Title:

FLASHDeconv: Ultrafast, high-quality feature deconvolution for top-down proteomics

Authors and affiliations:

Kyowon Jeong[1,2,†,*], Jihyung Kim[1,2,†], Manasi Gaikwad[3], Siti Nurul Hidayah[3], Laura Heikaus[3], Hartmut Schlüter[3], and Oliver Kohlbacher[1,2,4,5,6,*]

[1]Applied Bioinformatics, Department for Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany

[2]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

[3]Mass Spectrometric Proteomics, Institute of Clinical Chemistry and Laboratory Medicine, Campus Forschung, Universitätsklinikum Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany

[4]Center for Quantitative Biology, University of Tübingen, Sand 14, 72076 Tübingen, Germany [5]Biomolecular Interactions, Max Planck Institute for Developmental Biology, Max-Planck-Ring 4, 72076 Tübingen, Germany.

[6]Translational Bioinformatics, University Hospital Tübingen, Hoppe-Seyler-Str. 9, 72076 Tübingen, Germany

[†]These authors contributed equally to this work.

[*]Correspondence to: jeong@informatik.uni-tuebingen.de and oliver.kohlbacher@uni-tuebingen.de

Abstract

Feature deconvolution, the determination of intact proteoform masses, is crucial for top-down proteomics, but currently suffers from long runtimes and quality issues. We present FLASHDeconv, an algorithm based on a simple transformation of mass spectra, which turns deconvolution into the search for constant patterns thus greatly accelerating the process. We show higher deconvolution quality and two to three orders of magnitude faster execution speed than existing approaches.

Top-down (TD) proteomics has gained a lot of momentum for in-depth protein characterization and protein species analytics[1-7]. In contrast to bottom-up (BU) proteomics, where proteins are enzymatically digested and actually peptides are analyzed, the TD approach allows for the analysis of intact proteoforms (protein species arising from the same gene product via splice variants, genomic variation, post-translational modifications, degradation, etc.)[6, 8]. Due to the high mass of the analytes, however, in mass spectrometry (MS)-based TD approach (TD-MS), a single protein species can result in multiple MS features with different charge states and isotopic envelopes, making the signal structure highly complex and, in turn, the accurate determination of proteoform masses challenging[9]. Feature deconvolution (i.e., the determination of intact proteoform masses) is thus an essential step for TD data analysis[10, 11].

Current MS instrumentation and experimental protocols enable more complex analyses using TD proteomics, but in turn yield equally complex and large data sets[1, 2, 6, 7, 10, 12]. Hence, both quality and runtime of deconvolution algorithms have become an issue. Current deconvolution algorithms typically have long processing times of hours or days for complex data sets. Quality is another issue of these approaches. Mass artifacts and limitations with respect to mass or charge ranges reduce the methods' applicability[10, 11, 13].

We present FLASHDeconv, an algorithm for high-quality TD deconvolution two to three orders of magnitude faster than existing tools. The major speed-up of FLASHDeconv is achieved by very fast decharging (i.e., assigning charges to peaks) in the spectral deconvolution step (Fig. 1a). It is based on a simple transformation of mass spectra that transforms peak positions (m/z) to log m/z space. The variable (mass-dependent) patterns caused by different charge states of the same protein are turned into a mass-independent universal pattern by this transformation. Finding the occurrences of the universal pattern in a transformed mass spectrum can be done very efficiently by a single convolution calculation. FLASHDeconv also implements a deisotoping algorithm that works well for both isotopically resolved and unresolved regions and simple yet effective filtering/scoring schemes to reduce false positives (see online methods).

To evaluate its performance, we benchmarked FLASHDeconv with TD samples of varying complexity (three low complexity samples, Fig. 1b-e; and one highly complex sample, Fig. 2). The three low-complexity samples consist of individual proteins (Cyto - Bovine Cytochrome C, Fil - Filgrastim) and a standard protein mix (PIP - Pierce Intact Protein Standard Mix; see online methods for details). Cyto, Fil, and PIP data sets were acquired on a Quadrupole-Orbitrap at a resolution of 70,000 (Cyto and Fil data sets) and 17,500 (PIP). Due to their low masses (12,223.21 Da for Cytochrome C and

18,802.8 Da for Filgrastim) and high resolution, the spectra in the Cyto and Fil data sets contain isotopically resolved peaks. The PIP data set was acquired at a lower resolution of 17,500 to assess the deconvolution quality for unresolved isotope patterns. For the complex data set, we used a murine myoblast sample (unfractionated and generated by Q-Exactive HF at a resolution of 120,000) published previously[14]. This data set contains both isotopically resolved and unresolved peaks due to the wide mass range.

For comparison, Xtract[15] (in Thermo BioPharma Finder 3.1), ReSpect (Positive Probability Ltd, in Thermo BioPharma Finder 3.1), and Promex[16] (version 1.0.7017) were tested against FLASHDeconv for selected data sets: Xtract and Promex for Cyto, Fil, and myoblast; ReSpect for PIP and myoblast. Detailed sample generation procedure and tool parameters are found in the online methods.

Fig. 1b-e shows the results for the simple protein data sets (see Supplementary Data 1 for the full list of features). In Fig. 1b, it is shown that FLASHDeconv is 60-90 times faster than comparable tools. Per spectrum, the deconvolution took less than 6 ms (excluding spectrum loading time). Fig. 1c shows the deconvoluted spectra (deconvoluted masses and their intensities, merged for all RT range) for the isotopically resolved Cyto and Fil data sets. All three tools found a signal around the input protein masses within 3 Da mass

error. However, Promex reported additional harmonic masses (e.g., ½ and ⅓ of the input masses) for both data sets.

In the isotopically unresolved PIP data set, FLASHDeconv found all six input masses within 3 Da mass error (ranging from 9 to 70 kDa; Fig. 1d-e). ReSpect found five input masses, excluding the heaviest one. The features from each tool were also mapped to the RT-mass plane in Fig. 1e (feature map). The feature map from ReSpect shows that ReSpect reports thousands of features in addition to those directly relatable to the input masses. However, when we examined the mass and RT relations of the features, we found that at least 43% of them were classified as mass artifacts or isotopologues (see online methods). The results on the low-complexity samples demonstrate that FLASHDeconv accurately identifies the proteoform masses with high specificity and sensitivity for both isotopically resolved and unresolved data sets. Supplementary Fig. 1-2 provides the same information as Fig. 1b-e for replicate data sets, where the almost identical results were reproduced (see Supplementary Data 2-3 for the complete feature lists).

