**Proteoform identification using multiplexed top-down mass spectra**

Zhige Wang, Xingzhao Xiong, and Xiaowen Liu

1Deming Department of Medicine, Tulane University, New Orleans, Louisiana, 70112, United States

**Abstract**

Top-down mass spectrometry (TDMS) is the method of choice for analyzing intact proteoforms as well as post-translational modifications and sequence variations on proteoforms. In TDMS experiments, multiple proteoforms are often co-fragmented in tandem mass spectrometry (MS/MS) analysis, resulting in multiplexed TD-MS/MS spectra. As multiplexed TD-MS/MS spectra are more complex than common spectra generated from single proteoforms, these spectra pose a significant challenge for proteoform identification and quantification. Here we present TopMPI, a new computational tool specifically designed for the identification of multiplexed TD-MS/MS spectra. Experimental results showed that TopMPI significantly increased proteoform identifications and reduced identification errors in multiplexed TD-MS/MS spectral identification compared with existing tools.

**Introduction**

Top-down mass spectrometry (TDMS) has attracted increasing attention in proteomics due to its ability to directly examine intact proteoforms and characterize proteoforms with various alterations stemming from gene mutations, splicing variants, and post-translational modifications (PTMs) [1, 2]. Unlike bottom-up MS, in which the combinatorial patterns of PTMs on proteoforms are lost during enzymatical digestion of proteoforms prior to MS, TDMS enables the analysis of combinatorial patterns of PTMs and other variations on proteoforms and their biological functions [3].

In a typical TDMS experiment, MS1 spectra are generated for profiling proteoforms in the sample, and data-dependent acquisition (DDA) [3] or data-independent acquisition (DIA) [4] is employed to isolate and fragment specific proteoform ions within selected isolation windows to generate tandem mass spectrometry (MS/MS) spectra. Multiplexed MS/MS spectra containing fragments from multiple co-fragmented proteoforms are frequently observed, especially in top-down DIA-MS, adding a layer of complexity to the TDMS data analysis.

Many methods have been proposed for multiplexed DIA mass spectral identification in bottom-up proteomics [5-8]. When spectral libraries are available, library spectra are matched to bottom-up multiplexed mass spectra for peptide identification [6, 9]. Otherwise, multiplexed mass spectra are demultiplexed using the similarity of the retention time profiles of precursor and fragment ions [5], and the demultiplexed spectra are searched against a protein database for peptide identification.

Both database search and spectral library search have been used for identifying multiplexed DDA mass spectra in bottom-up proteomics [7, 8, 10]. A main difference between DIA and DDA multiplexed mass spectra is that the demultiplexing method for DIA-MS does not work for DDA-MS because fragment ion retention time profiles are not available for most multiplexed DDA spectra.

Several top-down DIA MS studies were reported recently [4, 11] and the demultiplexed demultiplexing method has been extended to analyze top-down multiplexed DIA-MS data [4]. But because of the limitations in top-down DIA MS data analysis, DDA MS is still the dominant method in top-down proteomics [2].

There is still a lack of software tools for identifying top-down multiplexed DDA mass spectra. Many computational tools, like MSPathFinder [12], ProSightPD [13],TopPIC[14], and TopMG [15], have been developed for spectral identification by database search. While these tools are efficient for identifying MS/MS spectra generated from single proteoforms, a common limitation is that they are not able to identify multiple proteoforms from multiplexed spectra. Spectral library-based method in bottom-up MS can be extended to top-down MS [8], but the methods rely on comprehensive spectral libraries. The database search methods [7, 10] for analyzing multiplexed bottom-up DDA-MS data are inefficient for analyzing multiplexed top-down DDA-MS data because proteoforms identified in top-down MS often contain unknown mass shifts and these database search methods in bottom-up MS are not designed for identifying proteoforms with unknown mass shifts.

To address this challenge, we present TopMPI (TOP-down mass spectrometry-based Multiple Proteoform Identification), a new software tool for identifying multiple proteoforms from multiplexed TD-MS/MS spectra by database search. TopMPI is capable of identifying multiple proteoforms with unknown mass shifts from multiplexed top-down DDA mass spectra and increasing proteoform identifications in complex biological samples compared with software tools for identifying single proteoforms from top-down DDA mass spectra.

