Comparing Top-Down Proteoform Identification: Deconvolution, PrSM Overlap, and PTM Detection

David Tabb, Kyowon Jeong, Karen Druart, Megan Gant, Kyle Brown, Carrie Nicora, Mowei Zhou, Sneha Couvillion, Ernesto Nakayasu, Janet Williams, et al.

Journal of proteome research

pubs.acs.org/jpr

Article

Comparing Top-Down Proteoform Identification: Deconvolution, PrSM Overlap, and PTM Detection

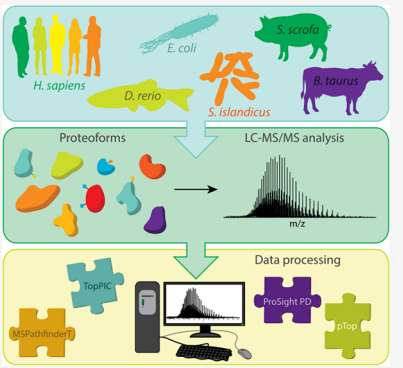
David L. Tabb, Kyowon Jeong, Karen Druart, Megan S. Gant, Kyle A. Brown, Carrie Nicora, Mowei Zhou, Sneha Couvillion, Ernesto Nakayasu, Janet E. Williams, Haley K. Peterson, Michelle K. McGuire, Mark A. McGuire, Thomas O. Metz, and Julia Chamot-Rooke*

Cite This: https://doi.org/10.1021/acs.jproteome.2c00673

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Generating top-down tandem mass spectra (MS/MS) from complex mixtures of proteoforms benefits from improvements in fractionation, separation, fragmentation, and mass analysis. The algorithms to match MS/MS to sequences have undergone a parallel evolution, with both spectral alignment and match-counting approaches producing high-quality proteoform-spectrum matches (PrSMs). This study assesses state-of-the-art algorithms for top-down identification (ProSight PD, TopPIC, MSPathFinderT, and pTop) in their yield of PrSMs while controlling false discovery rate. We evaluated deconvolution engines (ThermoFisher Xtract, Bruker AutoMSn, Matrix Science Mascot Distiller, TopFD, and FLASHDeconv) in both ThermoFisher Orbitrap-class and Bruker maXis Q-TOF data (PXD033208) to produce consistent precursor charges and mass determinations. Finally, we sought post-translational modifications (PTMs) in proteoforms from bovine milk (PXD031744) and human ovarian tissue. Contemporary identification workflows produce excellent PrSM yields, although approximately half of all identified proteoforms from these four pipelines were specific to only one workflow. Deconvolution algorithms disagree on precursor masses and charges, contributing to identification variability. Detection of PTMs is inconsistent among algorithms. In bovine milk, 18% of PrSMs produced by pTop and TopMG were singly phosphorylated, but this percentage fell to 1% for one algorithm. Applying multiple search engines produces more comprehensive assessments of experiments. Top-down algorithms would benefit from greater interoperability.



KEYWORDS: *bioinformatics, deconvolution, identification algorithms, post-translational modifications, proteoforms, top-down proteomics.*

INTRODUCTION

By detecting proteoforms without prior enzymatic digestion,[1] top-down proteomics analysis provides information that is highly complementary to bottom-up methods. Many classes of proteins provide too many or too few basic residues for trypsin digestion.[2,3] When multiple transcripts are possible for a gene, recognizing a particular isoform is simpler when examining the intact proteoform than when glimpsing only a few peptides.[4] When multiple post-translational modifications (PTMs) occupy a proteoform, assignment of PTMs to a specific proteoform is only possible by top-down proteomics, since digestion with a protease generally removes the linkage between different PTM sites.[5] For all their challenges, top-down proteome technologies are therefore essential for proteoform characterization.

Identifying MS/MS produced from intact proteoforms is a mainstay of top-down proteomics. Unlike the identification of proteolytic peptides by bottom-up proteomics, top-down identification relies upon extensive pre-identification signal processing, often called "deconvolution," to represent fragment ions as singly charged or neutral masses and to combine intensity among charge and isotopic variants for each fragment. Spreading the signal for each precursor among many product ions reduces signal-to-noise ratios for fragments, challenging the matching of sequence-predicted fragment masses to MS/MS scans.[6]

The proteoform matching step itself also faces unique challenges. Much longer polypeptide sequences combinatorically expand the number of potential PTM placements on a given sequence. For example, if a proteoform has ten sites that could bear a phosphorylation and the proteoform carries three phosphorylations, there are $_{10}C_3 = 120$ possible localizations to consider. Moreover, many proteoforms arise from the truncation of a base sequence, either through biological means (such as N-terminal methionine removal,[7] signal peptide cleavage,[8] or in vivo protease activity[9]) or by chemical ones (such as a "hot" source causing in-source dissociation[10] or hydrolysis during sample preparation/storage[11]). Allowing for

Received: October 19, 2022

Table 1. Experimental Data Employed in This Data Set[a]

study	instrument	author	experiments	median Z	MS count	MS/MS count
PXD003074-SULIS-ETD	Orbitrap Fusion	Vorontsov	6	11	5818	12,664
PXD003074-SULIS-HCD	Orbitrap Fusion	Vorontsov	6	9	6541	24,015
PXD020342-DANRE	Q-Exactive HF	Xu	18	5	32,505	31,905
PXD019247-ECOLI	Orbitrap Fusion Lumos	Dupré	6	9	27,878	44,609
PXD010825-PIG	Bruker maXis II	Brown	3	14[b]	7402	15,276
PXD019368-HUMAN	Bruker maXis II	Brown	8	21[b]	15,533	30,508
MSV000082070-BOVIN	Bruker maXis I	Vincent	18	15[b]	4662	5565
PXD031744-BOVIN	Orbitrap Eclipse	Zhou	4	11	8567	90,685
PXD031744-BOVIN	Orbitrap Fusion Lumos	Zhou	4	7	10,749	53,632
PXD031744-BOVIN	Q-Exactive HF	Zhou	4	6	17,582	35,281
PXD005420-HUMAN	Q-Exactive	Delcourt	12	7[b]	38,608	58,423

[a]"Experiments" enumerates the LC-MS/MS experiments in each set (CE-MS/MS was employed in PXD020342-DANRE). "Median Z" reports the median of precursor charge among all tandem mass spectra in a set. In PXD031744-BOVIN from the Fusion Lumos, two tandem mass spectra were collected for each precursor ion, one in HCD and one in ETD. [b]For these experiments, +1 precursors were excluded when computing the median.

these potential cleavages implies that top-down identification algorithms must apply "truncation rules" in much the same way as bottom-up identification algorithms apply "proteolytic cleavage rules." To allow the truncation of any number of residues from the N-terminus and any number of residues from the C-terminus in a top-down search is equivalent to a "non-specific" digestion in a bottom-up search. These large search spaces come with a cost in search time and in sensitivity.

Currently available top-down identification workflows differ substantially in their strategies for managing the challenges presented by top-down data. ProSight PD is a contemporary implementation of the 2003 ProSight PTM[12] algorithm built around a Poisson match-scoring model published in 2001.[13] Since its genesis, ProSight PD has sought to leverage protein annotation to curtail the larger search space of top-down identification; UniProtKB XML provides considerable information beyond naked sequences, such as observed or predicted signal peptide cleavage sites, coding SNPs, and modified residues with known mass shifts. MetaMorpheus, a search engine that operates in both top-down and bottom-up modes,[14] also makes use of UniProtKB XML databases but adds the capability to supplement UniProtKB PTMs with those discovered through bottom-up identifications in the same sample.

The spectral alignment algorithm matches deconvolved fragment ions to sequence-derived mass ladders, treating regions bearing PTMs as mass gaps in these alignments. It was originally deployed in top-down identification with MS-TopDown in 2008[15] and was followed by the publication of MS-Align+ in 2012.[16] Three contemporary top-down identification algorithms have built upon these foundations. The 2016 TopPIC Suite[17] was created with spectral alignment at its core, featuring built-in detection of mass shifts that are not pre-specified by the user. Also in 2016, the pTop software[18] (part of pFind Studio: http://pfind.org) incorporated machine learning to improve precursor deconvolution. MSPathFinderT was introduced in 2017 as part of the Informed Proteomics Suite,[19] incorporating HUPO-PSI formats for reading spectra (mzML[20]) and writing identifications (mzIdentML[21]). 2017 also saw the publication of the TopMG software,[22] added to TopPIC Suite to identify more heavily modified proteoforms.

In practice, most top-down laboratories develop expertise in a single top-down search engine and do not consider others. One of the key questions driving this study is how much more information might be obtained from existing data if more than one identification approach is used. The four algorithms compared here are: MSPathFinderT from Informed Proteomics, ProSight PD, pTop, and TopPIC. They can either be operated using included graphical user interfaces (all but MSPathFinderT) or via MASH Suite,[23] a universal interface for top-down proteomics software.

Previously published comparisons between proteoform identification software have generally been part of papers that introduce a new algorithm;[16−19,22] as a result, they may have inadvertently featured data types or evaluation methods that show the new algorithm in a disproportionately positive light. In this study, data from ThermoFisher Orbitrap-class and Bruker maXis Q-TOFs are used to compare the performance of four top-down proteomics software workflows. Various datasets from ThermoFisher instruments were used, including *Sulfolobus islandicus* (*S. islandicus*) and *Escherichia coli* (*E. coli*) cell lysates, brain tissue from male and female *Danio rerio* (*D. rerio*), *H. sapiens* serous ovarian tumors and de-fatted milk from *Bos taurus* (*B. taurus*) with clinical mastitis. To compare the performance for top-down deconvolution in data from Bruker Q-TOF instruments, this study relied on the analysis of data from *B. taurus* milk (maXis I), *Sus scrofa* (*S. scrofa*) cardiac material, and *Homo sapiens* (*H. sapiens*) embryonic kidney samples (both maXis II). With these datasets, the estimated PrSM yield and orthogonality of each search engine were assessed. Using this information, we recommend a multialgorithm approach for top-down proteoform identification and urge top-down software developers to facilitate the interoperability for identification software components.

