Published in final edited form as:

Proteomics. 2024 April ; 24(8): e2300234. doi:10.1002/pmic.202300234.

Spectral averaging with outlier rejection algorithms to increase identifications in top-down proteomics

Austin V. Carr,
Nicholas E. Bollis,
John G. Pavek,
Michael R. Shortreed,
Lloyd M. Smith
Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin, USA

Abstract

The identification of proteoforms by top-down proteomics requires both high quality fragmentation spectra and the neutral mass of the proteoform from which the fragments derive. Intact proteoform spectra can be highly complex and may include multiple overlapping proteoforms, as well as many isotopic peaks and charge states. The resulting lower signal-to-noise ratios for intact proteins complicates downstream analyses such as deconvolution. Averaging multiple scans is a common way to improve signal-to-noise, but mass spectrometry data contains artifacts unique to it that can degrade the quality of an averaged spectra. To overcome these limitations and increase signal-to-noise, we have implemented outlier rejection algorithms to remove outlier measurements efficiently and robustly in a set of MS1 scans prior to averaging. We have implemented averaging with rejection algorithms in the open-source, freely available, proteomics search engine MetaMorpheus. Herein, we report the application of the averaging with rejection algorithms to direct injection and online liquid chromatography mass spectrometry data. Averaging with rejection algorithms demonstrated a 45% increase in the number of proteoforms detected in Jurkat T cell lysate. We show that the increase is due to improved spectral quality, particularly in regions surrounding isotopic envelopes.

Keywords

averaging; outlier rejection; proteoform; proteomics; top-down

Correspondence: Lloyd M. Smith, 1101 University Ave, Madison, WI 53706, USA. smith@chem.wisc.edu.
Austin V. Carr and Nicholas E. Bollis contributed equally to this work.

CONFLICT OF INTEREST STATEMENT
The authors declare no conflicts of interest.

SUPPORTING INFORMATION
Additional supporting information may be found online https://doi.org/10.1002/pmic.202300234 in the Supporting Information section at the end of the article.

1 | INTRODUCTION

The identification and quantification of intact proteoforms in complex mixtures such as mammalian cell lysate is a challenging frontier of modern proteomics [1, 2]. It is accomplished by “top-down” proteomics, wherein intact proteoforms are ionized and analyzed by tandem mass spectrometry [3]. Two successive scans are executed, an initial precursor mass scan (MS1) to reveal the ion species that are present, followed by a secondary scan of fragment ions (MS2) generated in a gas phase dissociation process. The MS1 analysis yields the mass of the intact proteoform through deconvolution, while the MS2 fragmentation and analysis yields amino acid sequence information and post-translational modification (PTM) identification and localization.

Proteoform mass spectra can be highly complex. Large proteins are present as many isotopologues [4], such that an intact proteoform is observed not as a single peak but rather as a distribution of peaks referred to as the isotopic envelope. In addition to this isotopic complexity, the process of electrospray ionization (ESI) generates a charge state envelope. Splitting the signal into isotopic and charge state envelopes can severely reduce the abundance of individual peaks, making them difficult to distinguish from noise [4].

Making an identification with top-down proteomics requires successful feature detection: that is, the grouping of all isotopic peaks from each charge state of a single proteoform species (the peak set, or feature), and deconvolution to determine a species’ neutral mass [5]. The neutral mass of the ion is used to filter a list of candidate fragmentation spectra. Tools commonly used to generate deconvoluted masses for top-down spectra include ProSight [6], TopPIC Suite [7], pTop2 [8], Bruker DataAnalysis, and MetaMorpheus [9]. Of these programs, only ProSight, TopPIC Suite, and MetaMorpheus include both a built-in database search and ongoing software support. These three tools each use an approach similar to that of the THRASH deconvolution algorithm developed by Senko et al. [10] that requires resolved isotopic envelopes to predict charge state and monoisotopic mass. Deviations from theoretical isotopic distributions can cause miscalculations of charge state [11], leading to incorrect monoisotopic masses, and missed or false identifications during subsequent searching steps.

Due to the reliance of top-down proteomics search engine software on fully resolved isotopic envelopes, noise and artifacts can significantly influence top-down proteomics search results. Artifactual peaks can be generated by aberrant fragmentation events, background atmospheric ions introduced during electrospray ionization [12], or signal processing [13], causing errors in charge state assignment. Scan averaging is a widely used method to improve signal-to-noise ratios and has been found to drastically impact the sensitivity of mass deconvolution [14, 15]. However, achieving high signal-to-noise ratio improvements with averaging requires input values free from contamination by outlier measurements. Such outliers are often artifacts and their inclusion in the averaging can degrade the final, averaged spectrum [16].

We sought to identify and include a pre-processing method in the open-source proteomics search engine MetaMorpheus that could simultaneously improve signal-to-noise ratio of low abundance peaks, improve isotopic envelope distributions, and remove artifacts. Averaging with rejection algorithms developed by researchers in astronomy [16], enable the simultaneous improvement of signal-to-noise ratios and the removal of artifactual data. Astronomical measurements are commonly contaminated with high intensity but transient artifacts from cosmic rays, airplanes, satellites, and other light sources, analogous to contamination by chemical noise or signal processing artifacts in mass spectrometry data. We have implemented averaging with rejection algorithms for MS1 spectra, adapted from the astronomical image processing software PixInsight (https://pixinsight.com), to improve counts of protein spectral matches (PrSMs) and proteoform identifications in top-down proteomics experiments.

2 | METHODS

2.1 | Direct injection data

Isolated protein standard of and Somatotropin (Sigma Aldrich) were reconstituted in 50:50 ACN:H_20 to a concentration of 50 μM. Samples were injected by a syringe pump with a flow rate of 2 μL min^{-1} into the ESI source and scans were collected for ten minutes with a top five most intense data dependent acquisition method on a QE-HF Orbitrap (Thermo Scientific). MS1 spectra were collected with an orbitrap resolution of 120,000 (at m/z 400), maximum injection time of 100 ms, and automatic gain control of 3×10^6. MS2 spectra were collected with an orbitrap resolution of 30,000 (at m/z 400), a maximum injection time of 50 ms, and automatic gain control of 1×10^6.