The advantage of FLASHDeconv is even more pronounced on the complex data set (see Supplementary Data 4 for feature lists). Fig. 2a shows the run time comparison for the myoblast data set. FLASHDeconv was significantly faster (up to 330 times faster than Xtract) than the other tools; it

took only four minutes to process the whole data set (approx. 60 ms/spectrum).

We found that the numbers of reported features differed significantly from tool to tool (from around 600 for Xtract to about 20,000 for ReSpect). We then examined the masses and RT relations among the features as above (see online methods) to assess whether mass artifacts or isotopologues contribute to this large feature count variation between tools (Fig. 2b). For FLASHDeconv, only 8% of the detected features actually turned out to be artifacts or isotopologues, whereas that number reached almost 80% for some of the other tools. Excluding the detected artifacts, the largest number of features (~6,000) were reported by FLASHDeconv, demonstrating its high sensitivity and specificity. To see this in a different perspective, we compared the total ion current (TIC) chromatogram (from the raw spectrum file) with the intensities summed over masses (from the reported features) in Fig. 2c. FLASHDeconv showed almost perfect fit to the TIC. All other methods show a much lower agreement with the TIC. Good agreement between the TIC and the summed-up feature signal basically indicates that a large portion of the recorded signal could be assigned to the detected features. Poor agreement can occur when parts of the signal are not detected (causing sensitivity drop) and/or false positive features are erroneously included (causing specificity drop).

The feature maps in Fig. 2d and the feature mass distributions in Fig. 2e together show FLASHDeconv features are centered between 40-60 kDa while the others between 0-20 kDa or 0-40 kDa (see Supplementary Fig. 3 for the rounded monoisotopic mass overlap between tools in low mass region). Not surprisingly, Xtract and Promex exclusively reported masses lower than 20 kDa, as most masses larger than 20 kDa would be represented by isotopically unresolved peaks (for which Xtract and Promex are not optimized). Next, we compared our feature masses with the protein masses identified by BU searches for the same sample (also reported by L. V. Schaffer et al., 2018[14]). The distribution of these BU-identified protein masses (Fig. 2e top panel) was similar to the FLASHDeconv feature mass distribution, suggesting that the high masses reported by FLASHDeconv are indeed present in the sample. For more rigorous evidence, lastly we matched our feature masses against the BU identified protein masses (Fig. 2f; see online methods for matching method). About 74% of the feature masses were matched when a single modification was allowed, confirming most of our feature masses are likely to be genuine proteoform masses. The other tools showed substantially lower agreement than FLASHDeconv (see Supplementary Table 1 and 2).

In summary, FLASHDeconv can carry out high-quality deconvolution for diverse types of (simple and complex; and isotopically resolved and unresolved) TD-MS data sets at unprecedented speed and quality. With its

high speed and versatility, FLASHDeconv could significantly increase the potential of various TD analyses including native-MS and TD-BU integrated studies[4, 17]. Lastly, the very fast processing times for spectral deconvolution could be used for real-time mass deconvolution on MS instruments, enabling smarter and more efficient MS/MS fragmentation methods for TD-MS than are currently being used.

Code availability

FLASHDeconv is implemented in C++ as a part of OpenMS[18] and available as platform independent open-source software under a BSD three-clause license at https://OpenMS.org/FLASHDeconv.

Data availability

The raw files for Cyto, Fil, and PIP triplicate data sets have been uploaded to MassIVE (https://massive.ucsd.edu) and are available under accession number MSV000084001 or under the digital object identifier https://doi.org/10.25345/C59D26.

Acknowledgements

We are very thankful to Lloyd M. Smith for generously making murine TD&BU data sets publicly available. We thank to CinnaGen Co. for sharing Filgrastim protein and are grateful to Christoph Henrich, Bernard Delanghe, David Horn, and Jennifer Sutton from Thermo Fisher Scientific for insightful discussions. K.J. is grateful to Jong-Seo Kim and Jong-Eun Park for their advice on the paper. J.K., M.G., S.N.H., L.H., H.S., and O.K. acknowledge funding from the Horizon 2020 Marie Sklodowska-Curie Action ITN 2017 of the European Commission (grant 765502-A4B). O.K. acknowledges funding from BMBF (de.NBI, 031A535A).

Contributions

O.K. conceived the idea of the fast deconvolution algorithm for TD-MS data sets. K.J. and J.K. developed and implemented FLASHDeconv algorithm. M.G., S.N.H., L.H., and H.S. designed the experiment and performed sample preparation and data acquisition. O.K. and H.S. led the project and provided resources. K.J. and J.K. wrote the manuscript with input from all authors. All authors commented on and approved the paper.

Competing financial interests

The authors declare no competing financial interests.

References

1. Toby, T. K., et al. *Annu Rev Anal Chem (Palo Alto Calif)* **9**, 499-519 (2016).
2. Leney, A. C. and Heck, A. J. *J Am Soc Mass Spectrom* **28**, 5-13 (2017).
3. Skinner, O. S., et al. *Nat Chem Biol* **14**, 36-41 (2018).
4. Li, H., et al. *Nat Chem* **10**, 139-148 (2018).
5. Chen, B., et al. *Anal Chem* **90**, 110-127 (2018).
6. Schaffer, L. V., et al. *Proteomics* **19**, e1800361 (2019).
7. Donnelly, D. P., et al. *Nat Methods* **16**, 587-594 (2019).
8. Smith, L. M. and Kelleher, N. L. *Science* **359**, 1106-1107 (2018).
9. Cai, W., et al. *Mol Cell Proteomics* **15**, 703-714 (2016).
10. Marty, M. T., et al. *Anal Chem* **87**, 4370-4376 (2015).
11. Bern, M., et al. *J Proteome Res* **17**, 1216-1226 (2018).
12. Cesnik, A. J., et al. *J Proteome Res* **17**, 568-578 (2018).
13. Compton, P. D., et al. *Anal Chem* **83**, 6868-6874 (2011).
14. Schaffer, L. V., et al. *J Proteome Res* **17**, 3526-3536 (2018).
15. Zabrouskov, V., et al. *J Am Soc Mass Spectrom* **16**, 2027-2038 (2005).
16. Park, J., et al. *Nat Methods* **14**, 909-914 (2017).
17. Ntai, I., et al. *Mol Cell Proteomics* **15**, 45-56 (2016).
18. Rost, H. L., et al. *Nat Methods* **13**, 741-748 (2016).