EXPERIMENTAL PROCEDURES

ThermoFisher Orbitrap Data Sets

PXD003074 SULIS. Vorontsov et al. characterized lysine methylation and N-terminal acetylation of *S. islandicus* LAL 14/1 via top-down and bottom-up techniques on a ThermoFisher Orbitrap Fusion Tribrid instrument (Table 1).[24] The proteome was separated by strong anion exchange into six fractions analyzed once in ETD and once in HCD. The six ETD fractions included in this analysis (see Supporting Information Tables S1−S3 for a complete list of raw files for this and other sets) averaged 2111 MS/MS scans per raw file, while the six HCD fractions were approximately double that number, averaging 4003 MS/MS scans per raw file. The ETD data yielded an interquartile range of precursor charges from +8 to +16, while the HCD data yielded +7 to +14. The maximum precursor

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

charge detected in ETD was +50, while HCD rose to +44. The sequence database combined UniProtKB reference proteomes UP000013006 (*S. islandicus* LAL14/1: 2591 proteins) and UP000000625 (*E. coli*: 4438 proteins). The *E. coli* proteins, comprising 63% of all entries, were included as unlabeled decoys, since no *E. coli* was expected to be present in the samples.

PXD019247 ECOLI. Dupré et al. analyzed *E. coli* lysates to optimize the detection of pathogen-specific proteoforms on a ThermoFisher Orbitrap Fusion Lumos.[25] The six EThcD LC-MS/MS replicate experiments analyzed for this study averaged 7435 MS/MS scans per raw file. The interquartile of precursor charges from TopFD deconvolution ranged from +5 or + 6 to +14 (with a maximum Z of +41). The same sequence database was used for this data set as for the *S. islandicus* data above, with the SULIS sequences representing unlabeled decoy sequences.

PXD020342 DANRE. Xu et al.[26] employed capillary IEF-MS/MS rather than RPLC-MS/MS to investigate sexual dimorphism in brain tissue of zebrafish (*D. rerio*).[27] The four size exclusion chromatography fractions for each sex were analyzed in triplicate to produce a total of 24 raw files on a Q-Exactive HF. We omitted the first fractions from consideration because they contributed few PrSMs. Unlike the other studies, the zebrafish data incorporated iodoacetamide to alkylate cysteines after the reduction of disulfides. The 18 HCD experiments averaged 1576 MS/MS scans per raw file for female samples and 1969 for male samples. After TopFD processing, the precursor charge interquartile range spanned from +4 to +8 for female brain samples and from + 4 to +12 for the male samples. The UniProtKB zebrafish reference proteome, UP000000437, contained 47,204 sequences when downloaded during September, 2021. An additional 27 raw files under the same PXD accession examined similar samples by bottom-up methods, which enabled the creation of a parsimonious subset database via MSFragger 3.0[28] and IDPicker 3.1[29] containing 9432 sequences.

PXD031744 BOVIN. To create reference spectra for a phospho-proteoform mixture, we analyzed four samples of defatted *B. taurus* milk from an animal with clinical mastitis. One pair of "clinical quarter" and "healthy quarter" samples were collected both 13 days postpartum and 16 days postpartum. The "C13" (clinical, 16-day) sample was unusual for its inclusion of high somatic cell counts. Each of the four samples was analyzed in 100 min LC-MS/MS experiments on three ThermoFisher instruments: Orbitrap Eclipse, Q-Exactive HF, and Orbitrap Fusion Lumos, employing HCD for the first two instruments and a mix of HCD and ETD for each precursor on the last. The Orbitrap Eclipse samples were membrane-filtered, while the other instruments saw samples that had been ultracentrifuged. After TopFD deconvolution, the Orbitrap Eclipse set yielded an interquartile precursor charge range of +6 to +20, with 29 spectra reaching the maximum of +60 charge. The Fusion Lumos precursor charge interquartile spanned +5 to +10 (maximum of +44), while the Q-Exactive HF interquartile range was +4 to +10 (maximum of +56). Four additional experiments from the Q-Exactive HF and four from the Orbitrap Fusion Lumos collected MS scans at low resolution to improve sensitivity for larger masses; these are included in the PXD repository though they are not analyzed here. A Q-Exactive HF-X analyzed twenty-four fractions of a bottom-up TMT multiplex set spanning these and other milk samples. More detailed methods may be found in Supporting Information Text S1. The complete UP000009136 reference proteome contains 37,883 proteins. It was reduced to a subset FASTA of 5388 sequences via MSFragger and IDPicker 3.1 using the bottom-up TMT data. This database was used for a top-down search with ProSight PD, TopPIC, pTop, and MSPathFinderT, and any accession matched to any spectrum with high confidence by one of those four engines was retained to produce an accession list of 378 entries. The UniProt Subsetter described below was used to harvest this set of sequences to FASTA and UniProtKB XML databases used in the PTM search by all six algorithms (TopMG joining TopPIC, and ProSight PD employing both FASTA and XML inputs).

PXD005420 Human. Delcourt et al. evaluated human serous ovarian tumors through liquid micro-junction (LMJ) and parafilm-assisted microdissection (PAM).[30] The samples included necrotic/fibrotic tumor, tumor, and benign regions of interest. Our PTM characterization included only the tumor and benign LMJ and PAM analyses in technical triplicate, yielding a total of twelve Q-Exactive HCD raw files split among four samples: LB (LMJ-Benign), LT (LMJ-Tumor), PB (PAM-Benign), and PT (PAM-Tumor). The PAM analyses averaged more MS/MS (5728) than did the LMJ analyses (4009). Unlike the other ThermoFisher data sets, the TopFD deconvolutions of these files contained large numbers of singly charged precursor ions. After excluding those, the four samples yielded a charge interquartile range from +4 or +5 to +10, with a maximum overall of +28. We took advantage of the rapid search engines pTop and ProSight PD to shrink the full-size human reference database (UP000005640: 101,014 entries) to one containing just the accessions matched to any PrSM by either algorithm (558 entries). This reduced database was then used to support a PTM search allowing for up to four PTMs per PrSM.

Bruker Q-TOF Data Sets

PXD010825 PIG. Brown et al. identified proteins in *S. scrofa* cardiac material in evaluating the "Azo" surfactant for use in top-down proteomics on a Bruker maXis II instrument.[31] Raw directories containing "analysis.baf" files are available in ProteomeXchange as PXD033208. The three LC-MS/MS experiments were renamed PigHeartAzo-20170429, PigHeartAzo-20171029, and PigHeartAzo-20171107 to reflect their run dates. The first experiment ran 95 min, while the latter two ran 75 min. After conversion to mzML format in ProteoWizard 3.0,[32] the three experiments contained 5598, 5155, and 4523 MS/MS scans, respectively. Deconvolved msAlign files were created through a Visual Basic Script for Bruker DataAnalysis 5.3 described below, yielding precursor charge interquartile ranges of +8 to +23 for the first experiment, +8 to +25 for the second experiment, and +8 to +27 for the third. (The charge distributions reported here for Bruker experiments exclude +1 precursors.) The UP000008227 reference proteome for pig contained 49,792 proteins when isoform variants were incorporated. We derived a subset database by using a bottom-up proteome from the same laboratory (MSV000080621[33]) to find the parsimonious subset of these sequences that could be matched confidently to peptide-spectrum matches in MSFragger search followed by IDPicker protein assembly. The subset sequence database contained 3974 sequences.

PXD019368 HUMAN. Brown et al. evaluated cloud point extraction and size exclusion chromatography in human embryonic kidney (HEK293T) cells on a Bruker maXis II instrument.[34] Raw directories containing "analysis.baf" files are available in ProteomeXchange as PXD033208. The eight LC-

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

MS/MS experiments were divided equally into Tergitol and Triton cohorts. The Tergitol cohort contained fewer tandem mass spectra (14,365) than did the Triton cohort (16,143). The first two fractions for both cohorts contained far fewer tandem mass spectra (7674) than did the latter two fractions (22,834). Excluding singly charged precursors, the msAlign files created through Bruker DataAnalysis yielded an interquartile precursor charge range from +11 to +27. In order to create a subset database for the identification of these proteoforms, we employed the data set PXD017858,[35] searched in MSFragger and assembled in IDPicker, to produce a sequence database of 15,109 entries.

MSV000082070 BOVIN. Vincent et al. evaluated the Bruker maXis I instrument for LC-MS/MS phosphoproteomics using bovine milk samples.[36] We selected the "Holstein" and "Jersey" milk experiments from the raw data folders at MassIVE, excluding instrument methods 1, 4, and 12. These 40-min LC-MS/MS experiments averaged 309 MS/MS apiece, with Method 2 producing a low of 135 MS/MS for both types of milk and Method 11 producing a high of 365 MS/MS for both types of milk. Excluding singly charged precursors, the msAlign files created by Bruker DataAnalysis yielded an interquartile precursor charge range from +4 to +23. A 2016 bottom-up proteome by the same author from bovine milk (PXD002529[37]) was analyzed by MSFragger and IDPicker to generate a FASTA of 276 protein sequences that could then be used to identify proteoforms in the top-down set.