2.2 | Jurkat T lymphocyte cell lysate top-down dataset

This dataset was collected in 2020 to support a study of integrated top-down and bottom-up proteomics (MassIVE identifier: MSV000083768) [17, 18]. The samples were reduced, but not alkylated, before size fractionation by GELFrEE (Expedion) into ten fractions. Fractions were then analyzed by HPLC-ESI-MS/MS (nanoAcquity, Waters and QE-HF Orbitrap, Thermo Scientific). MS1 spectra were collected with three microscans at an Orbitrap resolution of 240,000 (at m/z 400) maximum injection time of 50 ms, and automatic gain control of1×10^6. The top eight most abundant precursors were selected for fragmentation by high energy collisional dissociation with a normalized collisional energy of 25 and an isolation width of 4 m/z. Each fraction was then calibrated with MetaMorpheus to account for instrumental drift and systematic errors. The eight calibrated mzML files will be referred to as the control dataset. Calibration was performed with the protease set to top-down, a precursor and product mass tolerance of 10-ppm, and methionine oxidation as a variable modification. The control dataset was then averaged with the experimentally optimized parameters and will be referred to as the averaged dataset.

2.3 | Neucode labeled proteins

Proteins extracted from a biopsied prostate tumor sample were extracted, reduced, and split into two portions. One portion was labeled with a $^{13}C_6$ $^{15}N_3$ isotopologue (NeuCode light) of a cysteine-reactive label, and the other was labeled with a D_9 isotopologue (NeuCode heavy) of the same cysteine-reactive label. These two isotopologues, the NeuCode light and NeuCode heavy labels, form a NeuCode pair with a mass difference between them of

0.0453 Da. The NeuCode labeled proteins were then recombined in a 2:1 ratio (heavy:light). Combining proteins labeled with these two isotopologues produces overlapping isotope distributions for otherwise identical cysteine-containing proteoforms (see Figure 5A for an example). The combined sample was then separated into 8 size-based fractions by PEPPI fractionation. All fractions were analyzed via HPLC-ESI-MS (Thermo Easy-nLC-Orbitrap Fusion Lumos). MS1 scans from 350 to 2000 m/z were collected in profile mode with 5 microscans, an AGC target of 800,000, and a maximum injection time of 50 ms.

Cysteine-containing proteoforms in these samples will exhibit overlapping isotope distributions. To discover all NeuCode pairs in the sample, the files are first deconvoluted in Thermo Protein Deconvolution 4.0, producing a list of masses. Each deconvoluted mass is processed as follows: averaging of surrounding MS spectra to improve S/N and produce more accurate isotope ratios, determining of all existing charge states for that mass, and then pairing of each isotope peak with its adjacent peaks to produce a set of possible experimental cysteine counts. Adjacent peaks are accepted as possible NeuCode pair peaks by intensity ratio filtering based on the expected 2:1 mixing ratio of NeuCode heavy to NeuCode light. Accepted possible NeuCode pair peaks are aggregated by rounded cysteine count, and the experimental cysteine count for the mass is assigned as the most frequently occurring cysteine count among accepted possible NeuCode pair peaks.

2.4 | Spectral averaging algorithm

The averaging with rejection task (Averaging Task) in MetaMorpheus consists of four steps: normalization, binning, outlier rejection, and averaging. Each intensity value in a spectrum is first normalized to the spectrum's total ion current. A binned m/z axis is generated using a predefined bin width. Next, for each collection of spectra to be averaged together, and for each intensity value in an m/z bin, the outliers within each m/z bin are determined and rejected by one of the outlier rejection methods incorporated in the Averaging Task. Finally, the remaining (unrejected) intensity values in each bin are averaged to produce the final, averaged spectrum.

Outlier rejection algorithms were adapted from pseudocode in the documentation of the astronomical image processing software PixInsight (https://pixinsight.com/doc/tools/ImageIntegration/ImageIntegration.html). The full details of each outlier rejection algorithm are presented in the Supplemental Information. For a given set of intensity values corresponding to an m/z measurement in a set of N consecutive MS1 mass spectra (an *intensity bin* at a given *m/z*), iterative rejection algorithms remove values outside $n\sigma_{\min}$-median (*intensity bin*) and median (*intensity bin*) $+n\sigma_{\max}$, where n is defined by user input, and σ is an estimate of the standard deviation of the intensity values in the *intensity bin*. During each iteration, σ and the median are calculated based on the non-rejected values of the previous iteration. The iterations continue until no more values are removed, or there are no more intensity values within the stack that fall outside the defined range.

Two spectral averaging algorithms are also incorporated in MetaMorpheus as the Averaging Task: one algorithm is used for data-dependent acquisition LC-MS, and the other for direct injection experiments. For LC-MS data, the user provides a number of scans to be averaged together (default = 5 scans), and each averaged scan in the output data file is a moving average of the original MS1 scans. For direct injection experiments, we have included options to average all spectra together, or, analogously to the options for LC-MS, to average every *N* spectra together. The output of the Averaging Task in MetaMorpheus is an mzML file that can be used in subsequent MetaMorpheus tasks or other software. Five other averaging algorithms are available as standalone code within mzLib (https://github.com/smith-chem-wisc/mzLib) as a convenience to users interested in exploring the full range of outlier rejection options.

2.5 | MetaMorpheus data processing workflow

MetaMorpheus was used to perform Calibration, Averaging, Global Post-Translational Modification discovery (G-PTM-D) [9, 19] and Search for both the control and averaged datasets as outlined in Figure 1. Task files containing the settings used for searches can be found in Supplemental Files S1–S4. The full analysis of parameters for each dataset can be found in Supporting Information but are briefly summarized here. The G-PTM-D and Search tasks were performed with the protease set to top-down, oxidation on M as a variable modification, and a 10-ppm mass tolerance for both the MS1 and MS2 spectra.

Additional PTMs set in the G-PTM-D task included a subset of common artifacts (deamidation and ammonia loss) and a subset of common biological modifications (phosphorylation, methylation, pyroglutamylation, palmitoylation, simple glycosylation) [19], with a full list available in Supporting Information. In the Search Task, options were selected for generating complementary ions and using internal fragment ions to improve post-translational modification localizations [20]. All searches were exact mass searches with a precursor and product mass tolerance of 10 ppm, a maximum assumed charge state of 80, and oxidation on methionine as a variable modification. For comparisons between workflows, each search was run separately for the data set, for example, control datasets were analyzed separately from the averaged datasets. All sets of PrSMs were filtered to 1% false discovery rate (FDR), a 0.01 *Q*-value [21], and an ambiguity level of 1 [22]. The Supplementary Workbooks containing the results of the MetaMorpheus searches can be found in the MassIVE repository associated with this analysis, MSV000092054.

2.6 | Deconvolution data processing

The control and averaged datasets were first centroided with MsConvert (v 3.0.22189) [23]. Both centroided datasets were deconvoluted with TopFD (v1.6.2) [24] and FLASHDeconv (OpenMS-3.0.0) [25] with a minimum charge of 1, maximum charge of 100, maximum mass of 100000 Da, and precursor window size of 4 m/z. The option to output TopFD result file types was selected within FLASHDeconv. Deconvolution software comparisons were made using the resulting ms1.align and .feature files were used for comparative analysis between averaged and control datasets.

