925.

(10) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **2019**, *16* (6), 509–518.

(11) Degroeve, S.; Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **2013**, *29* (24), 3199–3203.

(12) Zhou, X.-X.; Zeng, W.-F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S.-M.; Zhang, Z. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **2017**, *89* (23), 12690–12697.

(13) Schmidt, C.; Urlaub, H. Combining cryo-electron microscopy (cryo-EM) and cross-linking mass spectrometry (CX-MS) for structural elucidation of large protein assemblies. *Curr. Opin. Struct. Biol.* **2017**, *46*, 157–168.

(14) Yu, C.; Huang, L. Cross-linking mass spectrometry: an emerging technology for interactomics and structural biology. *Anal. Chem.* **2018**, *90* (1), 144–165.

(15) Gutierrez, C.; Chemmama, I. E.; Mao, H.; Yu, C.; Echeverria, I.; Block, S. A.; Rychnovsky, S. D.; Zheng, N.; Sali, A.; Huang, L. Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling. *Proc. Natl. Acad. Sci.* **2020**, *117* (8), 4088–4098.

(16) DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. protein Crystallogr.* **2002**, *40* (1), 82–92.

(17) Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **2020**, 1–14.

(18) Liu, M.-Q.; Zeng, W.-F.; Fang, P.; Cao, W.-Q.; Liu, C.; Yan, G.-Q.; Zhang, Y.; Peng, C.; Wu, J.-Q.; Zhang, X.-J. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nat. Commun.* **2017**, *8* (1), 1–14.

(19) Klein, J.; Zaia, J. Relative Retention Time Estimation Improves N-Glycopeptide Identifications by LC–MS/MS. *J. Proteome Res.* **2020**, *19* (5), 2113–2121.

(20) Choo, M. S.; Wan, C.; Rudd, P. M.; Nguyen-Khuong, T. GlycopeptideGraphMS: improved glycopeptide detection and identification by exploiting graph theoretical patterns in mass and retention time. *Anal. Chem.* **2019**, *91* (11), 7236–7244.

(21) Millikin, R. J.; Solntsev, S. K.; Shortreed, M. R.; Smith, L. M. Ultrafast peptide label-free quantification with FlashLFQ. *J. Proteome Res.* **2018**, *17* (1), 386–391.

Appendix I

Publications and Presentations

Publications

1. **Lu. L.***, Riley, N.M.*, Shortreed, M.R., Bertozzi, C.R.* & Smith, L.M.* (2020). Comprehensive identification and localization of O-glycopeptides in tandem mass spectra using MetaMorpheus. Nature Methods (Accepted In Principle)
2. Zhong, X., Yu, Q., Ma, F., Frost, D.C., **Lu, L.**, Chen, Z., Zetterberg, H., Carlsson, C., Okonkwo, O. and Li, L. (**2019**). HOTMAQ: A Multiplexed Absolute Quantification Method for Targeted Proteomics. Analytical Chemistry, 91(3), 2112-2119.
3. **Lu, L.**, Millikin, R. J., Solntsev, S. K., Rolfs, Z., Scalf, M., Shortreed, M. R., & Smith, L. M. (**2018**). Identification of MS-cleavable and noncleavable chemically cross-linked peptides with MetaMorpheus. Journal of Proteome Research, 17(7), 2370-2376.
4. Li, B., Li, H., **Lu, L.**, & Jiang, J. (**2017**). Structures of human O-GlcNAcase and its complexes reveal a new substrate recognition mode. Nature Structural & Molecular Biology, 24(4), 362.
5. Hu, C.W., Worth, M., Fan, D., Li, B., Li, H., **Lu, L.**, Zhong, X., Lin, Z., Wei, L., Ge, Y. and Li, L., Jiang, J.(**2017**). Electrophilic probes for deciphering substrate recognition by O-GlcNAc transferase. Nature Chemical Biology, 13(12), 1267.
6. **Lu, L.**, Fan, D., Hu, C. W., Worth, M., Ma, Z. X., & Jiang, J. (**2016**). Distributive O-GlcNAcylation on the highly repetitive C-terminal domain of RNA polymerase II. Biochemistry, 55(7), 1149-1158.
7. Zhu, F., **Lu, L.**, Fu, S., Zhong, X., Hu, M., Deng, Z., & Liu, T. (**2015**). Targeted engineering and scale up of lycopene overproduction in Escherichia coli. Process Biochemistry, 50(3), 341-346.
8. Liu, Q., Wu, K., Cheng, Y., **Lu, L.**, Xiao, E., Zhang, Y., Deng, Z. and Liu, T. (**2015**). Engineering an iterative polyketide pathway in Escherichia coli results in single-form alkene and alkane overproduction. Metabolic Engineering, 28, 82-90.
9. Liu, R., Zhu, F., **Lu, L.**, Fu, A., Lu, J., Deng, Z., & Liu, T. (**2014**). Metabolic engineering of fatty acyl-ACP reductase-dependent pathway to improve fatty alcohol production in Escherichia coli. Metabolic Engineering, 22, 10-21.
10. Zhu, F., Zhong, X., Hu, M., **Lu, L.**, Deng, Z., & Liu, T. (**2014**). In vitro reconstitution of mevalonate pathway and targeted engineering of farnesene overproduction in Escherichia coli. Biotechnology and Bioengineering, 111(7), 1396-1405.

Conference Presentations

1. **Lu, L.**, Shortreed, M. R., Millikin, R. J., Scalf, M., & Smith, L. M. Identification of N-glycopeptides with MetaMorpheus. Poster presentation at the 67th American Society for Mass Spectrometry (ASMS) Annual Conference, **June 2019**, Atlanta, GA
2. **Lu, L.**, Millikin, R. J., Solntsev, S. K., Rolfs, Z., Scalf, M., Shortreed, M. R., & Smith, L. M. Identification of MS-cleavable and noncleavable chemically cross-linked peptides with MetaMorpheus. Poster presentation at the 66th American Society for Mass Spectrometry (ASMS) Annual Conference, **June 2018**, San Diego, CA