Top-Down Identification Workflows

We selected four top-down search engines that were capable of truncation searches and could be installed for use on local computers (see Figure 1 and Table 2). Algorithms that did not completely fulfill these criteria (Perceptron,[38] Mascot Top-Down,[39] MetaMorpheus,[14] and PIITA[40] were not included in our comparison. Perceptron requires a high-end nVidia graphics card and several development kits for installation, which were not available to us. We attempted some Mascot Top-Down searches but were held back by having licensed a relatively old version of Mascot Server. MetaMorpheus was very close to inclusion but its truncation search was not completely implemented when searches were performed. Unfortunately, it appears the source code for PIITA is no longer retrievable.

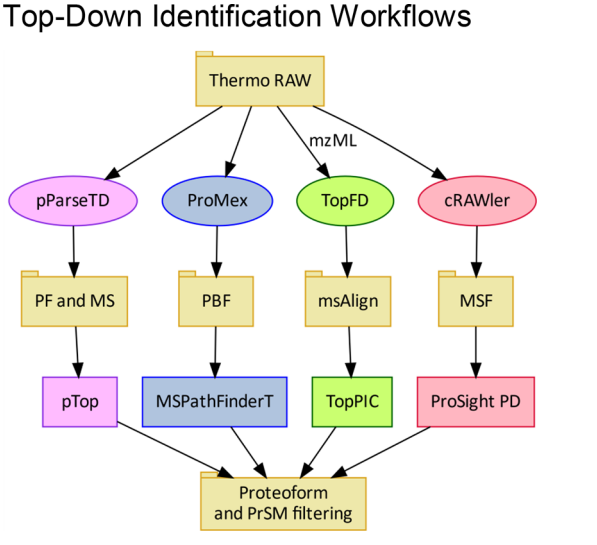


Figure 1. Four different software packages were used for PrSM identification. Each identification workflow is represented by a deconvolution engine (depicted as a colored oval) that outputs deconvolved MS/MS in a variety of formats (depicted as a file folder), which are then subjected to identification by a search engine (depicted as a colored box). The PrSMs produced by these four workflows could then be filtered to confident IDs and then reported in a common format for comparison.

Search engine outputs, configurations, and ProForma-formatted PrSM lists can be downloaded in Supporting Information File S1. ProForma is a standard proteoform annotation to communicate sequences,[41,42] PTMs, and localization data for a wide variety of modifications and crosslinks.

ProSight 4.0 in Proteome Discoverer 2.5. All searches employed the "PSPD Truncation Search with FDR" and "PSPD No FDR Consensus" workflows. The "ProSight PD High/High cRAWler" node deconvolved spectra through the Xtract algorithm. The cRAWler applies a maximum precursor charge of 80 and a maximum fragment charge of 30. The Subsequence Search Node was configured to allow up to 2 or 4 Maximum PTMs per isoform (under advanced settings) with no static modifications, and the mass tolerance default of 10 ppm for fragments was accepted without change. The precursor tolerance, however, was increased from 10 ppm to 1.1 Da, increasing both identification yield and the time required for each search. In the most recent version of the software, ProSight 4.2 for Proteome Discoverer 3.0, a comparable search would instead retain the 10 ppm precursor tolerance, supplementing it by setting the "Number of Off by 1's" to allow a single neutron slip. This newer approach is expected to reduce potential false positives that may be matched with high mass error tolerance. PrSM tables were exported from Proteome Discoverer in text (tab-delimited) format with three alterations from the default settings: "External Top-Down Displays" was disabled, while "Fragmentation Scan(s)" and "Original Precursor Charge" were enabled. Other than the default N-terminal acetylation, variable modifications were configured in the Database Manager at the time of FASTA or XML import.

TopPIC Suite 1.4.13. Because TopIndex requires considerable storage space when handling eukaryotic databases, we employed only TopFD (feature detection/deconvolution) and TopPIC (search engine) from the TopPIC Suite. PXD031744-BOVIN and PXD005420-HUMAN also employed the TopMG algorithm for PTM searching. TopFD deconvolved MS/MS from mzML format to msAlign format, using the "--skip-html-folder" option to curtail export of text files to represent each MS/MS. TopPIC matched proteoforms to tandem mass spectra in msAlign format, using the "--skip-html-folder" option to prevent writing text files to represent each PrSM, the "--combined-file-name" option to produce conjoint reports for each set of input files, and the "--num-shift" option to specify one gap in spectral alignment for PTM insertion (except in the PTM searches, where two gaps were allowed). The "--mod-file-name" option defined dynamic modifications in generating proteoforms as specified below in "PTM Handling." The default mass tolerance of 15 ppm was employed for both precursors and fragments (attempts to use 10 or 20 ppm instead resulted in less than a 2% difference in identified PrSMs). Default filtering of PrSMs and proteoforms retained only those with E-values below 0.01. PrSM tables were imported from the conjoint reports ending in "_ms2_toppic_prsm.tsv." PrSMs containing unlabeled mass shifts were retained only for the deconvolution study. In all other tests, the "--suppress" option in the ProForma exporter

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 2. Search Algorithms Employed with Maximum Charges and Mass Tolerances

algorithm	version	max Z	pre. tol.	frag. tol.	algorithm-specific
ProSight PD	4.0 for PD 2.5	80 (MS1)/30 (MS2)	1.1 Da	10 ppm	truncation search with FDR
TopPIC	1.4.13	30 (MS1 + MS2)	15 ppm	15 ppm	--num-shift 1
MSPathFinderT	1.1.7867	50 (MS1)/20 (MS2)	10 ppm	10 ppm	-ic 1 -tagSearch
pTop	1.2/2.0pre	30 (MS1 + MS2)	5.2 Da	15 ppm	use of pre-release builds

(described below) was employed to eliminate these from consideration.

Informed Proteomics 1.1.7867. We first used PBFGen to produce a PBF binary file optimized for chromatogram extraction from each raw file. ProMex was then applied to deconvolve the LC-MS features at different charge states, with a ceiling of +60. These first steps could be run without additional options because their default values were suitable for ThermoFisher FT-class experiments. MSPathFinderT handled MS/MS deconvolution and proteoform matching for Informed Proteomics, applying an additional charge ceiling of +50 for precursor ions and +20 for fragment ions. It was run in "-ic 1" mode, implying that only proteoforms representing a single cut in the peptide backbone would be identifiable (i.e. proteoforms that represented internal polypeptides, where both N-terminus and C-terminus were truncated could not be identified). The "-tda 0" option specified a search of target sequences only. Precursor and fragment mass tolerance defaulted to 10 ppm (experiments at 20 ppm yielded very similar numbers of PrSMs). The "-mod" option specified the modifications listed below in "PTM Handling" for each experiment, imposing a maximum of three PTMs per proteoform, except that a limit of four PTMs was employed for PTM searches of PXD031744 and PXD005420. Unlike TopPIC or ProSight PD, MSPathFinderT requires the N-terminal acetylation to be explicitly specified as a PTM, and the software accepts chemical compositions for its PTMs rather than explicit masses. PrSM tables were imported from the set of "_IcTda.tsv" reports, one for each input raw file. FDR filtering was handled inside the ProForma exporter described below, limiting to an E-value of less than 0.01 and requiring reported PrSM probabilities to be greater than 0.5.

pTop 2. pTop operation produced interesting challenges because the pParseTD incorporated in the pre-release pTop 2[43] produced sparse MS/MS peak lists, and the pTop search engine in pTop 1.2 failed to produce reasonable sensitivity in the PXD020342-DANRE and PXD031744-BOVIN sets. After configuring searches in the pTop 2 GUI, we executed the pParseTD.cfg using the pParseTD.exe from pTop 1.2 and executed the pTop.cfg using the pTop.exe from a pre-release of pTop 2. In all cases, we deactivated "Mixture Spectra" to prevent individual MS/MS scans from being assigned to multiple proteoforms. Deconvolution limited precursor charge to +30, using the default for this software. The pTop 2 software defaulted to 15 ppm fragment tolerance for fragments and a wide 5.2 Da tolerance for precursor masses. Like MSPathFinderT, pTop required that protein N-terminal acetylation be configured as a dynamic modification; the software does not anticipate this modification automatically. We limited PTMs per PrSM to four and used the default FDR threshold of 1%, filtering all PrSMs for a set conjointly rather than per raw file. PrSM tables were imported from the pTop_filtered.csv report. The PXD019247 experiment produced an unusual error for pParseTD. The deconvolution engine recognized the dissociation type as ETD (the experiments were EThcD, but pParseTD 1.2 does not support EThcD fragmentation), but the software misinterpreted the raw file to perceive the data as "ETDIT" (ion trap measurement) rather than "ETDFT" (Orbitrap measurement). We simply changed the file names to ETDFT, and then the searching proceeded normally.

PTM Handling

In addition to protein N-terminal acetylation and Met oxidation, the sets of dynamic modifications employed in searching each sample included the following:

- PXD003074 SULIS: Lys + C_1H_2
- PXD019247 ECOLI: (none)
- PXD020342 DANRE: Cys + $C_2H_3N_1O_1$
- PXD010825 PIG: (none)
- PXD019368 HUMAN: (none)
- MSV000082070 BOVIN: Ser + $H_1O_3P_1$, Thr + $H_1O_3P_1$, Tyr + $H_1O_3P_1$
- PXD031744 BOVIN: Ser + $H_1O_3P_1$, Thr + $H_1O_3P_1$, Tyr + $H_1O_3P_1$
- PXD005420 HUMAN: Ser + $H_1O_3P_1$, Thr + $H_1O_3P_1$, Tyr + $H_1O_3P_1$

For the final two sets, when ProSight PD was run with a UniProtKB XML rather than a FASTA, only Met oxidation was added; phosphorylations could only be considered if they were annotated in the UniProtKB XML.