Figure Legends

Figure 1. FLASHDeconv algorithm and analysis of simple data sets

a. FLASHDeconv consists of two separate processing steps: spectral deconvolution and feature finding. Spectral deconvolution finds monoisotopic masses from each TD-MS spectrum. The upper box illustrates the key idea for rapid deconvolution: if each peak position is transformed into log of its neutral m/z value, the charge pattern (the position pattern caused by different charge states of the same protein) in the transformed spectrum becomes independent of protein masses and hence is universal. Thus, decharging can be done by a single sliding this universal pattern along the transformed spectrum (see online methods) to identify matching peaks. Feature finding then reduces to finding agreeing deconvolved masses within given mass tolerance along the retention time (RT) direction (done by a mass trace detection algorithm[19]).

b. Run time comparison for simple data sets (Cyto, Fil, and PIP; see main text for abbreviations). FD stands for FLASHDeconv; XT for Xtract; RP for ReSpect; and PM for Promex. FD was up to 90 times faster than the other methods.

c. Deconvolved spectra generated by FD, XT, and PM for isotopically resolved pure protein data sets(Cyto - upper panel; Fil - lower panel). Dotted black lines indicate the expected exact masses (12,223.21 Da and 18,802.8 Da), and the colored digits specify the reported masses within 3 Da from exact input monoisotopic masses. For intensity estimates, MaxIntensity column was used for FD, Sum_Intensity for XT, and Abundance for PM. All three methods found the expected masses, but PM reported additional high-intensity low-harmonic mass artifacts (e.g., ½ and ⅓ input masses; marked with arrows).

d. Deconvolved spectra from FD and RP for the PIP (six standard intact protein mix) data set (isotopically unresolved). The colored digits specify the reported masses within 3 Da from input monoisotopic (for FD) or average (for RP) masses. For intensity estimation, Sum_Intensity was used for RP. FD found all six input protein masses. RP found five. The expected exact monoisotopic masses are 9,105.3, 11,858.0, 21,429.8, 28,963.7, 50,429.8, and 67,959.4 Da. The expected average masses are 9,111.5, 11,865.5, 21,442.6, 28,981.3, 50,459.7, and 68,001.2 Da.

e. Identified features mapped on the RT-mass plane (feature maps) from FD and RP for the PIP data set. Except for the five expected masses, RP also reported thousands masses, out of which 43% were identified as mass artifacts or isotopologues (see online methods for the artifact

detection method). See also Supplementary Fig. 1-2 for more results from the replicates.

Figure 2. Analysis of complex murine myoblast data set

- **a.** Runtime comparison for the murine myoblast data set. FD was drastically faster (25-330 times) than the other methods.
- **b.** The numbers of valid features (i.e., not detected as artifact) and detected mass artifacts/isotopologues from each tool (see online methods for the artifact detection method). The blue digits show the portion of the valid features. FD found almost 7,000 features, and only 8% of them were artifacts. The other tools generated significantly higher portion of artifacts (60-80%).
- **c.** For each tool, feature intensities were summed across all masses and plotted along RT direction. For reference, total ion current (TIC) chromatogram was also drawn on top. FD and the TIC showed almost perfect fit, demonstrating FD features cover most of eluted ions.
- **d.** The feature maps for each tool. For the 0-20 kDa region, all tools produced intense features (see Supplementary Fig. 3 for rounded monoisotopic mass overlaps in low mass region). XT and PM reported only few features for 20-100 kDa region, because most of the corresponding peaks would be isotopically unresolved.

e. The mass distribution of each tool as compared to the protein mass distribution obtained from bottom-up (BU) search (of the same myoblast sample). The distribution from FD is similar to the BU identified protein mass distribution.

f. The portion of the feature masses matched to the BU identified protein masses for each tool (see Supplementary Table S1). FD exhibits a much higher consistency with BU results than other tools.

Online methods

FLASHDeconv algorithm

FLASHDeconv consists of two sub-algorithms, spectral deconvolution and feature finding. FLASHDeconv performs fast mass deconvolution of each spectrum, and subsequent mass trace detection along retention time in the deconvolved spectra.

Spectral deconvolution algorithm

Decharging and harmonic artifact removal

FLASHDeconv reads mass spectra in HUPO-PSI-compliant mzML format[20] and then deconvolves each of these spectra individually. For this study, we used ProteoWizard[21] msconvert tool (without peak picking or filtration option) to convert raw spectrum files to mzML files. Goal of the decharging is to identify all peaks caused by the same protein mass, where the peaks are differing only with respect to their charge states. No initial peak filtration is applied by default since even low-intensity peaks can be important evidence for low-abundance features. Finding these groups of related peaks (peak groups) is achieved by searching for a mass-invariant charge pattern in log-transformed m/z space (see main text, Fig. 1 for the concept).

For each input spectrum **S,** FLASHDeconv computes a transformed spectrum $\mathbf{S}^*$ by subtracting the mass of the charge carrier (e.g., in positive

mode, a proton mass) and logarithmizing (natural base) the resulting value for each peak. We denote the mass of the charge carrier by $\boldsymbol{q}$. As m/z of a peak of charge $\boldsymbol{c}$ from mass $\boldsymbol{m}$ is given by $(\boldsymbol{m} + \boldsymbol{cq})/\boldsymbol{c} = \boldsymbol{m}/\boldsymbol{c} + \boldsymbol{q}$, the transformed peak position is at $\log(\boldsymbol{m}/\boldsymbol{c}) = \log(\boldsymbol{m}) - \log(\boldsymbol{c})$. Thus, for a series of peaks $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ from the same mass $\boldsymbol{m}$ with charges $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n$, the transformed peak locations are given by $\log(\boldsymbol{m}) - \log(\boldsymbol{c}_1)$, $\log(\boldsymbol{m}) - \log(\boldsymbol{c}_2)$, ..., $\log(\boldsymbol{m}) - \log(\boldsymbol{c}_n)$. We can thus define a universal charge pattern $\mathbf{U} := (-\log(\boldsymbol{c}_{min}), -\log(\boldsymbol{c}_{min} + 1), \ldots, -\log(\boldsymbol{c}_{max}))$ for a given charge range $[\boldsymbol{c}_{min}, \boldsymbol{c}_{max}]$ (two and 100 by default). Then, the peaks in the peak group of mass $\boldsymbol{m}$ should partially match (within tolerance) the universal charge pattern $\mathbf{U}$ positioned at $\boldsymbol{log}(\boldsymbol{m})$. Similar to the universal charge pattern $\mathbf{U}$, we define the harmonic charge patterns $\mathbf{H}_r$ for $\boldsymbol{r}$=½, ⅓, ⅖ as $\mathbf{H}_r := (-\log(\boldsymbol{c}_{min} + \boldsymbol{r}), -\log(\boldsymbol{c}_{min} + 1 + \boldsymbol{r}), \ldots, -\log(\boldsymbol{c}_{max} + \boldsymbol{r}))$. Harmonic charge patterns are used to filter out low harmonic artifacts (e.g., the masses having ½, ⅓, ¼, ⅕, or ⅖ times the true mass). Further filtration of (both high and low) harmonic artifacts are performed after peak groups are defined below.