**Methods**

**E. coli sample preparation**

E. coli K12 cells were pelleted through centrifugation at 5,000×g and 4° C for 5 min and washed with 5mL 1x PBS. Cell pellets were resuspended in 200uL 25 mM ammonium bicarbonate (ABC) buffer with the addition of 1x (v/v) protease inhibitor (EDTA free). Cells were lysed by using 0.1mm beads that were mixed with cell and ABC buffer with ratio 1: 1: 2 (v/v) and beating for 3 min. After beads beating, the cell lysate was centrifuged at 12,000×g and 4° C for 4 min to remove the insoluble debris. Then Amicon Ultra-0.5 centrifugal filter will be applied for desalting and concentrated by centrifuged at 14,000×g and 4° C for 20 min. 1 µL of 1M DTT was added to the lysate and allowed to react at 55°C for 45 minutes and then 2.5 µL of 1M IAA were added and allowed to react at room temperature for 30 minutes. The concentration of the lysate was measured using the Pierce BCA Protein Assay Kit.

**Top-down RPLC-MS/MS analysis**

A total of 300ng E. coli protein was analyzed using a Thermo Scientific (Waltham, MA, USA) Ultimate 3000 LC system with a C2 RPLC capillary column (100 μm i.d., 60 cm length, Richland, WA, CoAnn) coupled with a Thermo Orbitrap Lumos mass spectrometer (Waltham MA, USA). Mobile phase A was 0.1% FA in water. Mobile phase B was 0.1% FA, 50% acetonitrile with 50% isopropanol. A 98-min gradient (0-5 min 5%, 5-7 min for 5% to 35%, 7-10 min for 35% to 50%, 10-97 min for 50% to 80%, 97-98 min for 80% to 99% was applied with a flow rate of 400 nL/min.

MS1 scans were collected at a resolution of 240,000 (at 200 m/z) with 4 microscans and a scan range 720-1200 *m*/*z*. The top 6 precursors in each MS1 scan were selected for Higher-energy C-trap dissociation (HCD) MS/MS analyses with following settings: the precursor isolation window was 3 *m*/*z*, the normalized collision energy was 30%; the Automated Gain Control (AGC) target was 106, and the maximum injection time was 500 ms; the microscan number was 1; the resolution was 60,000 (at 200 m/z); and the scan range was 400-2000 *m*/*z*.

**Top-down mass spectral preprocessing**

Top-down MS raw data files were converted to mzML files using msconvert [16], and then the mzML files were analyzed by TopFD (version 1.7.6 and Supplemental Table S1 for parameter settings) [17] for spectral deconvolution and feature detection. For each reported proteoform feature with multiple charge states, a single charge proteoform feature (SCPF) was obtained for each charge state of the feature. For each MS/MS spectrum, the SCPFs observed in its isolation window were ranked based on their total peak intensities within the window, and the top two SCPFs were assigned to the spectrum.

**Primary precursor selection**

An MS/MS spectrum is treated as a multiplexing one if the ratio of the total peak intensities of the second most abundant and first most abundant SCPFs in the isolation window is higher than 𝛼, where 𝛼 is user specified parameter (the default setting of 𝛼 is 20%). For a multiplexed spectrum with two precursors, two rounds of database searches are conducted with the aim of identifying two proteoforms. The precursor used in the first round of database search is referred to as the primary precursor, and the other precursor the secondary precursor. Note that the primary precursor may have a lower intensity than the secondary precursor.

For an MS/MS spectrum *S* with two SCPFs *F*1 and *F*2, where *F*1 has a higher total peak intensity in the isolation window than *F*2, the primary precursor is determined as follows. Let *S*1 be the spectrum with precursor *F*1 and all fragment masses in *S*, and let *S*2 be the spectrum with precursor *F*2 and all fragment masses in *S* (Fig. 1a). The two spectra are searched against its corresponding protein sequence database for proteoform identification using TopPIC (version 1.7.6) without filtering of reported PrSMs. If TopPIC reports only one PrSM from the two spectra, then precursor of the spectrum with the PrSM is the primary one. If a PrSM is reported for each of two spectra, the precursor of the PrSM with more matched experimental fragment masses is chosen as the primary one.

we will first check if the two PrSMs are consistent and then determine the primary precursor.