Deconvolution Algorithms

As shown in Figure 1, each of the search engines comes prepackaged with its own deconvolution engine. It is not possible to mix-and-match most of these deconvolution and search tools. The output from pParseTD, for example, is not formatted as it would need to be for use in MSPathFinderT, TopPIC, or ProSight PD. Because TopPIC accepts deconvolved spectra in msAlign, a simple text format derived from Mascot Generic Format,[44] we sought to characterize the variability introduced by five deconvolution engines that could be exported to msAlign or MGF format (see Figure 2). We emphasize deconvolution in the context of Bruker maXis I and II data because Q-TOF instruments have not been as widely supported by software tools for top-down proteomics.

The test of deconvolution impact encompassed three Bruker Q-TOF datasets (PXD010825-PIG, PXD019368-HUMAN, and MSV000082070-BOVIN), as well as one ThermoFisher Orbitrap data set (PXD019247-ECOLI). Three deconvolution engines could be used on all sets (TopFD, FLASHDeconv, and Mascot Distiller), while two could be used only on data from a particular instrument vendor (ThermoFisher Xtract in ProSight PD or AutoMSn in Bruker DataAnalysis). TopFD and FLASHDeconv accepted data in peak-listed mzML format, writing their output to msAlign files; these could then be identified in the TopPIC identification engine. Mascot Distiller and ProSight PD started from raw mass spectrometry data, recording their outputs to Mascot Generic Format (MGF). The simple MGF2msAlign utility described below could then produce TopPIC-ready msAlign files. A Visual Basic Script for Bruker DataAnalysis (described below) performed feature detection from the raw Q-TOF data, deconvolved MS/MS

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

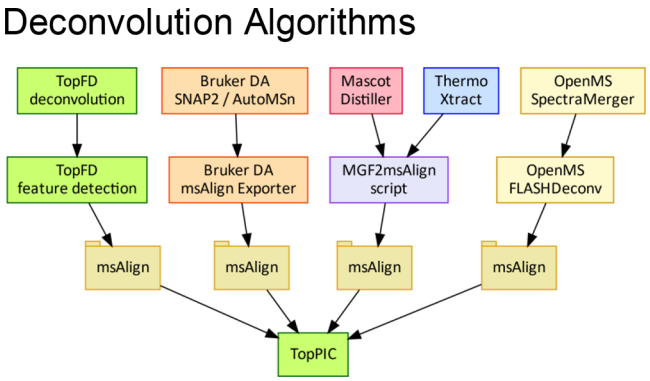


Figure 2. In order to characterize variation introduced by deconvolution, we generated output from five different deconvolution pathways for the same input data, directing these outputs to msAlign format for identification by TopPIC. Bruker maXis data were processed by all pipelines except for ThermoFisher Xtract, and ThermoFisher Orbitrap data were processed by all pipelines except for Bruker DataAnalysis.

peak lists, and exported msAlign files. In all cases, TopPIC 1.4.13 identified PrSMs from these msAlign deconvolved peak lists. All four sets of msAlign files for the two Bruker maXis II data sets have been posted to ProteomeXchange PXD033208 to accompany their raw data.

FLASHDeconv. FLASHDeconv (2.0 beta version)[45] was used to deconvolute all three Bruker datasets and the ThermoFisher PXD019247-ECOLI experiments. All raw files (".d" directories for Bruker and ".raw" files for ThermoFisher datasets) were centroided with MSConvert (version 3.0.20186), employing the vendor peakPicking filter, to produce input mzML files. For all datasets, the allowable charge range was configured to be 2−70, and mass range spanned from 50 to 60,000 Da. FLASHDeconv discarded MS2 spectra with precursors of high interference level (by applying min_precursor_snr threshold of 1.0). To maximize the signal-to-noise ratio of tandem mass spectra in all datasets, Gaussian weighted moving averaging was carried out using the SpectraMerger tool in OpenMS software.[46] Briefly, an MS/MS scan is collected with a precursor m/z value of x and inferred charge of c. The software computes the neutral mass of the precursor by subtracting a proton and multiplying by charge: $c(x\text{-}1.00727647)$. These neutral masses can then be compared among MS/MS scans to determine which have neutral masses within a precursor mass threshold (specified as 10 ppm for these experiments). Then the collected MS/MS scans are multiplied with Gaussian weights centered at the input MS/MS and summed to generate the output averaged MS/MS. Averaging is performed automatically within FLASHDeconv by setting the -merging_method option to 1. The version of FLASHDeconv employed in the time trial was from a pre-release build of Open MS 3.0, built on February 16, 2023.

Mascot Distiller 2.8.0.1. Mascot Distiller[47] was able to deconvolve tandem mass spectra from the Bruker Q-TOF datasets (PXD010825 PIG, PXD019368 HUMAN, and MSV000082070 BOVIN); its results on PXD019247 ECOLI, however, were of insufficient PrSM yield to be comparable to other deconvolution paths. The software operated on the Bruker ".d" directories containing "analysis.baf" files (made available at PXD033208). It was configured to allow a maximum charge of +40 (the highest value available for this software), with a default charge range of +10 to +20. Output was specified to be "fragment ions in MS/MS peak lists as MH+." See configuration options in Supporting Figure S1.

ThermoFisher Xtract in Proteome Discoverer 2.5. To generate an msAlign that represents Xtract deconvolutions, we opened each pdResult file representing a completed ProSight PD run to export complete peaklists (not just the spectra that were successfully identified) to MGF format (the File → Export → Spectra menu item). The MGF2msAlign script described below then converted the MGF file to msAlign. The maximum charge configured by the "High/high cRAWler" was +80.

Bruker DataAnalysis 5.3 msAlign Exporter. ProteomeXchange PXD033208 holds the msAlign files exported by a Visual Basic Script in Bruker DataAnalysis. Scripts are available via the GitHub named at the top of the "Formatting and Exporting Tools" section below or via https://github.com/dtabb73/Bruker-msAlign-exporter. Screenshots showing the configuration of the SNAP peakfinder and the AutoMSn compound detector appear in Supporting Figure S2. AutoMSn constructs a set of "Compounds" that combine information across a small range of retention times and that explain isotopic packets appearing across a small range of m/z values. The tandem mass spectra mapping to each compound are combined prior to deconvolution, yielding a smaller number of spectra that are higher in quality from summing signals together.

TopFD 1.4.13. The deconvolution engine in TopPIC Suite is TopFD ("feature detection)." TopFD attempts to include every MS/MS it receives as a deconvolved MS/MS in its output, and so spectra that have ambiguous or missing precursor ions will be retained by TopFD where they may be excluded by other workflows. It limits each output MS/MS to a single precursor charge state, in indeterminate cases recording a charge of 0. For the deconvolution tests, TopFD was allowed to infer up to a ceiling of +60 charge, but in all other cases, its default ceiling of +30 was retained.

UniDec 6.0.0.b3. The UniDec MS1 deconvolution engine has gained popularity for native and denatured proteoform data in recent years.[48] It was included only in time trials because it does not currently support MS/MS deconvolution. Algorithm defaults were employed in a test script supplied by the Michael Marty, allowing a ceiling of +100 for ions in the MS scans.

ThermoFisher Xtract in FreeStyle 1.8 SP2. Starting with version 1.8 of FreeStyle,[49] the "Xtract All" option can automatically deconvolve all MS and MS/MS scans in a raw file. It was included only in time trials because it did not support easy integration with the TopPIC search engine. The signal-to-noise requirement was reduced from a threshold of 3.0 to 1.0. The required number of detected charges was reduced to 1, and the charge range under consideration was set to +1 through +50.

Formatting and Exporting Tools

This study could only be conducted by creating support tools for reformatting PrSM outputs, extracting subsections of databases, and adapting between spectrum formats. All these tools have been made available under a common URL: https://gitlab.pasteur.fr/MSBio/TDP_comparative_tools.

ProForma Exporters. We created four software utilities in the C# language to export the following fields from each table of PrSMs: experiment filename, scan number, PrSM precursor charge, stripped UniProtKB accession, truncated sequence for PrSM without PTMs, integer sum of rounded PTM masses ("MassAdded)," ProForma v2 string representing sequence and PTMs,[42] and negative log E-Value for PrSM. These tools are available at the above GitLab hub or via https://github.com/

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

dtabb73/ProForma-Exporters. The tables output by these tools were the basis for estimating PrSM yield and for overlap assessments that comprise the bulk of this work. Some rows of ProSight PD PrSMs resulted in multiple rows printed to the ProForma reports. These cases represented proteoforms that were matched to multiple tandem mass spectra combined before the search. The utility applied the "high" criterion for ProSight PD, filtering PrSMs to include only those with a negative log E-value of 5 or higher (this provided better sensitivity at acceptable FDR since targeting a 1% FDR in the Consensus workflow produced an empirical FDR closer to 0%). MSPathFinderT output was filtered in the utility to limit PrSMs to an E-value below 0.01 and a reported probability above 0.5. Of the four algorithms, only TopPIC routinely reported ambiguity of PTM localization and unknown mass shift modifications; the format converter forced placement of the PTM at only the most C-terminal position to make the output more comparable to the other search engines. The utility's "--suppress" option for TopPIC was used throughout to remove all PrSMs containing any unknown mass shift (resulting in a substantial sensitivity reduction for TopPIC: see Supporting Information Text S2).