FLASHDeconv slides $\mathbf{U}$ and $\mathbf{H}_r$ along the transformed spectrum $\mathbf{S}^*$ to find suitable matching peak groups. Let $\mathbf{U}(\boldsymbol{x})$ denote the pattern $\mathbf{U}$ at the position $\boldsymbol{x}$ in $\mathbf{S}^*$ (likewise, $\mathbf{H}_r(\boldsymbol{x})$). Note that $\mathbf{U}(\boldsymbol{x})$ is matched by the peaks from

mass $\boldsymbol{exp}(\boldsymbol{x})$. Two peaks $\boldsymbol{p}$ and $\boldsymbol{p'}$ are called comparable if their intensity ratio is between ¼ and 4. We call two peaks $\boldsymbol{p}$ and $\boldsymbol{p'}$ are well-aligned if they are comparable and they match to any of two consecutive elements of $\mathbf{U}(\boldsymbol{x})$. A peak $\boldsymbol{p_h}$ is a low harmonic peak with respect to well-aligned $\boldsymbol{p}$ and $\boldsymbol{p'}$ if i) $\boldsymbol{p_h}$ is located between $\boldsymbol{p}$ and $\boldsymbol{p'}$, ii) $\boldsymbol{p_h}$ is comparable with $\boldsymbol{p}$ and $\boldsymbol{p'}$, and iii) $\boldsymbol{p_h}$ matches to any element of $\mathbf{H}_r(\boldsymbol{x})$. Then, two well-aligned peaks $\boldsymbol{p}$ and $\boldsymbol{p'}$ are further called signal peaks if they have no low harmonic peak.

While sliding a pattern along the transformed spectrum, FLASHDeconv stops at the location $\boldsymbol{x}$ where three consecutive signal peaks $\boldsymbol{p_1}$, $\boldsymbol{p_2}$, $\boldsymbol{p_3}$ are detected, that is, $(\mathbf{p}_1, \boldsymbol{p_2})$, and $(\boldsymbol{p_2}, \boldsymbol{p_3})$ are pairwise signal peaks. Then it recruits all peaks matched to $\mathbf{U}(\boldsymbol{x})$, forming a peak group of mass $\boldsymbol{exp}(\boldsymbol{x})$. A peak matched to an element -log($\boldsymbol{c}$) in $\mathbf{U}(\boldsymbol{x})$ now has the assigned charge of $\boldsymbol{c}$.

Even though the method described above efficiently collects peak groups and assigns charges to their peaks, still multiple peak groups represent either low or high harmonic artifacts. Also, quite often a single peak has multiple assigned charges. This happens mainly when a single peak has membership in multiple peak groups, often those representing harmonic artifacts. To resolve this, we assign an intensity score (described below) to each peak group. Briefly, the intensity score measures the total signal peak intensity subtracted by the total harmonic (both high and low) peak intensity.

Then, for each peak, we collect all peak groups containing the peak. The peak group of the highest intensity score is allowed to retain the peak. The peak charge is updated accordingly. At this point, the set of peak groups is called the candidate peak group set. The mass of a candidate peak group is determined by the most intense peak mass therein.

To define the intensity score of a peak group, we detect possible harmonic peaks in the peak group. The low harmonic peaks are already detected when finding signal peaks. To detect high harmonic peaks, first, we sort the peaks in ascending order of their assigned charges. Denote the $\boldsymbol{i}$-th peak in the sorted peak list as $\boldsymbol{p_i}$. For each peak $\boldsymbol{p_i}$, we take the peaks $\boldsymbol{p_{i-1}}$ and $\boldsymbol{p_{i+1}}$. The peak $\boldsymbol{p_i}$ is called a high harmonic peak if $\boldsymbol{p_i}$ is comparable but is not well-aligned with $\boldsymbol{p_{i-1}}$ and $\boldsymbol{p_{i+1}}$. For all collected signal, low harmonic, and high harmonic peaks in the peak group, the intensity score of a peak group is defined as its total signal peak intensity subtracted by its total low and high harmonic peak intensities.

The number of candidate peak groups is usually less than 2,000 even for complex spectra. Thus complicated processing of the candidate peak groups from this point on does not slow down FLASHDeconv significantly.

Peak group clustering for isotope envelopes (isotopically resolved case)

Each peak group has its mass, but some of them are from the isotopologues of the same protein. Moreover, some of the peaks in $\mathbf{S}^*$ may not have been selected for any peak group due to their low abundance even though they are isotopologues of existing peak groups. Thus, after defining candidate peak groups, we cluster peak groups and recruit additional peaks from plausible protein isotopologues.

This processing is done in a greedy fashion from the peak group with the smallest mass to the one with the largest mass. Let $\boldsymbol{m_{iso}}$ be the mass difference between ^{13}C and ^{12}C. Suppose we are processing a peak group **P** of mass $\boldsymbol{m}$. First, we examine if there are peak groups **P'** of mass $\boldsymbol{m} - \boldsymbol{m_{iso}}$ or $\boldsymbol{m} + \boldsymbol{m_{iso}}$. If no **P'** is observed or one of the intensity scores of **P'** is higher than that of **P**, we discard **P**. If **P** is retained, we merge peak groups of masses $\boldsymbol{m} + \boldsymbol{n}\boldsymbol{m_{iso}}$ for $\boldsymbol{n}$ = 0, 1, 2, ..., sequentially, until there is no such peak group left. At the same time, we also recruit isotopologue peaks. Denote the mass of charge carrier by **q**. For each peak in **P** of charge **c**, we recruit peaks in the raw spectrum **S** having m/z of $(\boldsymbol{m} + \boldsymbol{n}\boldsymbol{m_{iso}} + \mathbf{cq})/\mathbf{c}$ to **P** for $\boldsymbol{n}$ = 0, 1, 2, ..., until no such peak is found. We repeat the same procedure also for $\boldsymbol{n}$ = -1, -2,... This twoway scan ensures that the light (and less abundant) isotopologues are included in **P**. During this merging and recruiting, any peak added to **P** at $\boldsymbol{n}=\boldsymbol{i}$ is assigned the (tentative) isotope index $\boldsymbol{i}$, which can be negative. After

finishing the above procedure for **P**, all isotope indices are shifted so the minimum peak isotope index in **P** is 0 (still, tentative indices; see below). Each peak group now contains the peaks with different charges and different isotope indices that are likely to be from the same protein mass.