Two PrSMs are inconsistent if they are matched to the same protein or most of their matched fragment masses are shared. Let *M*1 and *M*2 are the sets of matched experimental fragment masses in the two PrSMs reported by database search, respectively. The two PrSMs are inconsistent if , where is a user specified parameter ( in the experiments). There are two possible cases for inconsistent PrSMs. First, the two features *F*1 and *F*2 are from two similar proteoforms of the same protein and most of their fragment masses overlap with each other. Second, the two features *F*1 and *F*2 are from two proteins, and one of the two PrSMs is incorrect.

If the PrSMs of *S*1 and *S*2 are consistent, then precursor with more matched experimental fragment masses in the reported PrSM is chosen to be the primary one. Otherwise, the primary precursor is selected using the following method. If the PrSM of *S*2 contains *𝛾* more matched theoretical fragment masses than that of *S*1, where *𝛾* is a user specified parameter (*𝛾* = 3 in the experiments), and the number of unknown mass shifts in the PrSM of *S*2 is no more than that in the PrSM of *S*1, then precursor *F*2 is the primary one. Otherwise, precursor *F*1 is the primary one.

**Proteoform identification by multiplexed top-down DDA mass spectra**

TopMPI uses a method similar to CharmeST [10] to identify two proteoforms from a multiplexed top-down DDA mass spectra (Fig. 1b). For an MS/MS spectrum *S* with a primary precursor *F*1 and a second precursor *F*2, a two-round search is used to identify two proteoforms from *S*. In the first found, the primary precursor *F*1 and the fragment masses in spectrum *S*are searched against the corresponding protein sequence database concatenated with a decoy database using TopPIC (version 1.7.6). If the *E*-value of the best PrSM reported by TopPIC is less than 0.01, the PrSM is kept and the matched fragment masses in *S* are removed. Otherwise, the PrSM is not reported and no fragment masses in *S* are removed. In the second round, the secondary feature and the remaining fragment masses in *S* are searched against the same target-decoy protein sequence database using TopPIC. Finally, the PrSMs reported in the first and second rounds are filtered by 1% spectrum-level false discovery rate (FDR), separately.

**Results**

**Overview of TopMPI**

TopMPI is designed to identify two proteoforms from a multiplexed top-down DDA MS/MS spectrum with two precursor ions. The first function of TopMPI is to determine the order of the two precursor ions for database search based proteoform identification (Fig. 1a). A common error in multiplexed top-down mass spectral identification is that the most abundance precursor and the fragment masses from another precursor are combined to identify an incorrect proteoform [18]. The objective of the function is to choose the precursor with the correct database search identification as the primary (see Methods). The other precursor is assigned as the secondary precursor. After the primary and secondary precursors are determined, TopMPI uses a method similar to CharmeST [10] to identify two proteoforms from a multiplexed top-down DDA mass spectra (Fig. 1b). TopMPI first uses the primary precursor and all the fragment masses to identify a PrSM by database search). Then it removes from the spectrum the matched fragment masses in the reported PrSM and generates a new spectrum using the secondary precursor and the remaining fragment masses for spectral identification (see Methods).

**Precursor intensity and spectral identification**

The *E. coli* K12 data with 10320 MS/MS spectra (see Methods) was used to study how the precursor intensities in the isolation window affect the identification of multiplexed spectra. After MS data preprocessing, TopPIC (version 1.7.6 and parameter settings in Supplemental Table S2) was used to search the MS/MS spectra against the UniProt *E. coli* K12 proteome database (version September 7, 2023; 4530 entries) concatenated with a decoy database of the same size. After the first PrSM was reported, the fragment masses matched to the proteoform was removed and a new MS/MS spectrum was generated using the second most abundant precursor and the remaining fragment masses and search against the same target-decoy database using TopPIC with the same parameter settings. The PrSMs reported by the first precursor and the second precursor were filtered with a 1% spectrum-level FDR separately. A total of 1991 and 170 PrSMs were reported from first and second precursors, respectively, and two PrSMs were reported for 86 spectra. We computed the intensity ratio of the second precursor and the first precursor for each of the 86 spectra (Supplemental Fig. S1), and the distribution shows all the ratios are more than 20%. Because of this, the parameter 𝛼 for finding multiplexed spectra was set to 0.2 (see Methods). That is, in the experiments, a spectrum was treated as a multiplexed one only if the intensity ratio of the second and first most abundant precursors was no less than 20%.