UniProt XML Subsetter. While many tools exist to choose a subset of FASTA sequences from a comprehensive proteome database, we needed to select a subset of entries from a UniProtKB XML. We therefore developed a Python script in two different modes: (1) to extract XML sequences corresponding to those in a subset FASTA file or (2) to extract both XML sequences and FASTA sequences based on a provided list of UniProtKB accessions.

MGF2msAlign. Since the MGF files from Xtract and Mascot Distiller contain singly charged monoisotopic masses for fragment ions, we created a script to convert these ion masses into uncharged monoisotopic masses by subtracting the mass of a proton from them. Normally, MGF files record precursor charge and monoisotopic precursor m/z values in the header for each MS/MS. Computing an uncharged intact mass for the proteoform (a necessary field in the msAlign format) required that we subtract the mass of a proton from the precursor m/z and then multiply the result by the precursor charge.

RESULTS AND DISCUSSION

The assessments in this study were intended to characterize the current state-of-the-art for top-down identification algorithms for deconvolution, PrSM yield under controlled FDR, overlap of proteoform identification, and detection of post-translational modifications. Our investigations spanned eight data sets, with each undergoing different iterations of the following top-down identification engines: ProSight 4.0 in Proteome Discoverer 2.5, TopPIC Suite 1.4.13, Informed Proteomics 1.1.7867, and pTop 2 (July 2, 2017, build of pTop.exe).

Making comparable identifications from these four search engines required many steps that we automated through support tools (see Formatting and Exporting Tools section). The tests of identification sensitivity, FDR control, and overlap required that we be able to compare identifications directly. For example, determining that a particular MS/MS identified by TopPIC matches an identification from pTop requires that we can compare the raw file name and scan number attributed to those PrSMs, and each of those tool's records that information in a different format. The search engines also abbreviate UniProtKB accessions differently. The formats by which they communicate proteoforms (implying sequence truncations, the identities of PTMs, and their localizations) are inconsistent. Finally, MSPathFinderT in Informed Proteomics outputs all PrSMs to its output tables, so filtering criteria must be applied to those PrSMs before comparison to other search engines. We created ProForma Exporters in the C# programming language to output identification data from all four search engines in identical formats, using the ProForma v2 specification from HUPO-PSI to communicate proteoforms.

Identification Yield and FDR

Proteomics users often use the terms "sensitivity" and "specificity" to explain the goals they want to achieve in identifying MS/MS scans produced by their experiments. From a statistical point of view, sensitivity can only be calculated if we know the identities of all spectra that were correctly identified, both above and below the threshold. Computing specificity requires that we know the identities of all spectra that were incorrectly identified, both above and below the threshold. Instead, we seek to maximize the estimated number of true positives from a search (the "yield") while holding the FDR to a pre-specified ceiling. In the case of this study, we sought to identify as many spectra as possible while producing an FDR that was controlled to approximately 1% (software typically yields far fewer proteoform-spectrum matches if the thresholding attempts to eliminate all false matches).

As bottom-up proteome informatics developed during the first decade of the 2000s, researchers often turned to data sets derived from mixtures of known proteins to tune software to separate good peptide-spectrum matches from bad. The "ISB-18," for example, was a mixture of 18 purified proteins that had been analyzed in many replicates for eight different mass spectrometers.[50] The "Aurum" set from the University of Michigan was created to help tune the identification of MALDI-TOF/TOF tandem mass spectra.[51] The top-down proteomics community has begun assembling resources of this type, as well. The 2014 pilot project by the Consortium for Top-Down Proteomics, for example, incorporated data from seven different laboratories for human histone H4.[52] Defined protein mixtures are again being analyzed as a way to assess CZE separation and Q-TOF instrumentation for top-down technologies.[53] Such data would seem ideal for testing identification algorithms, but they do come with drawbacks. Even if the proteins included in a mixture are known, the proteoforms representing each protein may remain uncertain. Moreover, defined mixtures intended to contain a dozen proteins sometimes contain twice as many proteins as expected due to accompanying impurities.[54] Even a reference protein mixture, such as the NCI-20 or Sigma UPS1,[55] however, will not generate as large a diversity of proteoforms as a biofluid, let alone a cell lysate. In stepping away from defined mixtures, we lose the ability to say which spectra were correctly identified and which were incorrectly identified, irrespective of thresholds (and thus computing sensitivity or specificity), but we greatly increase the diversity of spectra to be identified in each experiment.

Most top-down identification algorithms were originally developed and tuned for handling ThermoFisher Orbitrap-class data. The data sets used here to evaluate these identification algorithms, include 6 ETD and 6 HCD LC-MS/MS experiments for *S. islandicus* (PXD003074-SULIS), 6 EThcD experiments for *E. coli* (PXD019247-ECOLI), and the 18 HCD runs for *D. rerio* (PXD020342-DANRE) (see Table 1). The *D. rerio* set differed from the *S. islandicus* and *E. coli* sets in more ways than just the complexity of its genome. It was the only experiment that employed iodoacetamide to alkylate Cys

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

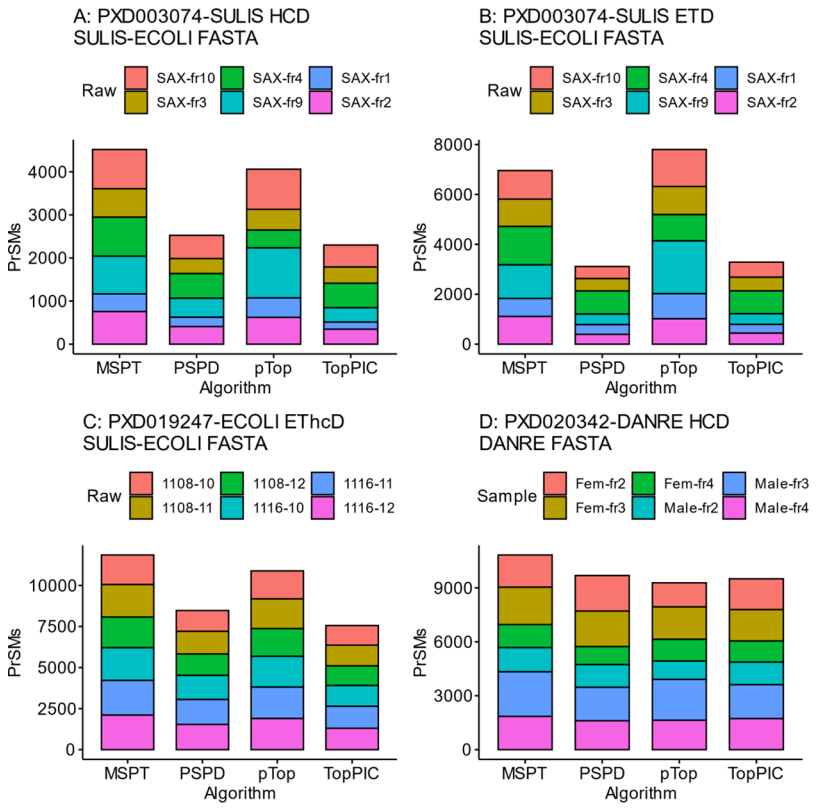


Figure 3. In panels A−C, each color signifies the identifications yielded from one of six different raw files. ProSight PD has been abbreviated "PSPD," while MSPathFinderT (part of the Informed Proteomics suite) has been abbreviated "MSPT." In PXD020342-DANRE, the triplicates for each fraction of each sex were combined to a single color (18 raws become six samples). The ability to produce many PrSMs for a given raw file varies due to many factors, particularly in how one threshold "good" from "bad" identifications. The PrSM yield for TopPIC reflects the removal of PrSMs containing unanticipated PTMs (this effect is quantified in Supporting Information Text S2).

residues, it was the only one to employ CE rather than LC upstream of the instrument, and it was generated by a ThermoFisher Q-Exactive HF while the other two were produced in either an Orbitrap Fusion (PXD003074-SULIS) or an Orbitrap Fusion Lumos (PXD019247-ECOLI).

The four identification algorithms (see Figure 1) can be considered two pairs; ProSight PD and pTop are both configured and launched from their integrated Microsoft Windows graphical user interfaces, and both employ a target-decoy strategy for estimating the false discovery rate of the PrSMs. TopPIC and MSPathFinderT both implement a mass spectrum alignment strategy for finding mass shifts that are interpreted through lists of enumerated PTMs, and both directly estimate the expectation value of the best match for each MS/MS by an adaptation of the generating function strategy embodied in the MS-GF+ bottom-up search engine.[56]

Estimated Yield

All search configurations, original search outputs, and filtered ProForma outputs can be downloaded as Supporting Information File S1.

In each of the four test sets shown in Figure 3, either MSPathFinderT or pTop yields the highest number of confident PrSMs. In PXD020342-DANRE, the difference between the highest yield and lowest yield is considerably smaller. The first three tests all employed a sequence database spanning the complete reference proteomes in FASTA format for both ECOLI and SULIS, but only the SULIS tests incorporated methylation of Lys as a PTM. TopPIC and MSPathFinderT both automatically processed the ECOLI EThcD data as if they were produced via ETD rather than EThcD (considering only c−z fragments rather than a mix of c−z and b−y fragments). In pTop2, both ETD and EThcD modes were tested with the ECOLI set, and ETD mode produced the larger yield. Only ProSight PD appears to have benefited from its EThcD identification mode.