Peak group clustering for isotope envelopes (isotopically unresolved case)

In cases where isotopologues are not baseline resolved (in m/z direction) in raw profile spectrum, the corresponding peaks in the centroid spectrum (converted by mscovert in our case) often have inaccurate m/z positions. Furthermore, the peaks from distinct isotopologues are located so close to each other that they are within the mass tolerance. Thus, the above peak group clustering procedure results in the peak group **P** containing multiple occurance of the identical peaks with different isotope indices causing the scoring in the next step ineffective. To avoid this, we first retain only unique peaks in **P**. Then, for each unique peak, the isotope index **n** is assigned such that the distance between the peak mass and the mass of the **n**-th isotopologue (i.e., $\boldsymbol{m} + \boldsymbol{n}\boldsymbol{m}_{\boldsymbol{iso}}$) is minimized.

Filtration/scoring of peak groups and determination of the final isotope indices

To reduce false-positive masses, we filter out peak groups based on their per charge or per isotope intensity distribution. In this process, the determination of the final isotope indices (and thus the monoisotopic mass) is also carried out.

For a peak group, let the charge intensity for **c** be the summed intensity over the peaks of charge **c** in the peak group. We say a charge is intense if its charge intensity is higher than 25% of the max charge intensity. Let $\mathbf{m_i}$ be the maximum number of intense charges that continuously increment by **i**. For example, if charges 3, 5, 6, 7, 9, 10 are intense, $\mathbf{m_1} = 3$ (charges 5, 6, 7) while $\mathbf{m_2} = 4$ (charges 3, 5, 7, 9). A peak group is filtered out if $\mathbf{m_1} < \max(3, \mathbf{m_2}, \mathbf{m_3}, \ldots, \mathbf{m_7})$. This filtration dramatically reduces remaining high harmonic artifacts.

Then the charge intensities are fitted to a Gaussian function using Caruana's algorithm[22], and if the cosine similarity between the charge intensities and the fitted Gaussian is less than 0.4 (user-specified parameter), we discard the peak group. Before the fitting, we find the smallest and largest charges such that their intensities exceed 2% of the maximum charge intensity. The charge intensities between the found charges (including zero charge intensities) are used as the input to Caruana's algorithm. If Caruana's

algorithm yields a negative variance, the peak group is disregarded. Otherwise, the cosine similarity is calculated.

The isotope intensity for $\boldsymbol{i}$ (the summed intensity over the peaks of isotope index $\boldsymbol{i}$ in the peak group) is then used for further filtration. The isotope intensities are compared against the averagine distribution of the (tentative) monoisotopic mass of the peak group[23]. Let $\boldsymbol{w}$ be the isotope index of the maximum abundance in the averagine distribution. And denote the isotope intensity for $\boldsymbol{i}$ as $\mathbf{I}_{\boldsymbol{i}}$. As the isotope indices of the peaks might not be accurate, we give a large isotope index offset Δ from $-\boldsymbol{w}$ to $\boldsymbol{w}$ and pick the offset Δ^* that maximizes the cosine similarity between the isotope intensities $\mathbf{I}_{\boldsymbol{i}+\boldsymbol{off}}$ (for $\boldsymbol{i}=-\Delta, \ldots, 2\boldsymbol{w}-\Delta$) and the avergine distribution. If the resulting maximum cosine similarity is less than 0.4 (user-specified), the peak group is filtered out. Otherwise, the isotope indices of all peaks in the peak group are shifted by Δ^* (giving the final isotope indices). Monoisotopic mass is decided accordingly with the mass and the final isotope index of the most intense peak in the peak group.

Use of RT-adjacent spectral information

Similar to Xtract and ReSpect, FLASHDeconv can use information from spectra close in retention time. Given a spectrum $\mathbf{S}$ at RT $\boldsymbol{rt}$ and an RT

window Δrt (user specified), we retain all the average masses of peak groups obtained from spectra at RTs between $rt - \Delta rt$ and rt. When processing the spectrum **S**, these masses are used to generate additional peak groups for the candidate peak groups; for these masses, the peak groups are formed and added even if there is no three consecutive signal peaks in **S**. The next steps (i.e., assigning unique charges to peaks, isotope envelope clustering, and peak group filtration) are equally applied to these additional peak groups.

Feature finding by mass trace detection

Once all spectra have been deconvolved as described above (and if there are multiple spectra), FLASHDeconv searches for features along the retention time. This idea goes back to the notion of 'mass trace detection' frequently used in quantitative mass spectrometry. We use the term feature here to describe all the signal arising from the same proteoform. Integrating this signal over retention time and summing it up across charge states/isotopes reduces the complexity of the data set and results in a feature primarily characterized by its retention time, mass, and intensity. Thus, feature detection is based on the notion of a mass trace in the deconvolved spectra.

We use a robust mass trace detection algorithm[19] based on a Gaussian kernel density estimation of the local mass error and as implemented in

OpenMS. For mass tolerance, twice the input tolerance to FLASHDeconv was used. The “re-estimate mass tolerance” function was activated. For each feature, the charge intensities and isotope intensities are defined similar to those for peak groups and are used for filtration of features (instead of peak groups) as above, but with higher cosine similarity score threshold (0.6, user-specified parameter).

Cyto (Bovine Cytochrome C) and Fil (Filgrastim) data set acquisition

For the Cyto data set, Bovine Cytochrome C (P62894 UniProt accession number) was purchased from Sigma-Aldrich (Darmstadt, Germany). For the Fil data set, Filgrastim was kindly provided by CinnaGen Co. HPLC-grade H_2O, acetonitrile (ACN) and formic acid (FA) were purchased from Merck (Darmstadt, Germany).

Each protein was dissolved with 0.1% FA in H_2O to the concentration of 0.5 μg/μL. Reversed-phase liquid chromatography analysis was performed with an ultra-pressure liquid-chromatography (UPLC) system (ACQUITY, Waters, Manchester, UK). 0.1 % FA in H_2O was used as mobile phase A and 0.1% FA in ACN was used as mobile phase B. 1 μL of sample were loaded onto a reversed-phase column (monolithic Proswift RP-4 Analytical 1 x 50 mm; Thermo Scientific, Bremen, Germany) and washed for 1 min with 2% mobile

phase B. The proteins were eluted at a constant flow rate of 0.1 μL/min with a linear gradient increasing to 25% B in 5 min, 40% B in 5 min and 70% B in 3 min and hold in 70% B for 1 min. Column temperature was set to 30°C.