**Precursor selection for PrSMs**

For a multiplexed TD-MS/MS spectrum with two precursors *F*1 and *F*2, where *F*1 is from protein *P*1 and *F*2 is from protein *P*2, it is possible that the precursor *F*1 and the fragment ions of *F*2 are incorrectly matched to a proteoform of *P*2. These errors are called precursor alignment errors [18].

To assess when such errors are introduced in top-down spectral identifications, we generated four evaluation data sets from the *E. coli* data set as follows. After MS data preprocessing, TopPIC (version 1.7.6 and parameter settings in Supplemental Table S3) was used to search the MS/MS spectra against the UniProt *E. coli* K12 proteome database (version September 7, 2023; 4530 entries) and identified 1094 PrSMs with an E-value cutoff of 10-6, for which the intensity of the top SCPF was more than 85% of the total peak intensity sum of all the SCPFs for 358 PrSMs. The 385 PrSMs were divided into two data sets: MATCH contained 199 PrSMs without any unknown mass shifts and SHIFT contained 159 PrSMs each with an unknown mass shift. For each spectrum in MATCH and SHIFT, we randomly increased or decreased the precursor charge state by 1 and further increased the precursor mass by 79.966 Da. The resulting two data sets are referred to as MATCH-ERROR and SHIFT-ERROR respectively. The two data sets were used to assess PrSMs with an incorrect precursor mass.

The mass spectra in four data sets were searched against the UniProt *E. coli* proteome database using TopPIC (version 1.7.6 and parameter settings in Table S4). With an E-value cutoff of 0.01, the MATCH-ERROR and SHIFT-ERROR dataset reported 112 and 47 PrSMs respectively, out of which 111 (99.1%) and 47 (100%) were matched to the same proteins compared with those in MATCH and SHIFT. The only PrSM in MATCH-ERROR that was matched to a different protein compared with that in MATCH might be an incorrect identification. For each spectrum identified by both MATCH and MATCH-ERROR, we computed (1) the decrease of the number of matched experimental fragment masses in the PrSM in MATCH-ERROR compared with MATCH, in which two masses with different charge states from the same fragment were treated as two different masses; (2) the decrease of the number of matched theoretical fragment masses in the PrSM in MATCH-ERROR compared with MATCH, in which two masses with different charge states from the same fragment were treated as one mass. (3) the increase of the E-value with log transformation (base 10) of the PrSM in MATCH-ERROR compared with MATCH. Similarly, the SHIFT and SHIFT-ERROR data sets were processed similarly (Fig. 3). The differences between SHIFT and SHIFT-ERROR is more left-skewed than that of MATCH and MATCH-ERROR. And the incorrect precursor increased the number of matched fragment masses by at most ???. Based on this results, if the two precursors of a multiplexed MS/MS spectrum are matched to the same protein and the PrSM of the second most abundant precursor contains >=3 (set the default setting for parameter *𝛾*  to 3) more matched fragment masses compared with the PrSM of the most abundant precursor, the PrSM of the most abundant precursor is treated as a precursor assignment error, the second abundant precursor is chosen as the primary precursor to fix the error (see Methods).

**Evaluation on pseudo-multiplexed MS/MS spectra**

To evaluate the performance of TopMPI for identifying multiplexed spectra, we generated an evaluation set of pseudo-multiplexed MS/MS spectra by combining two non-multiplexed MS/MS spectra identified from the first replicate of the *E. coli* data set. TopPIC (version 1.7.6) was used to search the MS/MS spectra against the UniProt *E. coli* K12 proteome database (version September 7, 2023; 4530 entries) and identified 1779 PrSM with an E-value cutoff of 0.01 (parameters settings detailed in Supplemental Table S5). As some of the identified MS/MS spectra were multiplexed ones, we filtered the identified spectra using the SCPFs observed in their isolation windows in MS1 spectra. For each of the 1779 MS/MS spectra, we ranked the SCPFs observed in its isolation window based on the sum of the peak intensities of the feature in the window. If the intensity sum of the top SCPF was less than 85% of the total peak intensity sum of all the SCPFs, the MS/MS spectrum was removed, resulting in 498 spectra. A pair of spectra in the 498 spectra is a matched pair if (1) the charge states of their top SCPFs are different, (2) the distance between the average *m/z* values of their SCPFs were no more than 1.5 m/z, and (3) their proteoform identifications were from two different proteins. A total of 725 matched spectrum pairs were found in the list of 498 spectra, and a pseudo-multiplexed MS/MS spectrum was generated for each matched spectrum pair by combining the fragment masses of two spectra with their original mass intensities. The SCPF of the spectra with more matched fragment masses among the pair was chosen as *F*1, while the SCPF of the other spectra was chosen as *F*2 (see Methods). Because TopPIC does not use mass intensities in database search-based proteoform identification, the mass intensities do not affect the results of database search.