The SULIS set is also valuable for evaluating the relative performance of identification algorithms in HCD versus ETD. While all four algorithms identified more PrSMs in HCD than in ETD experiments, the Orbitrap Fusion required more time to acquire each ETD MS/MS (summing two microscans) than it did to acquire an HCD MS/MS (one microscan), generating 12,664 ETD MS/MS and 24,015 HCD MS/MS in the same chromatographic time. The four algorithms identified a range from 13 to 32% of all HCD spectra and a range from 18 to 36% of all ETD spectra. The SULIS data tentatively argue that top-down algorithms perform better on ETD spectra than on HCD spectra, but the ETD spectra were generally from higher signal-to-noise precursor ions than the HCD spectra because the instrument was fragmenting fewer overall precursor ions. Only pTop identified roughly the same fraction of MS/MS in both ETD and HCD experiments (32%). The other three algorithms identified higher fractions of MS/MS in the ETD set.

FDR Estimation

For *S. islandicus* and *E. coli* experiments, we used a single FASTA database that contained all 2591 proteins from the SULIS reference proteome (UP000013006) and all 4450 proteins from the ECOLI reference proteome (UP000000625). Any match to an *E. coli* protein in *S. islandicus* experiments would be

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

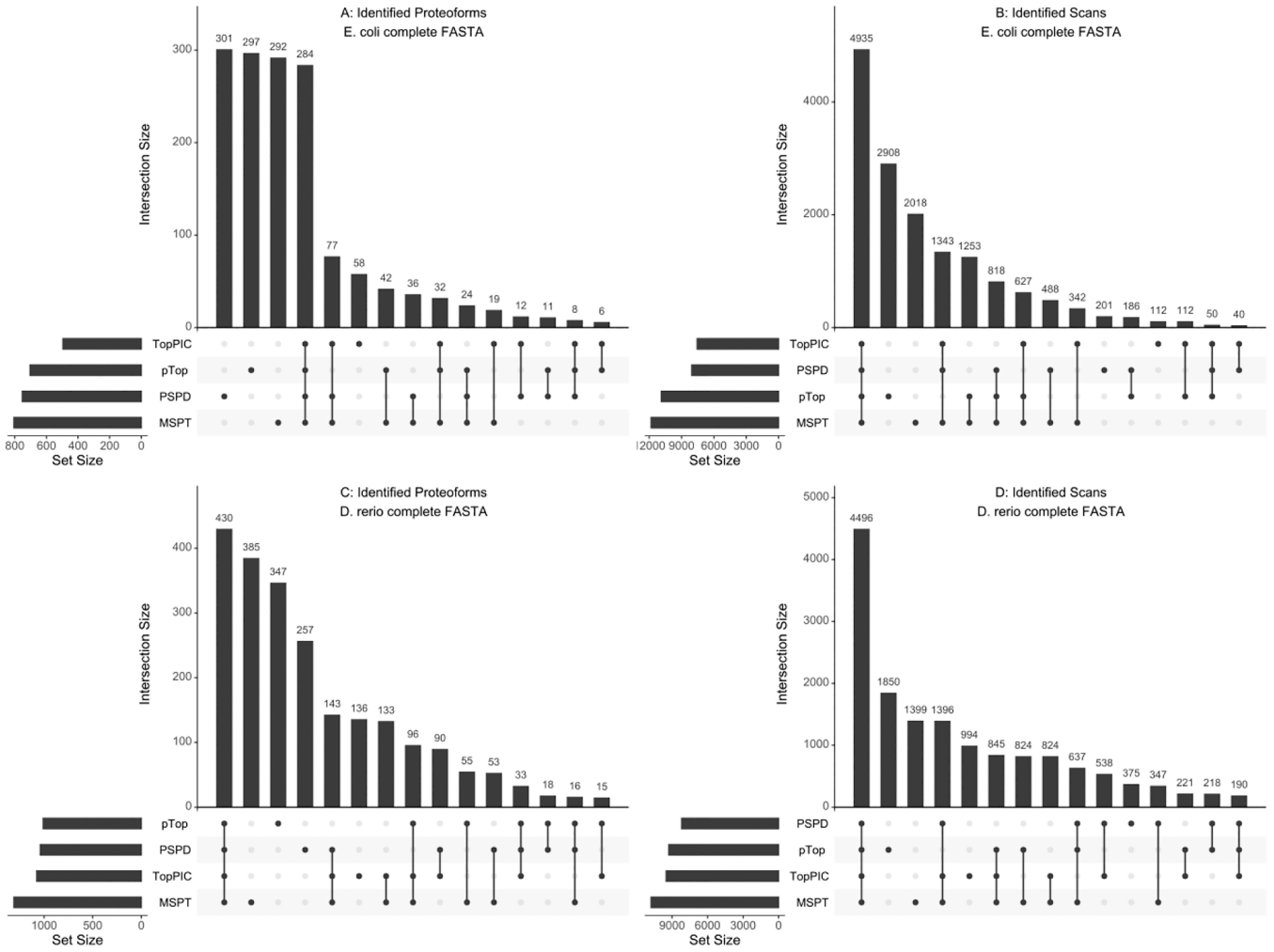


Figure 4. UpSet diagrams show the intersection of proteoforms (panels A and C) and identified spectra (panels B and D) in the *E. coli* and *D. rerio* experiments. A vertical bar may be thought of as a region of a Venn Diagram, with the areas accounting for the most items sorted to the left. The dots on the vertical lines show which search engines detected a given proteoform or identified a particular spectrum. Similar plots for *S. islandicus* appear in Supporting Figure S3.

considered a known false PrSM, and vice versa. These known false matches could be used to estimate the number of false matches hidden among the plausible accessions based on the ratio of sequence counts for the two species. (We tested these two reference proteomes for shared sequences by seeking the ortholog pair in ProteinOrtho 6 that produced the highest BLASTP bitscore;[57] the longest run of identical amino acids for M9UBS4_SULIS and NARG_ECOLI was 14 amino acids in length.) Our intent with this SULIS-ECOLI database was to limit PrSM-level FDR. LeDuc et al. have noted that this type of decoy strategy is much less effective at controlling isoform- or protein-level false discoveries.[58]

At first, ProSight PD was tested with its consensus step targeting a 1% FDR via target-decoy analysis. In practice, however, the SULIS and ECOLI data sets showed an effective FDR near 0% for this filtering, compromising its sensitivity. As a result, ProSight PD PrSMs were filtered for "high confidence" instead. See the "ProForma Exporters" section above to see other filtering details.

Different PrSM filtering routes for each of the four workflows led to very comparable PrSM FDRs, as re-estimated using the false hits from either ECOLI or SULIS. In several cases for SULIS, the estimated FDR rose above 1%; MSPathFinderT yielded a set of PrSMs at a 1.7% FDR for the HCD experiments, ProSight PD PrSMs gave an empirical FDR of 2.6% for ETD and 1.6% for HCD, and pTop produced an estimated FDR of 1.2% for the HCD experiments. A similar pattern appeared in the *E. coli* experiments, where MSPathFinderT delivered an effective FDR of 1.10% and pTop yielded 1.05%, but ProSight PD gave FDRs of 2.3 and 2.8% for XML and FASTA sequence inputs, respectively. If the ProSight PD truncation search is limited to 10 ppm precursor tolerance rather than 1.1 Da precursor tolerance, its FDR falls below 1%. Estimates of FDR using target-decoy search may be compromised if the number of database entries is low or the number of identified tandem mass spectra is low.[59] The ECOLI and SULIS experiments employed thousands of sequences in the FASTA and yielded thousands of confidently identified PrSMs, reducing potential error in FDR estimates.

A discussion of why a search may report multiple PrSMs from an individual MS/MS is in Supporting Information Text S3.

Although MSPathFinderT and pTop workflows yielded impressive numbers of PrSMs, these search engines are

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

accompanied by some challenges. The Informed Proteomics suite that includes MSPathFinderT is not currently under active development, and its deconvolution and search efficiency lags behind the other tools. A user may wait more than twice as long for the results from ProMex/MSPathFinderT as from other workflows (see Supporting Information Text S4 for time trials of all four workflows). Our testing of pTop 2 software represented a "mix-and-match" pairing the pParseTD deconvolution engine from version 1.2 with an early build of the pTop search engine from version 2.0. The December 2022 release version of pTop2 with a fixed pParseTD and modified pTop arrived after the finalization of the search results for this study.

Because the *E. coli* UniProtKB database annotates many PTM-modified residues, we detected differences between the ProSight PD searches when using FASTA and UniProtKB XML. The use of UniProtKB XML instead of FASTA produced 3.1% more PrSMs at the "high" filtering criterion. 725 proteoforms were found in common between the searches, 31 were found only with the UniProtKB XML database, and 28 were found only with the FASTA database. The UniProtKB XML search loses some proteoforms that appear in the FASTA search because its search space was larger by including all UniProtKB-annotated PTMs and coding SNPs, reducing PrSM yield. The proteoform that represented the most PrSMs gained in the UniProtKB XML search was P0A7N9 (50S ribosomal protein L33), with the N-terminal Met clipped and the Ala at position two methylated (rather than acetylated). By itself, this proteoform accounted for 106 PrSMs, making it the eighth most frequent proteoform in the ProSight PD XML search. L-Beta-methylthioaspartic acid (+46 Da on Asp) was also a frequently observed modification due to its occurrence in 30S ribosomal protein S12.