The eluting proteins were ionized via ESI and analyzed with Quadrupole-Orbitrap mass spectrometry (Q-Exactive, Thermo Scientific, Bremen, Germany) in a positive mode with 20.0 eV in-source CID. Full scan mass spectra were acquired at mass range 500 – 4,000 m/z, with a resolution of 70,000 and an AGC target of 3,000,000.

PIP (Pierce Intact Protein Standard Mix) data set acquisition

Pierce Intact Protein Standard Mix containing six standard proteins namely Protein G, Protein AG, IGF-I LR3, Thioredoxin, Carbonic Anhydrase II, and Exo Klenow was purchased from Thermo Scientific (Bremen, Germany). The lyophilised powder of the mixture was reconstituted in 100 μL HPLC-grade water to a final concentration of 0.76 μg/μL. LC-MS analysis was performed as mentioned above with 0.1% FA water as mobile phase A, 0.1% FA acetonitrile as mobile phase B, and column temperature was set to 30°C. 1μL of sample was loaded onto a reversed-phase column (monolithic Proswift RP-4 Analytical 1 x 250 mm, Thermo Scientific, Bremen) and washed 2 min with 5% mobile phase B. The proteins were eluted at a constant flow rate of

0.2 μL/min with a linear gradient increasing to 25% B in 5 min, 45% B in the next 27 min. The MS data was obtained on a quadrupole-orbitrap mass spectrometer operated in positive mode with 20.0 eV in-source CID at a mass range of 500 to 3,000m/z, MS1 only at 17,500 resolution with 5 microscans and AGC target of 3,000,000.

Tool parameters

FLASHDeconv

FLASHDeconv was used with default parameters for all data sets. By default, the mass range is set to 1-100 kDa and the charge range is 2-100. The isotope cosine threshold and charge intensity cosine threshold are set to 0.4 (spectrum level) and 0.6 (feature level). Mass tolerance is 10 ppm. The RT window is automatically set to 1% of total gradient time per data set.

Xtract

"Intact Protein Analysis" experiment type was used for all data sets. We used the processing method "Default SW Xtract" with the following modifications. "Output Mass Range" was set to 1-100 kDa for all data sets. In the case of "Charge Range", 2-100 was used for the simple (Cyto, Fil, and PIP) data sets, but 2-50 for the myoblast data set, because Xtract often stopped working with 2-100 range for the myoblast data set. "Automatic Sliding Window

Parameter Values" option was turned on for the simple data sets. But for the myoblast data set, the option was deactivated and "Target Avg Spectrum Width" was set to 3 minutes.

ReSpect

"Intact Protein Analysis" experiment type was used for all data sets. We used the processing method "Default SW ReSpect" with the following modifications. The mass ranges (both "Output Mass Range" and "Model Mass Range") were set to 1-100 kDa and "Charge Range" to 2-100 for all data sets. "Target Mass" was set to 50 kDa. "Automatic Sliding Window Parameter Values" option was turned on for the simple data sets. But for the myoblast data set, the option was deactivated and "Target Avg Spectrum Width" was set to 3 minutes.

Promex

For all data sets, the mass range was set to 1-100 kDa, and the charge range was set to 2-60 (the maximum charge for Promex is 60).

Mass artifact and isotopologue detection in a feature list

Given a set of features $\mathbf{F}$, to test if a feature $\boldsymbol{f}$ in $\mathbf{F}$ with mass $\boldsymbol{m}$ is a mass artifact or one of isotopologues, we first collected all features in $\mathbf{F}$ such that the RT overlap between the feature and $\boldsymbol{f}$ exceeds 60% of the RT span of $\boldsymbol{f}$.

Secondly, the features having lower intensity than $\boldsymbol{f}$ are filtered out. Denote the remaining feature set by $\mathbf{F}_f$. To determine, for example, if $\boldsymbol{f}$ is a low harmonic artifact, we tested if there exists a feature in $\mathbf{F}_f$ that has mass of $\boldsymbol{cm}$ for $\boldsymbol{c}$ = 2, ..., 100 (within 10 ppm tolerance) allowing up to 10 isotope error. If such a feature exists, we declared $\boldsymbol{f}$ as a low harmonic artifact feature. More formally, $\boldsymbol{f}$ is a low harmonic artifact if there exists a feature in $\mathbf{F}_f$ whose mass equals $\boldsymbol{c}(\boldsymbol{m} + \boldsymbol{k}\boldsymbol{m}_{iso})$ within 10 ppm tolerance for $\boldsymbol{c}$=2,...,100 and $\boldsymbol{k}$= -10, ..., 10, where $\boldsymbol{m}_{iso}$ denotes the mass difference between ^{13}C and ^{12}C. For high harmonic artifacts, the mass term $(\boldsymbol{m} + \boldsymbol{k}\boldsymbol{m}_{iso})/\boldsymbol{c}$ was used in the place of $\boldsymbol{c}(\boldsymbol{m} + \boldsymbol{k}\boldsymbol{m}_{iso})$.

Charge off by one artifact features[11] occur when a charge $\boldsymbol{c}$+1 or $\boldsymbol{c}$−1 is assigned to a charge $\boldsymbol{c}$ peak. They mainly arise when the charge assignment is based on the peak m/z distance between isotopologues because the measurement of the distance for highly charged peaks is often inaccurate, not to mention the isotopically unresolved peaks. But even when the charge assignment is based on the charge pattern, the artifact can be generated if two peaks with consecutive charge states are close to each other (e.g, below 4 Th), due to peak m/z variations. Thus, to detect the charge off by one artifacts, we chose the charge states $\boldsymbol{c}^*$ (between 2 and 100) such that $\boldsymbol{m}/\boldsymbol{c}^*$ -

$\mathbf{m}/(\boldsymbol{c}^*+1)<4$. For all such $\boldsymbol{c}^*$, we tested if features having masses $(\boldsymbol{m}+\boldsymbol{k}\boldsymbol{m}_{iso})*(\boldsymbol{c}^*+1)/\boldsymbol{c}^*$ or $(\boldsymbol{m}+\boldsymbol{k}\boldsymbol{m}_{iso})*(\boldsymbol{c}^*-1)/\boldsymbol{c}^*$ are present in $\boldsymbol{F}_f$ (10 ppm mass tolerance allowed) for $\boldsymbol{k}$= -10, ..., 10. If present, we declared $\boldsymbol{f}$ to be a charge off by one artifact.