We searched the 725 pseudo multiplexed spectra against the same UniProt *E.* coli proteome database (version September 7, 2023; 4530 entries) concatenated with a shuffled decoy database of the same size using TopPIC with *F*1, TopPIC with *F*2, and finally TopMPI using the same parameters, which are given in Supplemental Table S4.

With 1% spectrum-level FDR, TopPIC searched with *F*1 reported 724 PrSMs, out of which 722 PrSMs matched to the protein consistent with its SCPF, whereas the remaining 2 PrSMs matched to the protein of the incorrect SCPF. Similarly with 1% spectrum-level FDR, TopPIC searched with *F*2 reported 719 PrSMs, out of which 625 PrSMs matched to the protein consistent with its SCPF, 87 PrSMs matched to the protein of the incorrect SCPF, and the remaining 7 PrSMs matched to a protein not in the pair. Lastly, with 1% spectrum-level FDR filtered separately for the primary PrSMs and the secondary PrSMs, TopMPI reported PrSM pairs for 713 spectra and single PrSMs for 12 spectra. In this process, 14 spectra reassigned their primary SCPF from *F*1 to *F*2. Among the total 1438 single PrSMs reported, 1431 PrSMs matched to the protein consistent with its SCPF, 3 PrSMs matched to a protein of the incorrect SCPF, and the remaining 4 PrSMs matched to a protein not in the pair. It is worth noting that all 3 PrSMs matched to the protein of the incorrect SCPF came from the primary PrSMs, whereas all 4 PrSMs that matched to a protein not in pair were results of the secondary PrSMs. We then examined the E-value distribution of the original PrSMs that was taken to form their respective pseudo-multiplexed spectra, showing that TopMPI tends to miss PrSMs with low confidence identifications for the original non-multiplexed spectra (Fig. 4). How about errors in precursor assignment?

We further generated evaluation data sets of pseudo-multiplexed MS/MS spectra from the *E. coli* data set, in which the number of fragment masses from one proteoform was controlled. To generate a pseudo-multiplexed MS/MS spectrum, we first select a non-multiplexed experimental spectrum as the base spectrum and another non-multiplexed spectrum as the noise spectrum and then add some fragment masses of the noise spectrum to the based spectrum. A spectrum is a candidate noise spectrum of a base spectrum if (1) the distance between the SCPF average *m/z* values of the noise spectrum and the base spectrum is no more than 20 m/z, (2) the proteoform identifications of the base and noise spectra are from two different proteins, and (3) after removing from the noise spectrum fragment masses matched to the b- or y-ions in the base spectrum, the number of fragment masses in the noise spectrum is no less than 2 times of that in the base spectrum.

The 498 non-multiplexed spectra were used as base spectra to generate the pseudo multiplexed spectra. For 325 of the 498 spectra, we found one candidate noise spectrum without unexpected modifications and another candidate noise spectrum with an unexpected modification. The 325 spectra and their matched noise spectra were used to generated 650 spectral pairs: 325 pairs for a base spectrum and a noise spectrum without unexpected modifications and 325 pairs for a base spectrum and a noise spectrum with an unexpected modification. The 650 pairs were divided into four groups: group 1: both base and noise spectra have no unexpected modifications, group 2: only the noise spectrum has an unexpected modification; group 3: only the base spectrum has an unexpected modification, group 4: both base and noise spectra have unexpected modifications.

The 650 spectrum pairs were used to generate 10 pseudo multiplexed spectrum data sets, each of which corresponds to a noise ratio (*r* = 20%, 40%, …, 200%). The data set with noise ratio *r* contained 650 spectra, each of which contains all masses in the base spectrum and randomly selected round(*rn*) fragment masses in the noise spectrum, where *n* is the number of masses in the base spectrum and round() is the rounding function. TopPIC (version 1.7.6 parameter settings in Supplemental Table S5) was used to search the 11 MS data sets against the UniProt *E. coli* proteome database (version September 7, 2023; 4530 entries). Note that a mass shift ±20075 Da was allowed as 20059.86 Da was the largest difference between the precursor masses of a base spectrum and its corresponding noise spectrum.