The use of "subset" databases from bottom-up proteomics can reduce the time required to search proteoforms from large eukaryotic proteomes and greatly accelerate PTM search (as employed in the PTM section below). Note, however, that this approach precludes the identification of very small proteins, or proteins bearing a limited number of Lys/Arg residues that could be missed in bottom-up analyses because of their biochemical properties. *D. rerio* data analysis faced challenges of both a large sequence database and a need for added variable PTMs. First, its proteome of 47,204 sequences could be reduced to 9432 sequences through the identification of bottom-up spectra accompanying the top-down files. Second, iodoacetamide induced many proteoforms where Cys side-chains were incompletely labeled, slowing identification considerably (see PTM section). The fivefold reduction in number of sequences through subsetting certainly accelerated identification (see Supporting Information Text S5 for time trials of database subsetting), but it did come at a cost. In all four search engines, the number of PrSMs was slightly lower in the subset database searches than it was in the complete database (the biggest difference was −646 PrSMs in pTop, while the smallest was −58 PrSMs in TopPIC). When compared with the overall numbers of PrSMs identified by these engines, these losses are probably manageable (9281 in the complete FASTA search by pTop and 9504 by TopPIC). These losses may have been mitigated if we had employed non-parsimonious protein inference in producing our subset database. From a biochemical perspective, it appears that at least some proteoforms do not easily yield to proteolysis, necessitating top-down proteomics for detection.

Identification Overlap

It should be noted that a sensitive proteoform identifier does not necessarily identify all the scans identified by a less-sensitive algorithm plus some additional set. Instead, these four search engines produced a very complex set of overlaps at both proteoform and spectrum level in the *E. coli* and *D. rerio* experiments (see Figure 4). Of 15,154 distinct scans identified by at least one search engine in *D. rerio*, only 4496 (30%) were identified by all four algorithms. Of 15,433 distinct scans identified in *E. coli*, 4935 (32%) were identified by all algorithms. MSPathFinderT was the most sensitive engine for both *D. rerio* and *E. coli*, identifying 72 or 77% of all these identifiable spectra, respectively. This implies that even if a researcher could predict which algorithm would be most sensitive on a data set, plenty of identifiable spectra would remain unidentified if no other algorithms were employed. By contrast, if the two least sensitive algorithms are used to identify the *D. rerio* set, 13,033 PrSMs (84%) would be identified; *the "worst" pair of algorithms delivers better PrSM sensitivity than the "best" choice for a single algorithm.* We caution that accepting the superset of PrSMs for two different algorithms is quite likely to produce elevated FDR relative to either of the searches that contribute to it. Nonetheless, top-down proteomics identification could benefit from strategies, such as post-search re-scoring[60,61] to combine the information from multiple scoring strategies in recognizing reliable PrSMs. "Voting" models and other data integration approaches from bottom-up proteome informatics[62,63] would make it possible to boost the proteoforms that are consistent among algorithms and downplay proteoforms that are found by only one algorithm.

Since each proteoform may be observed at different precursor charges or be duplicated in multiple MS/MS, top-down experiments offer some degree of redundancy. In *D. rerio*, the 15,154 PrSMs match 2207 distinct proteoforms, giving an average of 6.9 spectra per proteoform; identifying any one of them will add the proteoform to the list of identifications. In *E. coli*, this redundancy is even higher because the six experiments are essentially technical triplicates collected on two dates; on average, 10.2 spectra per proteoform were identified by at least one of the four algorithms. Of all 2207 proteoforms for *D. rerio*, 19.5% were identified by all four engines. In *E. coli*, 20.1% of the 1499 distinct proteoforms were identified unanimously. The advantages seen for PrSMs in using two search engines apply for proteoforms, as well. Combining results for the two least sensitive search engines yields a set of 1553 distinct proteoforms in *D. rerio*, while the best individual performer yielded 1311.

Two search engines may agree that a spectrum has been identified and yet disagree in its interpretation. We considered three different levels of PrSM agreement: Did the algorithms agree on (a) the precursor charge, (b) the accession from which the sequence is drawn, (c) the proteoform sequence and amount of mass contributed by modifications? These types of agreement can be considered in the context of the five-level proteoform classification scheme proposed by Smith et al.[64] In all cases, we considered only the summed nominal mass of post-translational modifications, not their localization, so level 1 proteoform level agreement was not attempted in this study. While we always associate a possibly truncated sequence with its protein database accession, we ignored gene associations. Two of these search engines may agree entirely on truncated sequence and mass of added PTMs for a given spectrum and yet associate those sequences with different accessions, particularly if multiple isoforms for that protein-coding gene exist and contain the same

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

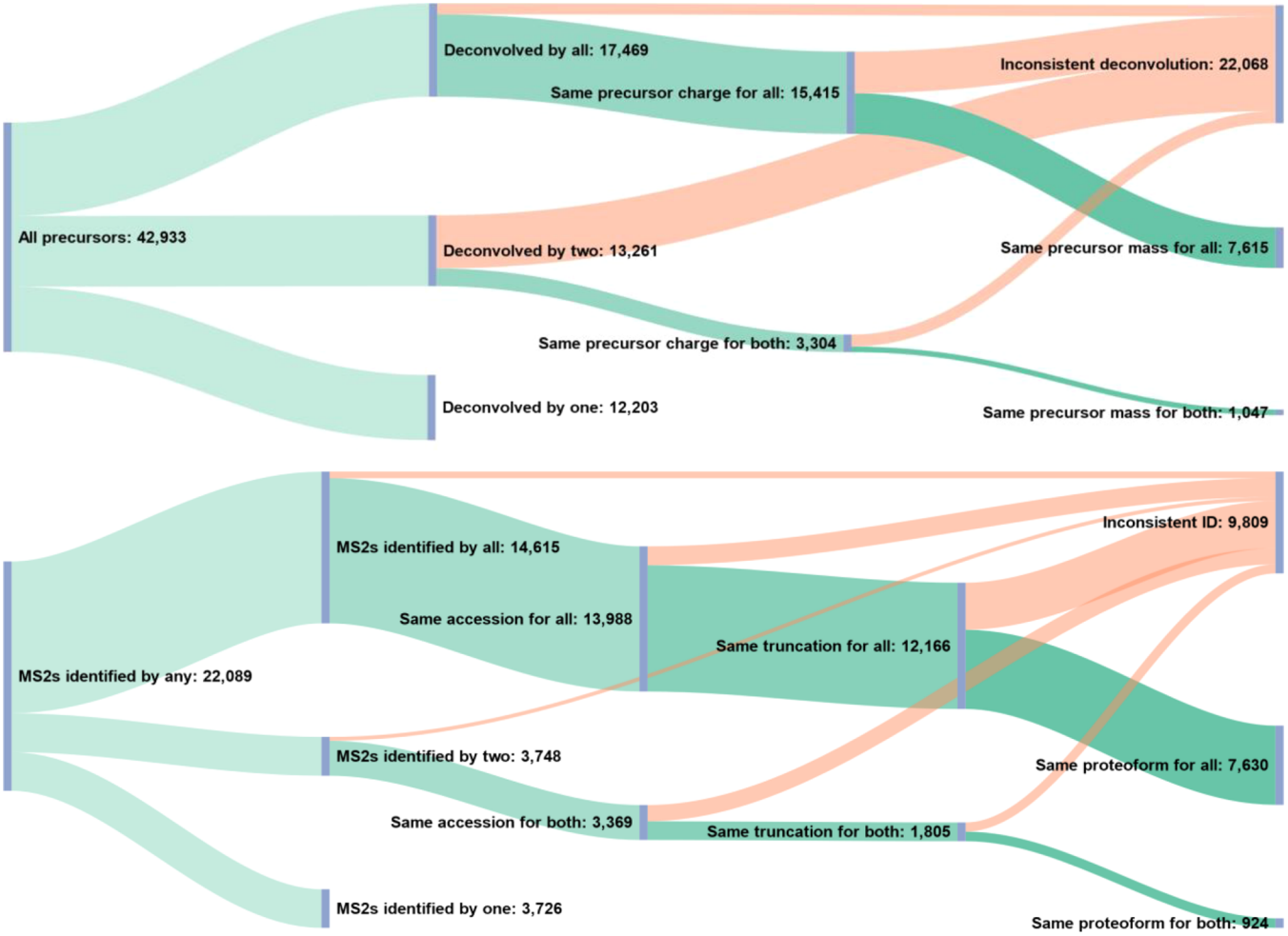


Figure 5. These Sankey diagrams visualize the flow of the *E. coli* MS/MS scans through the process of deconvolution and of identification. Deconvolution methods may disagree on whether a precursor can be deconvolved at all, which charge the precursor represents, and the monoisotopic mass attributed to the precursor. Identification methods may disagree on whether an MS/MS has been identified successfully, which database accession produced the sequence, and the truncation and PTM decoration borne by the proteoform.

truncated sequence. Of all four pipelines, only TopPIC reports all protein accessions that contain a given truncated proteoform sequence, and that feature was not yet present in version 1.4.13, used here. Since we do not take PTM localization and gene-of-origin into account when comparing proteoforms, the best-case agreement our strategy can achieve is level 3 on the Smith proteoform level classification system.