3 | RESULTS

Averaging Task Parameter Screening.

While use cases for each of the averaging with outlier with rejection algorithms are well-known in the astronomy community [26], reasonable default parameters were unclear for mass spectrometry data. To determine a default algorithm and settings to be used in MetaMorpheus, averaging on data from standard protein, direct injection data sets was performed using each of the 7 averaging methods tested. We evaluated the effect of over 9000 unique combinations of parameters and 7 different averaging with rejection algorithms. The parameters included a bin size ranging from 0.001 Th to 1 Th, number of MS1 scans averaged, choice of outlier rejection algorithm or no outlier rejection, and outlier detection settings. The parameters varied and the possible values used for each parameter are shown in Table S1.

Before the averaging step in astronomical imaging software, the images are “aligned” such that each pixel corresponds to the same point in space across repeated measurements. This is possible due to the static nature of imaging large objects at long distances over relatively short time scales. However, measurements in mass spectrometry are dynamic, with peaks that can shift from scan to scan. Therefore, intensity bins are created for each m/z value along a common axis to achieve a similar alignment procedure. The effect of bin size on a variety of factors was determined using a range of bin sizes from 0.001 Th to 1 Th (Figure 2A–C). To determine how bin size affects the averaged spectra, the theoretical most abundant isotopic peak was calculated based upon the elemental composition of the proteoform for each charge state. For each spectrum, the most abundant peak in each charge state was counted if it was above 5% relative intensity and within a 10-ppm error of the theoretical most abundant peak. Figure 2A shows that the number of most abundant peaks within 10-ppm of their theoretical m/z is maximized at a bin size of 0.01 Th. The ppm error between theoretical and experimental isotopic peaks is also minimized when using a bin width of 0.01 Th (Figure 2B). Figure 2C provides further supporting evidence for selecting an 0.01 Th bin size as a reasonable default by demonstrating more isotopic envelopes are charge-state resolvable, or successfully deconvoluted to the correct charge state, per averaged spectrum than the other bin sizes tested and the unaveraged spectra. These results are reasonable because a large bin size is more likely to average multiple peaks together whereas a smaller bin size will provide greater mass accuracy but at the cost of increasing the likelihood of splitting peaks between multiple bins as shown by the large distribution in scores for each metric for a bin size of 0.001 Th in Figure 2A–C.

After the proper bin size has been determined, the next step is to optimize the outlier rejection process, which is depicted in Figure 2F–I for a single isotopic envelope extracted from five consecutive MS1 spectra (Figure 2F). In Figure 2H, the outlier rejection procedure is illustrated for three distinct peaks, highlighting two instances (blue and green) where lower abundance peaks are rejected. Excluding the lower abundance peaks increases the averaged intensity of the green and blue peaks resulting in the isotopic distribution more closely matching the expected theoretical isotopic distribution than averaging without outlier rejection. The peak highlighted in purple demonstrates a scenario wherein one low abundance peak and one high abundance peak are rejected, with the resultant averaged peak remaining consistent whether outlier rejection is applied or not. These examples illustrate the utility and effectiveness of outlier rejection.

The greatest improvement was obtained when using the sigma clipping and averaged sigma clipping iterative rejection algorithms. The effects of these algorithms were further studied by varying the acceptable deviation both above and below the median before an intensity value is rejected as an outlier. The effects of unique combinations of min/max σ values upon the ppm error of theoretical most abundant isotopic peaks is shown in Figure 2D,E. Rejecting values that deviated slightly below the median (0.5σ) and values that deviated dramatically above the median (3.5σ) yielded the greatest improvement suggesting that in regions of reproducible signal peak density, averaging with outlier rejection can remove low abundance noise and artifact peaks.

The results of the parameter assessment suggest using a bin width of 0.01 Th, the averaged sigma clipping rejection method, a $\sigma_{\min}$ of 0.5, and a $\sigma_{\max}$ of 3.5. The default value chosen for N, the number of consecutive scans to be averaged, differs for direct injection compared to LC-MS. The direct injection experiments suggest reasonable defaults to be used for online LC-MS experiments. We selected the sigma clipping outlier rejection algorithm to be used as the default for LC-MS experiments because of its similar performance to the optimal results in the direct injection experiment. The bin size of 0.01 Th and $\sigma_{\min}$ and $\sigma_{\max}$ values were also used as the default LC-MS settings. In MetaMorpheus, defaults are implemented for direct injection and LC-MS experiments, while also allowing users to set a custom $\sigma_{\min}$, $\sigma_{\max}$, bin size, or outlier rejection algorithm if desired.

3.1 | Analysis of complex mixtures

We compared top-down search results of a GELFrEE fractionated sample with three processing methods: calibration alone and calibration followed by averaging with or without a rejection algorithm. The full PrSM and proteoform search results from each condition can be found in Supplemental Workbooks 1–6. MetaMorpheus proteoform identifications correspond to the highest scoring PrSM for each proteoform. A similar number of unique primary sequences were found across all fractions regardless of the pre-processing method used (Figure 3A). Utilizing averaging with no rejection decreased the number of PrSMs per fraction compared to calibration only, while the number of PrSMs obtained by averaging with rejection was greatly increased (Figure 3B). Across the analyzed fractions, averaging with rejection outperformed simple averaging (Figure 3C). Cumulatively, averaging with rejection increased the number of identified proteoforms by 45% compared to both calibration only and averaging with no rejection (Figure 3D).

The steep decline in the number of identifications in the fractions containing the higher mass proteins (Fractions 9 and 10) is indicative of the upper mass limit inherent to deconvolution algorithms requiring isotopic resolution to calculate monoisotopic mass. Because the resolution required to resolve isotopic peaks of high mass proteins quickly exceeds the instrument’s capabilities, the isotopic peaks are unable to be used to determine charge state and monoisotopic mass for large proteins. A complete analysis of the molecular weight of proteoform spectral matches by fraction can be found in Figure S2. Figure S3 demonstrates that averaging with rejection enhances sensitivity and discriminative power, leading to the identification of previously undetected proteoforms. This gain in identification results primarily arises from the salvaging of ions that were previously unidentifiable in chimeric spectra, where multiple precursors coexist within the isolation window, and is explored in Figures S4 and S5. The distributions of MetaMorpheus scores, which are the quality scores for how well the peaks in an MS2 scan match with a theoretical fragmentation spectrum by number of matched fragment ions, are shown in Figure 3E for each of the processing conditions. The score distributions of unique proteoform identifications at 1% FDR are similar between the calibrated only and the averaging with rejection processing conditions. Proteoforms at 1% FDR had a score of 25 or higher. Averaging without rejection incorporates outliers into the averaged MS1 spectrum resulting in deconvolution errors. Errors in the deconvoluted mass and/or charge prompt the search algorithm to attempt matching the experimental spectrum to database proteins of an incorrect mass, rather than to proteins with a mass corresponding to the selected species for fragmentation. Consequently, this increases the likelihood of identifying a decoy proteoform and decreases the average score of proteoform identifications. These combined effects result in lower scores and fewer identifications at 1% FDR (Figure 3E, blue).