We plotted the change in protein identification accuracy against the percentage of noise mass fragments added to the base spectrum (Fig. 5). When the noise ratio in the noised pseudo multiplexed spectra increases, the identification accuracy of the noise protein of the spectra in group 3 increases faster than that in group 1. Similarly, the identification accuracy of the noise protein of the spectra in group 4 increases faster than that in group 2. As expected, the unexpected modification in the base spectrum in groups 3 and 4 reduces the protein identification accuracy in database search and increases the influence of the noise fragment masses in the pseudo multiplexed spectrum as the percentage of PrSMs identified as the protein of the noise spectrum increased. The percentage of noise protein identification of group 1 increases faster than group 2 as the noise ratio increases, showing that a multiplexed spectrum with a noise spectrum whose PrSM does not have unexpected modifications is more likely to be incorrectly identified compared with a noise spectrum whose PrSM have an unexpected modification. A similar conclusion can be drawn comparing group 3 and group 4.

**Comparison of TopMPI and TopPIC**

We compared spectral identifications of TopMPI and TopPIC on the first replicate of a CZE-MS/MS data set of yeast proteins (Pride ID: ????) reported by ??? [19]. The data file contained ??? MS/MS spectra with an isolation window of ???. TopFD and TopPIC (version 1.7.6) were employed to analyze the data set against the UniProt yeast proteome database (version March 3, 2023; 6,727 entries), with parameter settings detailed in Supplemental Tables S1 and S6. With a 1% spectrum-level FDR, TopPIC reported 4,544 PrSMs at 1% spectrum-level FDR and ??? proteoform at 1% proteoform-level FDR.

Among these, 123 SCPFs were identified as contradictory, where the associated PrSMs were assigned to different proteins. By applying TopMPI, we reassigned the primary precursor for 114 MS/MS spectra, resulting in an increased identification of 4,563 PrSMs and 2,128 unique SCPFs, while reducing the number of contradictory SCPFs to 87. When identifications from a second round of analysis were incorporated, the total number of PrSMs increased to 5,714, corresponding to 2,484 unique SCPFs, of which 126 were identified as contradictory. However, this result suggests that the second-round identifications under a 1% FDR still include inaccuracies.

**Reproducibility in Demultiplexed Proteoform Identification**

**TODO**

**Comparison of two FDR Computation Methods**

As two PrSMs may be reported from a multiplexed MS/MS spectrum, to accurately assess the false discovery rate (FDR) when using the target-decoy approach as two single PrSMs can be reported per spectrum, we compared two different FDR computation methods using three different datasets, the pseudo-multiplexed spectra dataset, the *E. coli* dataset, and then the ovarian cancer dataset. The first method is to compute FDR on a single-PrSM level whereas the second computational method will compute FDR separately for the first reported PrSM and the second reported PrSM. Both *E. coli* datasets were searched against the UniProt *E. coli* proteome database concatenated with a shuffled decoy database of the same size and so was the ovarian cancer dataset against the UniProt human proteome database.

In Table 1 we present the comparative results of the two FDR computation methods. We can observe that by separating the FDR computation between the first and second reported PrSM per spectrum, we would be able to report more identifications than doing the FDR computations combined for both real-world datasets. As for the pseudo-multiplexed dataset, since all 1095 spectra were able to report a targeted identification as their first reported PrSM providing an FDR rate of 0, it would then lead to an overestimation of the FDR if we were to combine the first reported PrSMs with the second reported ones, resulting in reporting PrSMs with terrible identification scores such as 579.82. Therefore, it is evident that the first reported PrSMs should be separated from the second reported PrSM in FDR computations.

An E-value recalculation method was also proposed to exclude the matched fragments of the second reported PrSM when calculating the E-value for the first reported PrSM, as it would only be fair since the matched fragments of the first reported PrSM were removed when trying to identify a second PrSM. Four different variations of this method were tested, and we would remove the matched fragments of the second reported PrSM from the spectra of the first reported PrSM on the following sets of second reported PrSMs:

* Method 0: Null set
* Method 1: The set of second reported PrSMs with a filter of 1% FDR
* Method 2: The set of second reported PrSMs with an E-value cutoff of 0.01
* Method 3: The set of second reported PrSMs with no cutoff

Method 0 will act as the baseline method, and the four methods were tested on the *E. coli* dataset against the UniProt proteome database concatenated with a shuffled decoy database of the same size. All reported PrSMs were filtered with 1% FDR.