Lastly, to see if $\boldsymbol{f}$ has other features corresponding to its isotopologues, we simply checked if features having masses $\boldsymbol{m} - \boldsymbol{m}_{iso}$ or $\boldsymbol{m} + \boldsymbol{m}_{iso}$, within 10 ppm tolerance, are present in $\boldsymbol{F}_f$. If present, $\boldsymbol{f}$ is considered to be one of the isotopologues. We repeat the above procedure for each feature in the list $\boldsymbol{F}$.

Matching feature masses against protein masses identified by bottom-up searches

Feature masses from complex data sets were matched to the protein masses identified by BU search, from the same myoblast sample. To collect the protein masses to be matched, we took the UniProt (downloaded in April, 2019) entries identified by BU search in L. V. Schaffer et al., 2018[14]. The (unmodified) protein masses were then calculated using their amino acid sequences. For the matches with modifications, 18 relatively abundant modification types found by PTMCurator[24] and two protein N-terminal modifications (Met-loss and Met-loss+Acetyl) were considered (see

Supplementary Table 2). We performed two matchings: one with zero modification and the other with up to one modification per protein.

Matching between a feature mass and the set of all theoretical proteoform masses (subject to the above modification criteria) was done by using the multimap data structure. For each possible proteoform mass, the multimap takes the mass as key and the proteoform(s) having the mass as value. A feature is said to be matched to a proteoform if the key (i.e., mass) of the proteoform is within 10 ppm tolerance from the feature mass, allowing one isotope error.

Online methods references

19. Kenar, E., et al. *Mol Cell Proteomics* **13**, 348-359 (2014).
20. Martens, L., et al. *Mol Cell Proteomics* **10**, R110 000133 (2011).
21. Kessner, D., et al. *Bioinformatics* **24**, 2534-2536 (2008).
22. Caruana, R. A., et al. *Anal Chem* **58**, 1162-1167 (1986).
23. Senko, M. W., et al. *J Am Soc Mass Spectrom* **6**, 229-233 (1995).
24. Khoury, G. A., et al. *Sci Rep* **1** (2011).

Supplementary information

Supplementary Figure 1. Analogue of Figure 1b-e for the simple data sets (Cyto, Fil, and PIP) technical replicate 1.

Supplementary Figure 2. **Analogue of Figure 1b-e for the simple data sets (Cyto, Fil, and PIP) technical replicate 2.**

Supplementary Figure 3. **Venn diagram for the overlap of the rounded monoisotopic masses between tools in low mass region for the myoblast data set.**

Supplementary Data 1. **The outputs from FLASHDeconv, Xtract, ReSpect, and Promex for the simple data sets (Cyto, Fil, and PIP).**

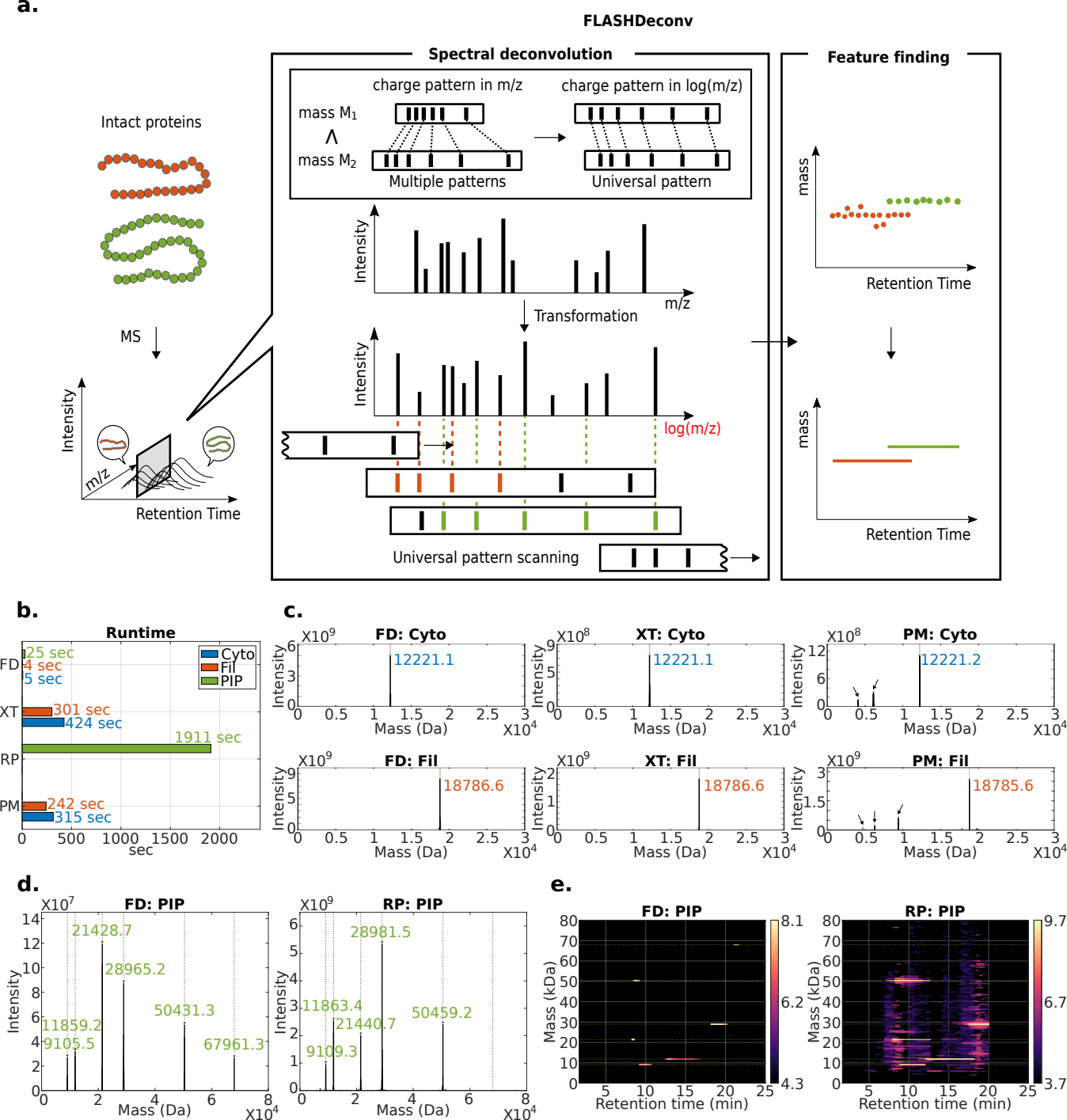
Supplementary Data 2. **The outputs from FLASHDeconv, Xtract, ReSpect, and Promex for the simple data sets (Cyto, Fil, and PIP) technical replicate 1.**