3.2 | Characterizing improvements in deconvolution

The effects of averaging with and without outlier rejection on the deconvolution process were tested by analyzing the counts and quality of deconvoluted features by TopFD and FLASHDeconv. The number of masses reported in the MS1 Align deconvolution output increased by 29% and 149% for FLASHDeconv and TopFD, respectively (Figure 4A). The most dramatic increase was found within the final two fractions, which contained the largest proteins. As both algorithms rely upon accurate isotopic spacing, and isotopic resolution decreases as a function of increasing mass, these results suggest that averaging spectra may improve the detection of isotopically resolved peaks. The total number of features detected remained constant for TopFD and slightly increased for FLASHDeconv (Figure 4B). The detected features were assessed using a method described by Jeong et al. to determine if a feature belonged to a class of artifacts including high harmonic masses, low harmonic masses, and isotopologues. For both deconvolution software and averaging conditions, the count of artifactual features remained constant (Figure 4C).

Isotopic peak spacing was evaluated for all fractions using the additional output files that FLASHDeconv exclusively provides. Plotting the number of peaks found per identified neutral mass shows the increases in the average number of peaks per deconvoluted mass, with significant gains in the most spectrally complex fractions (fractions 6–8) (Figure 4D). FLASHDeconv also reports signal-to-noise measurements for the mass and charge of each deconvoluted feature. The signal-to-noise measurement used in FLASHDeconv is defined by Jeong et al. [25]. The charge state signal-to-noise ratio is the proportion of signal belonging to a single cluster of isotopic peaks, and the mass signal-to-noise ratio is defined as the proportion of signal belonging to all charge states corresponding to a single, deconvoluted mass. Both the charge and mass signal-to-noise ratio for deconvoluted masses at 1% FDR increased by 100% and 470% after averaging without outlier rejection and with outlier rejection, respectively (Figure 4E). These results provide additional evidence that averaging with outlier rejection increases the similarity of experimental isotopic envelopes to that of theoretical isotopic distributions. Taken together, the data presented in Figure 4 suggests that the Averaging Task, with or without outlier rejection, does not increase the number of deconvoluted features. Instead, the evidence suggests that the quality of the deconvoluted features is improved. Notably, this enhancement allows for the resolution of a greater number of isotopic peaks per feature. Additionally, the implementation of outlier rejection results in a remarkable improvement in both the charge and mass signal-to-noise ratios as measured by FLASHDeconv. These enhancements collectively contribute to a substantial increase in top-down proteoform identifications, underscoring the critical role of outlier rejection in the spectral averaging algorithm.

3.3 | Applying averaging with rejection to isotopically labeled proteoforms

The improvement in the performance of deconvolution algorithms relying on isotopic resolution suggested that Averaging with Rejection could be applied to isotopically labelled proteins. Isotopic labeling of proteoforms using NeutronEncoding (NeuCode) can be used to count amino acids, providing additional information to aid in proteoform identification [27]. In this section, an application of averaging with rejection for proteins labeled with a NeuCode alkylating reagent to enable cysteine counting is detailed. Light and heavy NeuCode labeled proteoforms will elute together, at the same nominal mass, but with a mass difference in the tens of millidaltons, requiring high resolving power to differentiate two overlapping, distinct, isotopic envelopes, as shown in Figure 5A. Applying averaging with outlier rejection doubled the number of identified charge states and quadrupled the number of accepted NeuCode pairs while only modestly increasing the estimated false discovery rate of the NeuCode pairs (Figure 5B).

The observed improvement in NeuCode pair discovery with the application of averaging with outlier rejection likely stems from the improved accuracy of isotope ratios observed in averaged spectra. An important filter in the differentiation of real and false NeuCode pair isotopic envelopes is the intensity ratio between the NeuCode heavy and NeuCode light labeled species, which should reflect the mixing ratio (typically 2:1). The intensity ratio between the heavy and light distributions in the unaveraged spectra of Figure 5A varies across the distribution and does not match the expected 2:1 ratio. The intensity ratio between the heavy and light distributions in the averaged spectra is closer to the expected value of 2:1 than in the unaveraged spectra.

4 | CONCLUSIONS

We have described the implementation of averaging with outlier with rejection algorithms in MetaMorpheus and demonstrated a broad utility for top-down and intact mass proteomics applications demonstrating a substantial increase in the number of proteoforms identified in discovery mode top-down proteomics, direct injection, and intact mass proteoform identification with NeuCode. Analyzing complex samples from a previously published experimental dataset allowed us to evaluate the improvements obtained by using MetaMorpheus top-down proteomics search with the newly implemented Averaging Task.

Notably, averaging with rejection increased the number of identified proteoforms by 45% compared to either calibration only or averaging with no rejection. The analysis presented here yielded 2789 proteoforms at 1% FDR, a dramatic increase compared to the 711 proteoforms identified in the original study [28]. The refinement of statistical models to validate chimeric identifications, particularly those revealed by signal averaging, is an important area for future research in top-down proteomics. We posit that improved signal averaging could also increase the quantitative precision of label-free data, especially in challenging scenarios like single-cell proteomics, where low signal-to-noise ratios pose a persistent challenge. Additionally, the application of improved signal averaging to MS2 spectra originating from the same proteoform species holds promise for constructing more robust spectral libraries. This suggests a broader applicability of improved averaging techniques beyond mere identification, extending its potential impact to multiple aspects of mass spectrometry-based proteomics methodologies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding information

National Institute of General Medical Sciences (NIGMS); National Heart, Lung, and Blood Institute, Grant/Award Number: R01HL149966; National Cancer Institute, Grant/Award Number: R01CA193481; National Institute of General Medical Sciences, Grant/Award Numbers: R35GM126914, T32GM135066

DATA AVAILABILITY STATEMENT

The raw MS data files for the direct injection data set have been deposited to ProteomeXchange with the provisionary dataset identifier 1-202230530-116057. The top-down data set was accessed via MassIVE using the identifier MSV000083768. The files used in this analysis and their associated MetaMorpheus output were deposited to MassIVE under the identifier MSV000092054.