Table 2 illustrates the results of the recalculation. As the results have suggested, the recalculation of E-values by removing the matched peaks of the second reported PrSM does not improve the score of target identifications better than decoy ones, in fact, it degrade the performance of E-values. Therefore it’s better to not conduct the recalculation.

**Identification Results Compared with Vanilla TopPIC**

To evaluate the performance of MSDeplex against the standard TopPIC suite pipeline, all three *E. coli* replicate datasets were searched against the UniProt *E. coli* database for proteoform identification using the proposed demultiplexing method and were compared against the list of identifications obtained using TopPIC suite (v1.7.6).

Using the target-decoy approach concatenated with a shuffled decoy database along with a 1% proteoform level FDR cutoff, we see that MSDeplex was able to increase the number of proteoform identifications by 27% and protein identifications by 7% on average.

**Larger Datasets**

To evaluate the performance of demultiplexing on larger and more complex datasets containing thousands of proteoforms, an SEC-RPLC-CZE-MS/MS SW480 cells dataset with triplicates containing a total of 52 fractions was introduced and analyzed using the pipeline [].

**References**

1. Catherman, A.D., O.S. Skinner, and N.L. Kelleher, *Top Down proteomics: facts and perspectives.* Biochem Biophys Res Commun, 2014. **445**(4): p. 683-93.

2. Roberts, D.S., et al., *Top-down proteomics.* Nat Rev Methods Primers, 2024. **4**(1).

3. Gregorich, Z.R. and Y. Ge, *Top-down proteomics in health and disease: challenges and opportunities.* Proteomics, 2014. **14**(10): p. 1195-210.

4. Basharat, A.R., et al., *TopDIA: A Software Tool for Top-Down Data-Independent Acquisition Proteomics.* Journal of Proteome Research, 2024.

5. Tsou, C.-C., et al., *DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics.* Nature Methods, 2015. **12**(3): p. 258-264.

6. Demichev, V., et al., *DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput.* Nat Methods, 2020. **17**(1): p. 41-44.

7. Wang, J., P.E. Bourne, and N. Bandeira, *MixGF: spectral probabilities for mixture spectra from more than one peptide.* Mol Cell Proteomics, 2014. **13**(12): p. 3688-97.

8. Zhang, W., et al., *ChimST: An Efficient Spectral Library Search Tool for Peptide Identification from Chimeric Spectra in Data-Dependent Acquisition.* IEEE/ACM Trans Comput Biol Bioinform, 2021. **18**(4): p. 1416-1425.

9. Sinitcyn, P., et al., *MaxDIA enables library-based and library-free data-independent acquisition proteomics.* Nat Biotechnol, 2021. **39**(12): p. 1563-1573.

10. Dorfer, V., et al., *CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction.* J Proteome Res, 2018. **17**(8): p. 2581-2589.

11. Sun, C., et al., *A Full Window Data Independent Acquisition Method for Deeper Top-down Proteomics.* bioRxiv, 2024: p. 2024.11.08.622616.

12. Park, J., et al., *Informed-Proteomics: open-source software package for top-down proteomics.* Nat. Methods, 2017. **14**: p. 909.

13. Zamdborg, L., et al., *ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry.* Nucleic Acids Res., 2007. **35**: p. W701.

14. Kou, Q., L. Xun, and X. Liu, *TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization.* Bioinformatics, 2016. **32**(22): p. 3495-3497.

15. Kou, Q., et al., *A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra.* Bioinformatics, 2017. **33**(9): p. 1309-1316.

16. Kessner, D., et al., *ProteoWizard: open source software for rapid proteomics tools development.* Bioinformatics, 2008. **24**(21): p. 2534-6.

17. Basharat, A.R., et al., *TopFD: A Proteoform Feature Detection Tool for Top-Down Proteomics.* Anal Chem, 2023. **95**(21): p. 8189-8196.

18. Jeong, K., et al., *Precursor deconvolution error estimation: The missing puzzle piece in false discovery rate in top-down proteomics.* Proteomics, 2024. **24**(3-4): p. e2300068.

19. !!! INVALID CITATION !!! 17.