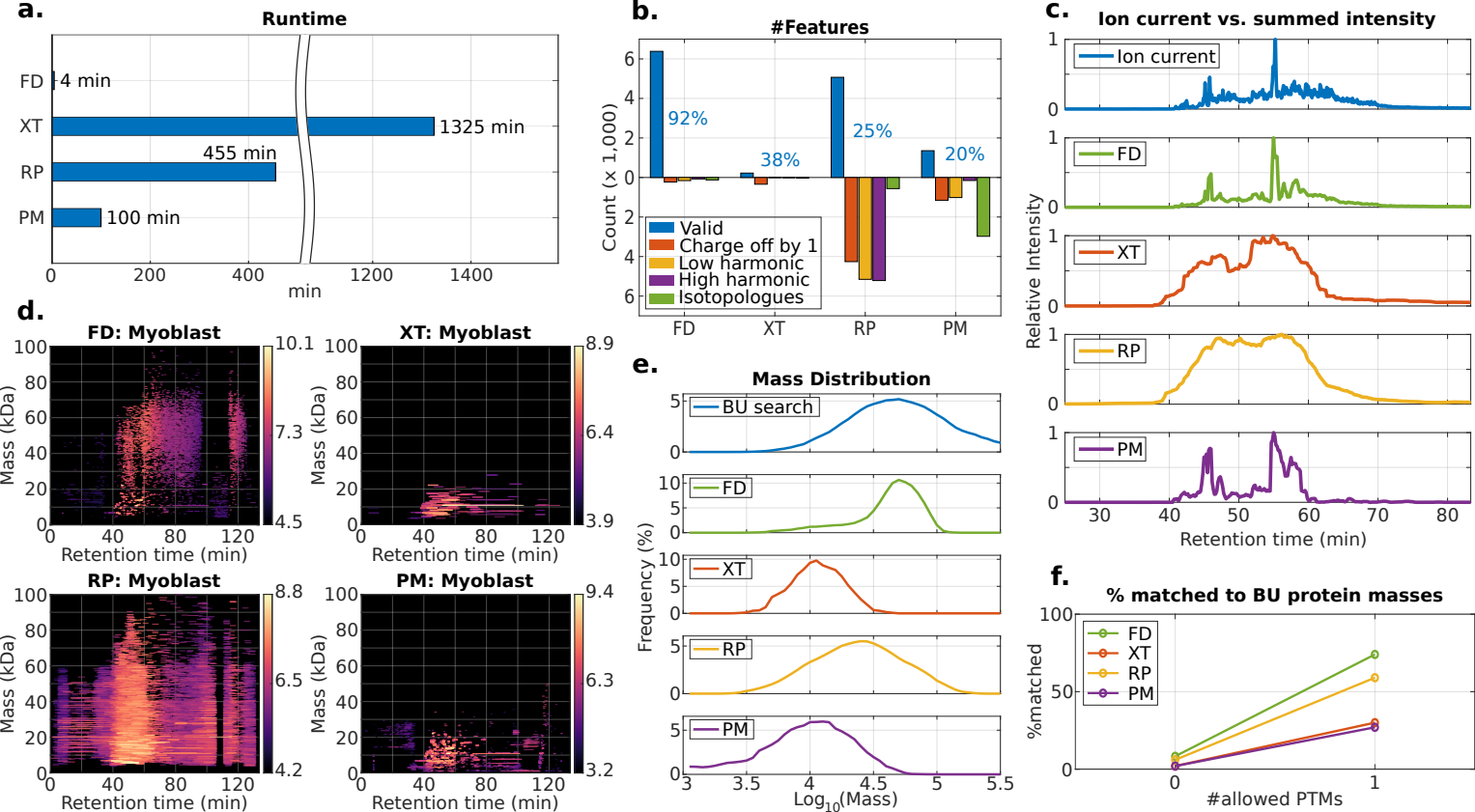
Supplementary Data 3. **The outputs from FLASHDeconv, Xtract, ReSpect, and Promex for the simple data sets (Cyto, Fil, and PIP) technical replicate 2.**

Supplementary Data 4. **The outputs from FLASHDeconv, Xtract, ReSpect, and Promex for the myoblast data set.**

Supplementary Table 1. **Features that matched to the protein masses identified by bottom-up (BU) search for the myoblast data set.**

Supplementary Table 2. **The list of modifications considered in the feature-BU identified protein mass matching.**


a.
FLASHDeconv
Spectral deconvolution
Feature finding
Intact proteins
MS
Intensity
m/z
Retention Time
charge pattern in m/z
charge pattern in log(m/z)
mass M1
∧
mass M2
Multiple patterns
Universal pattern
Intensity
m/z
Transformation
Intensity
log(m/z)
Universal pattern scanning
mass
Retention Time
mass
Retention Time
b.
Runtime
Cyto
Fil
PIP
FD
25 sec
4 sec
5 sec
XT
301 sec
424 sec
RP
1911 sec
PM
242 sec
315 sec
0
500
1000
1500
2000
sec
c.
FD: Cyto
12221.1
XT: Cyto
12221.1
PM: Cyto
12221.2
FD: Fil
18786.6
XT: Fil
18786.6
PM: Fil
18785.6
Intensity
Mass (Da)
d.
FD: PIP
21428.7
28965.2
50431.3
11859.2
9105.5
67961.3
RP: PIP
28981.5
11863.4
21440.7
50459.2
9109.3
Intensity
Mass (Da)
e.
FD: PIP
RP: PIP
Mass (kDa)
Retention time (min)
8.1
6.2
4.3
9.7
6.7
3.7


a.
Runtime
FD
4 min
XT
1325 min
455 min
RP
PM
100 min
0
200
400
1200
1400
min
b.
#Features
Count (x 1,000)
92%
38%
25%
20%
Valid
Charge off by 1
Low harmonic
High harmonic
Isotopologues
FD
XT
RP
PM
c.
Ion current vs. summed intensity
Ion current
FD
XT
RP
PM
Relative Intensity
Retention time (min)
d.
FD: Myoblast
XT: Myoblast
RP: Myoblast
PM: Myoblast
Mass (kDa)
Retention time (min)
e.
Mass Distribution
BU search
FD
XT
RP
PM
Frequency (%)
Log10(Mass)
f.
% matched to BU protein masses
FD
XT
RP
PM
%matched
#allowed PTMs