Abbreviations:

FDR	false discovery rate
G-PTM-D	global post-translational modification discovery
LC-MS	liquid chromatography mass spectrometry
MS	mass spectrometry
MS1	precursor mass spectra
MS2	fragmentation mass spectra
PrSM	protein spectral match
PTM	post-translational modification

REFERENCES

1. Melby JA, Roberts DS, Larson EJ, Brown KA, Bayne EF, Jin S, & Ge Y (2021). Novel strategies to address the challenges in top-down proteomics. Journal of the American Society for Mass Spectrometry, 32(6), 1278–1294. 10.1021/jasms.1c00099 [PubMed: 33983025]
2. Zhou H, Ning Z, Starr E, Abu Farha M, & Figeys D (2012). Advancements in top-down proteomics. Analytical Chemistry, 84(2), 720–734. 10.1021/ac202882y [PubMed: 22047528]
3. Kelleher NL, Lin HY, Valaskovic GA, Aaserud DJ, Fridriksson EK, & McLafferty FW (1999). Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. Journal of the American Chemical Society, 121(4), 806–812. 10.1021/ja973655h
4. Compton PD, Zamdborg L, Thomas PM, & Kelleher NL (2011). On the scalability and requirements of whole protein mass spectrometry. Analytical Chemistry, 83(17), 6868–6874. 10.1021/ac2010795 [PubMed: 21744800]
5. Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, & Pevzner PA (2012). Protein identification using top-down spectra. Molecular & Cellular Proteomics, 11(6), M111.008524. 10.1074/mcp.M111.008524
6. Zamdborg L, LeDuc RD, Glowacz KJ, Kim Y-B, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, & Kelleher NL (2007). ProSight PTM 2.0: Improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Research, 35(suppl_2), W701–W706. [PubMed: 17586823]
7. Kou Q, Xun L, & Liu X (2016). TopPIC: A software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics, 32(22), 3495–3497. 10.1093/bioinformatics/btw398 [PubMed: 27423895]
8. Sun RX, Luo L, Wu L, Wang RM, Zeng WF, Chi H, Liu C, & He SM (2016). pTop 1.0: A high-accuracy and high-efficiency search engine for intact protein identification. Analytical Chemistry, 88(6), 3082–3090. 10.1021/acs.analchem.5b03963 [PubMed: 26844380]
9. Solntsev SK, Shortreed MR, Frey BL, & Smith LM (2018). Enhanced global post-translational modification discovery with MetaMorpheus. Journal of Proteome Research, 17(5), 1844–1851. 10.1021/acs.jproteome.7b00873 [PubMed: 29578715]
10. Horn DM, Zubarev RA, & McLafferty FW (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. Journal of the American Society for Mass Spectrometry, 11(4), 320–332. 10.1016/S1044-0305(99)00157-9 [PubMed: 10757168]
11. Kaur P, & O’Connor PB (2006). Algorithms for automatic interpretation of high resolution mass spectra. Journal of the American Society for Mass Spectrometry, 17(3), 459–468. 10.1016/j.jasms.2005.11.024 [PubMed: 16464606]
12. Bromirski AKM (2018). Selecting the best Q Exactive Orbitrap mass spectrometer scan mode for your application (Vol. White Paper). Thermo Scientific.
13. Mathur R, & O’Connor PB (2009). Artifacts in Fourier transform mass spectrometry. Rapid Communications in Mass Spectrometry, 23(4), 523–529. 10.1002/rcm.3904 [PubMed: 19142849]
14. Smith S (2013). Digital signal processing: A practical guide for engineers and scientists. Elsevier.
15. Robey MT, Utley D, Greer JB, Fellers RT, Kelleher NL, & Durbin KR (2023). Advancing intact protein quantitation with updated deconvolution routines. Analytical Chemistry, 95(40), 14954–14962. 10.1021/acs.analchem.3c02345 [PubMed: 37750863]
16. Maples MP, Reichart DE, Konz NC, Berger TA, Trotter AS, Martin JR, Dutton DA, Paggen ML, Joyner RE, & Salemi CP (2018). Robust Chauvenet outlier rejection. The Astrophysical Journal Supplement Series, 238(1), 2. 10.3847/1538-4365/aad23d
17. LV S, RJ M, MR S, S. M, & S. LM (2020). Improving proteoform identifications in complex systems through integration of bottom-up and top-down data. Journal of Proteome Research, 19(8), 3510–3517. 10.1021/acs.jproteome.0c00332 [PubMed: 32584579]
18. Dai Y, B. K. E., Schaffer LV, Miller RM, Millikin RJ, Scalf M, Frey BL, Shortreed MR, & Smith LM (2019). Constructing human proteoform families using intact-mass and top-down proteomics with a multi-protease global post-translational modification discovery database. Journal of Proteome Research, 18(10), 3671–3680. [PubMed: 31479276]

19. Li Q, Shortreed MR, Wenger CD, Frey BL, Schaffer LV, Scalf M, & Smith LM (2017). Global post-translational modification discovery. Journal of Proteome Research, 16(4), 1383–1390. 10.1021/acs.jproteome.6b00034 [PubMed: 28248113]

20. Rolfs Z, & Smith LM (2021). Internal fragment ions disambiguate and increase identifications in top-down proteomics. Journal of Proteome Research, 20(12), 5412–5418. 10.1021/acs.jproteome.1c00599 [PubMed: 34738820]

21. Miller RM, Millikin RJ, Rolfs Z, Shortreed MR, & Smith LM (2023). Enhanced proteomic data analysis with MetaMorpheus. In Burger T (Ed.), Statistical analysis of proteomic data: Methods and tools (pp. 35–66). Springer US.

22. Smith LM, Thomas PM, Shortreed MR, Schaffer LV, Fellers RT, LeDuc RD, Tucholski T, Ge Y, Agar JN, Anderson LC, Chamot-Rooke J, Gault J, Loo JA, Paša-Tolić L, Robinson CV, Schlüter H, Tsybin YO, Vilaseca M, Vizcaíno JA, . . . Kelleher NL (2019). A five-level classification system for proteoform identifications. Nature Methods, 16(10), 939–940. 10.1038/s41592-019-0573-x [PubMed: 31451767]

23. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, . . . Mallick P (2012). A cross-platform toolkit for mass spectrometry and proteomics. Journal of Proteome Research, 30(10), 918–920. doi: 10.1038/nbt.2377

24. Basharat AR, Zang Y, Sun L, & Liu X (2022). TopFD—A proteoform feature detection tool for top-down proteomics. Preprint, 10.1101/2022.10.11.511828

25. Jeong K, Kim J, Gaikwad M, Hidayah SN, Heikaus L, Schlüter H, & Kohlbacher O (2020). FLASHDeconv: Ultrafast, high-quality feature deconvolution for top-down proteomics. Cell Systems, 10(2), 213–218.e6. doi: 10.1016/j.cels.2020.01.003 [PubMed: 32078799]

26. L. S, & A. P (2014). ImageIntegration. Reference Documentation, 2023. https://pixinsight.com/doc/tools/ImageIntegration/ImageIntegration.html

27. Dai Y, Shortreed MR, Scalf M, Frey BL, Cesnik AJ, Solntsev S, Schaffer LV, & Smith LM (2017). Elucidating *Escherichia coli* proteoform families using intact-mass proteomics and a global PTM discovery database. J Proteome Res, 16(11), 4156–4165. 10.1021/acs.jproteome.7b00516 [PubMed: 28968100]

28. Dai Y, Buxton KE, Schaffer LV, Miller RM, Millikin RJ, Scalf M, Frey BL, Shortreed MR, & Smith LM (2019). Constructing human proteoform families using intact-mass and top-down proteomics with a multi-protease global post-translational modification discovery database. Journal of Proteome Research, 18(10), 3671–3680. 10.1021/acs.jproteome.9b00339 [PubMed: 31479276]

29. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, & Pfister H (2014). UpSet: Visualization of intersecting sets. IEEE Transactions on Visualization and Computer Graphics, 20(12), 1983–1992. 10.1109/TVCG.2014.2346248 [PubMed: 26356912]

30. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, & Old WM (2010). Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. Journal of Proteome Research, 9(8), 4152–4160. 10.1021/pr1003856 [PubMed: 20578722]

31. Maronna RAA (2019). Robust statistics: Theory and methods (with R.) (2nd ed.). Wiley. [Piscataqay, New Jersey]: IEEE Xplore [2019].

Significance Statement

The purpose of this study is to reduce the stochastic nature of mass spectrometry measurements and remove artifactual peaks. We did so through new spectral averaging with outlier rejection methods adapted from astronomical image processing. Integrating a rejection algorithm with averaging eliminates artifactual peaks, improving the quality of the deconvolution results and subsequently enabling the identification of 45% more proteoforms than not using averaging with rejection. These results pave the way toward a better understanding of spectral preprocessing methods to increase top-down proteoform identifications. The averaging with rejection algorithm described is implemented in the freely available and open-source proteomics search engine MetaMorpheus.

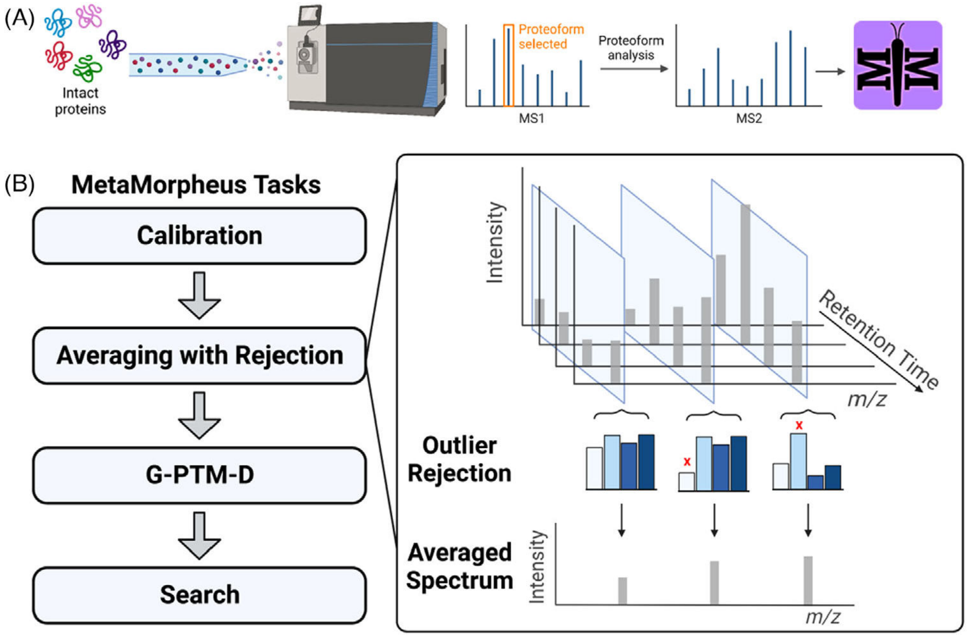


FIGURE 1.

Workflow. (A) Intact, denatured protein ions are analyzed by Orbitrap mass spectrometry in data-dependent acquisition mode and analyzed by MetaMorpheus. (B) The MetaMorpheus tasks used for pre-processing, post-translational modification discovery, and search.

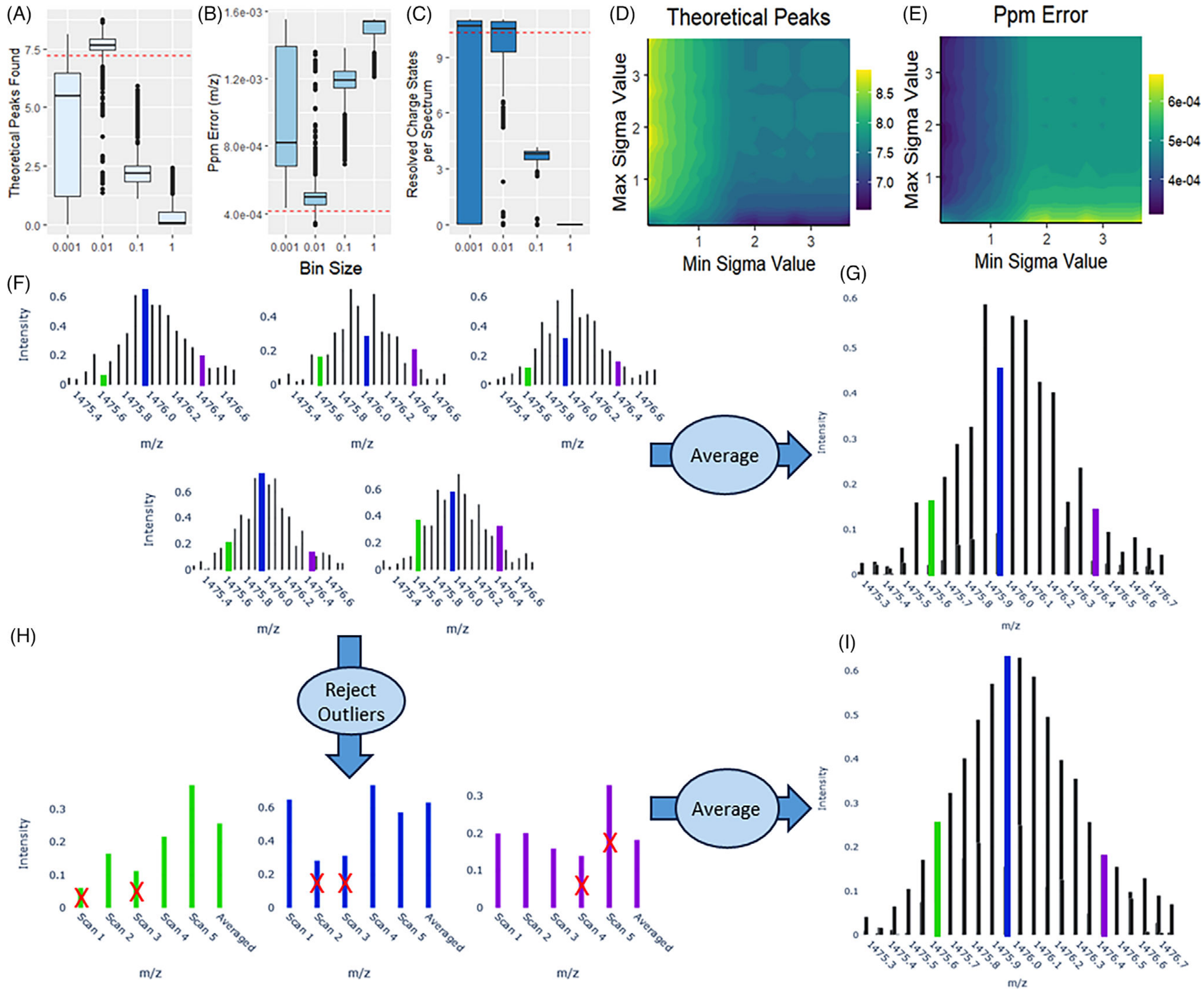


FIGURE 2.

Effect of Averaging Parameters on directly injected isolated protein standards. (A) The effect of m/z bin size on the number of theoretical most abundant isotopic peaks above 5% relative intensity, (B) the ppm error between each theoretical and corresponding experimental most abundant isotopic peak, and (C) the number of charge state resolvable isotopic envelopes per spectrum with the dashed red line representing the value for the original, unaveraged, spectra. (D) The number of theoretical most abundant peaks found and (E) their ppm error as a function of the minimum and maximum acceptable dispersion (sigma). (F) Five representative mass spectra were chosen to illustrate the averaging process with and without rejection. Isotopic peaks highlighted in green, blue, and purple. (G) Result of averaging without applying a rejection algorithm. (H) Representation of the rejection process for the highlighted isotopic peaks from each of the five representative mass spectra, with the x indicating a rejected peak. (I) Final result of averaging following the application of a rejection algorithm.

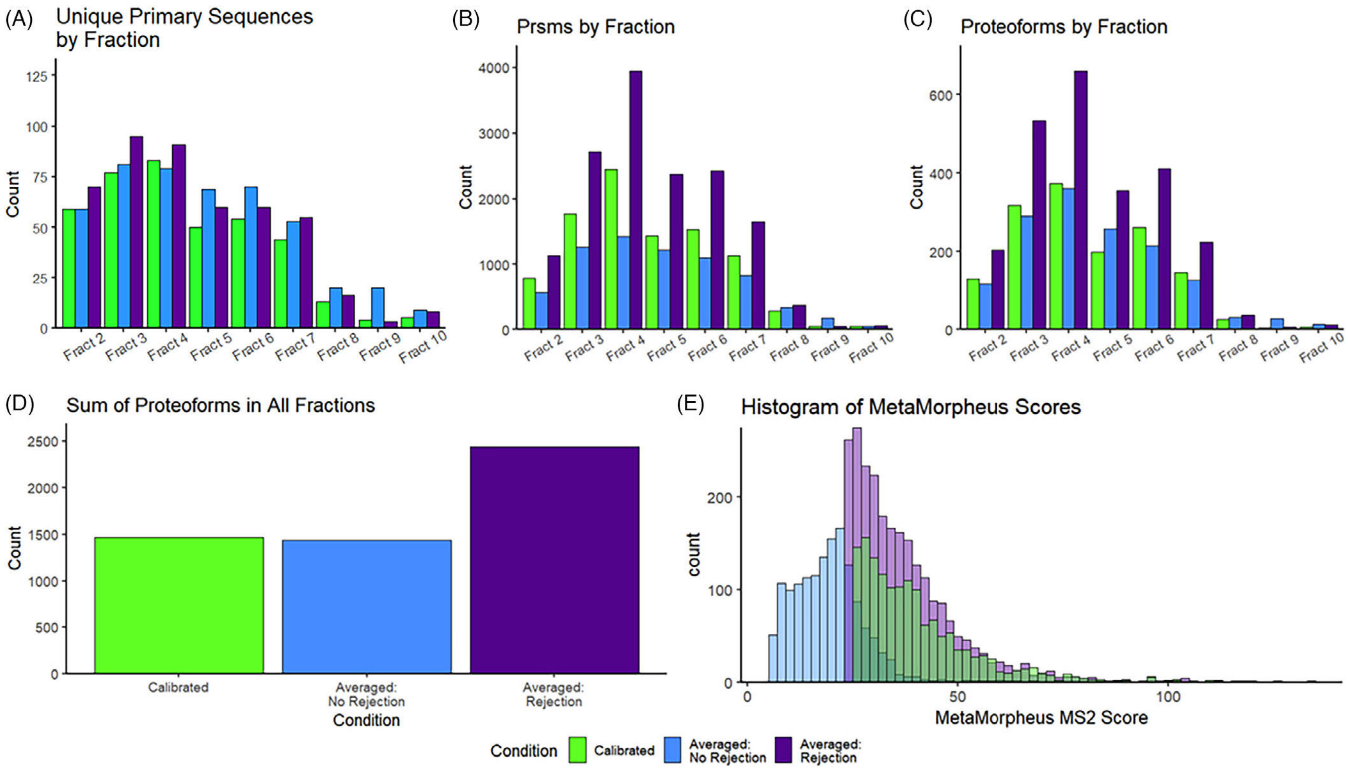


FIGURE 3.

(A) Count of unique primary sequences identified in each fraction. (B) Count of PrSMs identified in each fraction. (C) Count of proteoforms identified in each fraction. (D) Sum of proteoforms identified in all fractions. (E) Histogram of the MS2 scores.

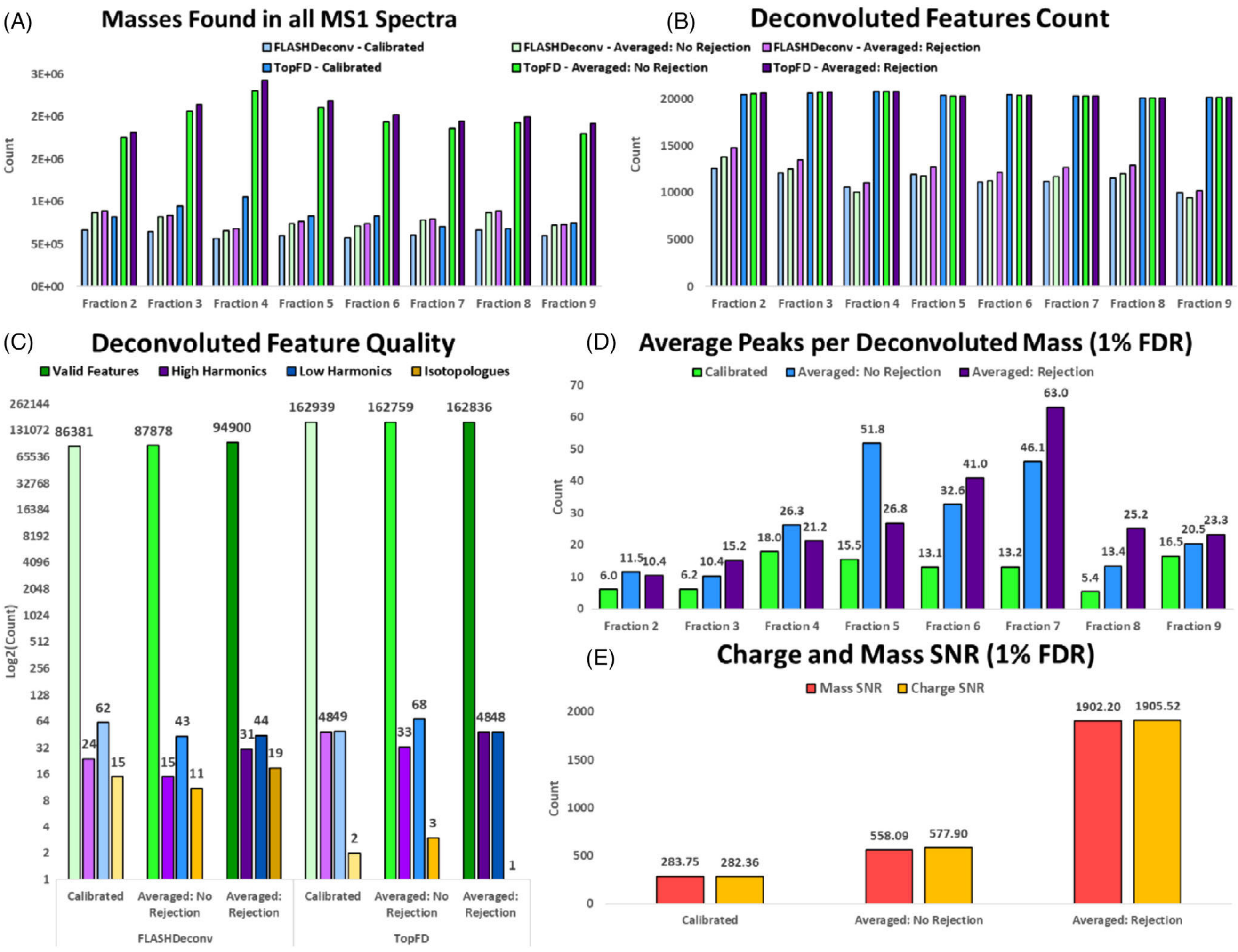
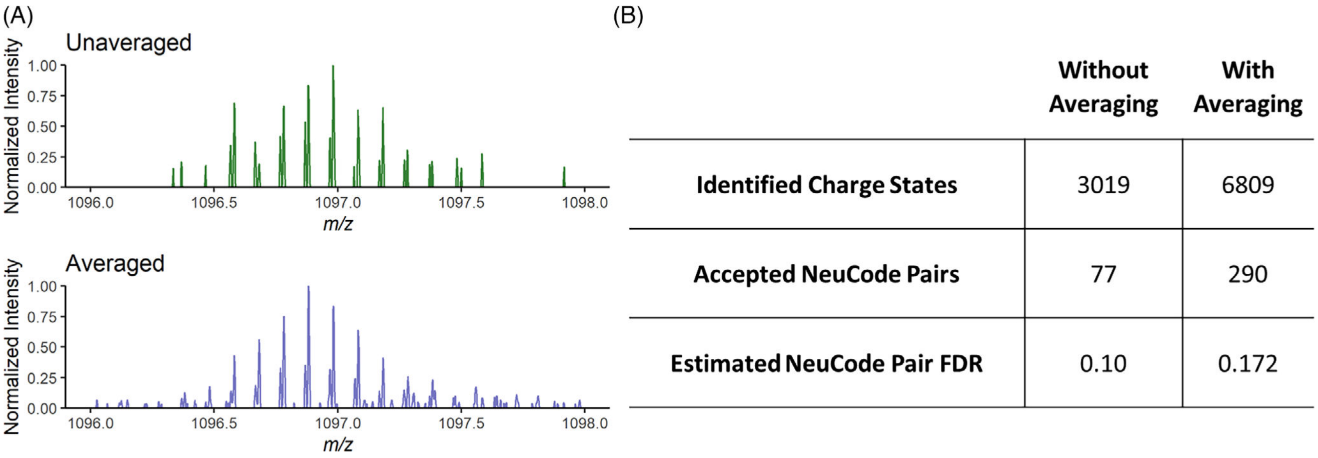


FIGURE 4.

Analysis of Deconvolution Results for the Jurkat top-down dataset. (A) Comparison of masses found within MS1 spectra per fraction as reported in MS1 align files from both TopFD and FLASHDeconv. (B) Comparison of deconvoluted features as reported in feature files from both TopFD and FLASHDeconv. (C) The number of valid and artifactual features detected for averaged and unaveraged spectra files from both TopFD and FlashDeconv. (D) Comparison of the average isotopic peaks per deconvoluted mass for targets and targets at 1% FDR as reported by FLASHDeconv. (E) Comparison of the mass signal-to-noise and charge signal-to-noise as reported by FLASHDeconv.



	Without Averaging	With Averaging
Identified Charge States	3019	6809
Accepted NeuCode Pairs	77	290
Estimated NeuCode Pair FDR	0.10	0.172

FIGURE 5.

(A) Top: Original, unaveraged spectrum from NeuCode experiment. Bottom: Spectrum after averaging. (B) Table of identified charge states accepted NeuCode pairs, and estimated NeuCode Pair FDR with and without using averaging as part of the pre-processing workflow.