For each pairing of search engines in *S. islandicus*, *E. coli*, and *D. rerio*, we created an "inner join" of raw file names and scan numbers to construct temporary tables of MS/MS scans that both search engines identified. Supporting Information Table S4 reports the percentage of these spectra-in-common for which the precursor charge was identical, the accession was identical, and the combination of truncated sequence and rounded modification mass added was identical. The *E. coli* experiments give a near-ideal result. The precursor charge matched for 93 to 99% of the jointly identified spectra, accessions matched for 97 to almost 100% of the spectra, and proteoforms matched for 86 to 97% of the spectra. The other sets held some anomalies, though. The *S. islandicus* experiments (both HCD and ETD) revealed that precursor charge state inference by Xtract in the ProSight PD workflow frequently diverged from the other algorithms, agreeing in precursor charge only 71−80% of the time for these mutually identified spectra. In *D. rerio*, accession matching was markedly lower than for the other species, ranging from 83 to 91%, probably reflecting that isoforms in the sequence database increase the opportunity for disagreements in identification (and algorithms may differ in which accession is reported when multiple accessions contain the PrSM sequence). The proteoform agreement tables showed diminished concordance for pTop versus other algorithms in *S. islandicus* (43−69%) and *D. rerio* (77−89%) compared to what it produced in *E. coli* (86−94%). Even when search engines agree that a particular spectrum has been identified, they may disagree considerably in detailing the proteoform a tandem mass spectrum represents.

Deconvolution Contributes to PrSM Inconsistency

The term "deconvolution" is frequently used to describe many roles for top-down analysis at both MS and MS/MS levels, including the following:[6]

- feature detection: recognizing persistent precursor chromatograms in MS signals,
- charge state inference: recognizing isotope spacing or ions at multiple charges,
- charge reduction: representing all fragments at neutrality or unit charge, and
- deisotoping: combining intensity across an isotopomer envelope.

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

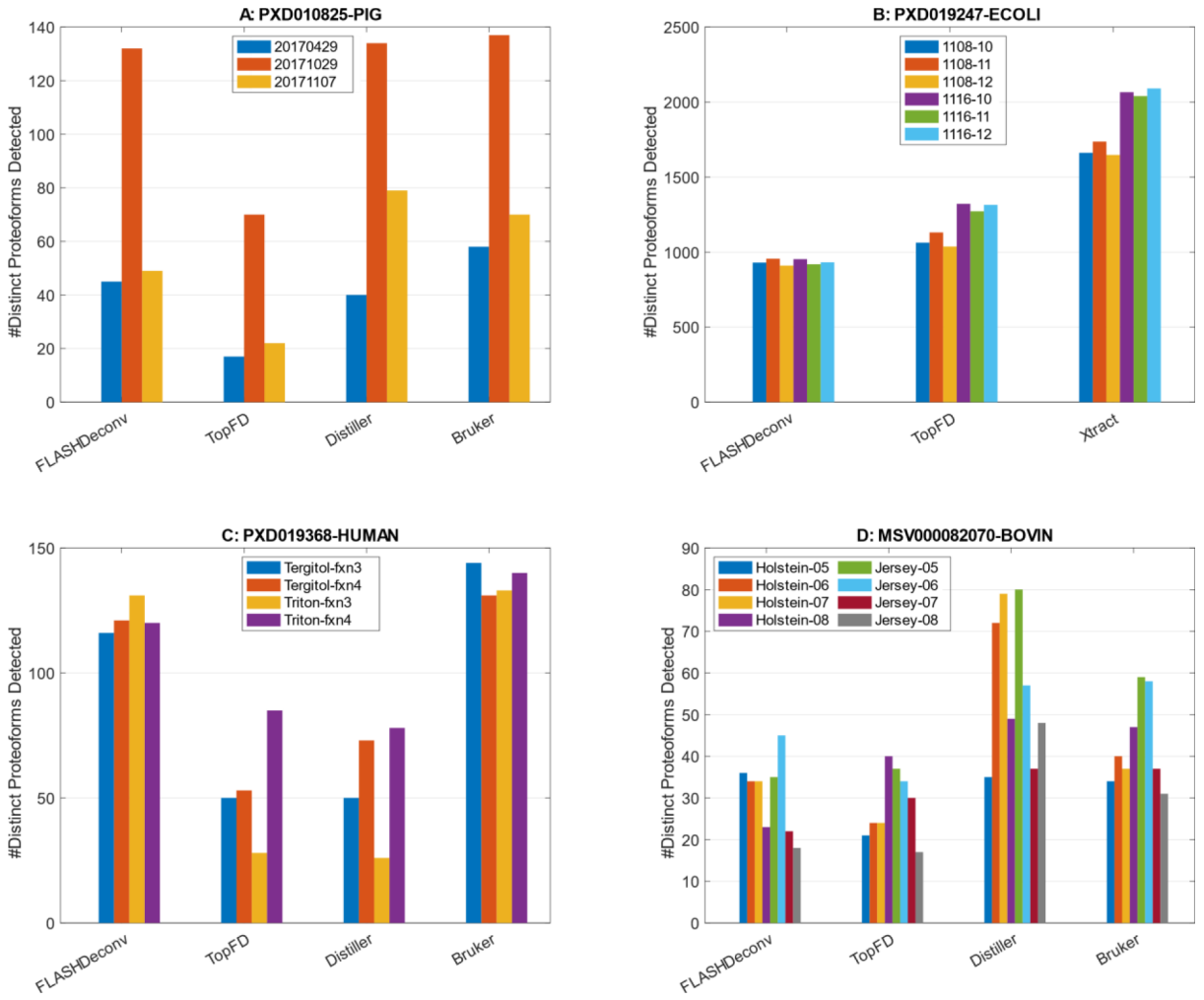


Figure 6. Influence of the deconvolution algorithm on the number of distinct proteoforms detected in our four datasets. Each color in a panel represents a particular LC-MS/MS experiment. Panels (A, C, and D) represent Bruker maXis I or II data, while Panel (B) represents Thermo Fusion Lumos experiments. Mascot Distiller was excluded from Panel (B) because it did not yield comparable identification performance.

Because each of the four search engines is packaged with a different deconvolution engine, both deconvolution and identification may be opening the door to variability. We held the search engine constant (using TopPIC) while varying the deconvolution engine to determine the variability contributed by deconvolution alone. We interrogated three Bruker maXis I and II data sets and examined the ThermoFisher *E. coli* set described above, using these deconvolution systems: Xtract as employed in ProSight PD (for *E. coli* only), AutoMSn from Bruker DataAnalysis (for all but *E. coli*), TopFD from TopPIC Suite, Matrix Science Mascot Distiller, and FLASHDeconv.

An important factor appeared early in the analysis of the Q-TOF data. Bruker's AutoMSn approach detects that many MS/MS scans have been produced from the same "component," and it merges these signals together prior to deconvolution. As a result, the number of peak lists it exports is far smaller than what TopFD exports, since TopFD aspires to report each MS/MS individually to its output, whether or not it can detect their precursors in the MS. FLASHDeconv integrated functions from OpenMS SpectralMerger to emulate the Bruker software's behavior. Meanwhile, Mascot Distiller frequently found multiple precursor masses in the MS1, so it reported some MS/MS scans multiple times. In processing the ThermoFisher data, Xtract employed a Sliding Window Algorithm[65] to average multiple MS scans preceding and succeeding a particular MS/MS to increase signal-to-noise ratio for the precursor ion. However, combining replicated MS/MS scans was quite rare among the Xtract peak lists. Supporting Information Figure S4 illustrates the disparity observed in these MS/MS scan counts.

Inferring precursor charge for thousands of precursors at a wide variety of signal-to-noise ratios and in the context of thousands of other proteoforms, which can possibly overlap, is highly error-prone. When deconvolution algorithms are unable to infer charge successfully, some will write a 0 (neutral) or + 1 charge as an error code while others will omit the corresponding MS/MS from output. Plotting the charge state distribution of these LC-MS/MS experiments revealed considerable shifts from engine to engine (see Supporting Information Figure S5). The number of fragment masses appearing in deconvolved MS/MS scans also showed considerable variability (see Supporting Information Figure S6).

To compare the outputs of deconvolution engines, we looked for three types of agreement: (A) Multiple deconvolution engines included this MS/MS scan in the output. (B) Multiple deconvolution engines agreed on the inferred precursor charge for this MS/MS scan. (C) Multiple deconvolution engines agreed on the precursor mass for this MS/MS scan within 15 ppm. These deconvolution outputs were then identified in the

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

TopPIC search engine. We sought four different types of agreement for the resulting identifications: (D) The search results from different deconvolution engines indicated that this MS/MS scan was successfully identified. (E) The search results from different deconvolution engines matched this MS/MS scan to the same protein accession. (F) The search results from different deconvolution engines matched this MS/MS scan to the same truncation of the protein sequence. (G) The search results from different deconvolution engines matched this MS/MS scan to the same summed nominal mass of PTMs.

Figure 5 illustrates the progress of MS/MS scans for six *E. coli* experiments through these stages. The *E. coli* experiments were chosen for visualization because MS/MS averaging was not employed by any of the deconvolution methods (making the fate of individual scans easier to track). Mascot Distiller was omitted from this visualization because its outputs did not lead to comparable levels of identification for these experiments. The remaining candidates included TopFD, FLASHDeconv, and ThermoFisher Xtract (as implemented in ProSight PD workflows).

The upper part of Figure 5 shows that 28% of all the *E. coli* MS/MS scans were deconvolved by only one deconvolution engine; for these spectra we cannot evaluate reproducibility further. At the other extreme, 41% of all scans were deconvolved by all three engines; these are likely to be the MS/MS scans that offer the highest signal-to-noise precursor ions. Of the spectra reported by all three engines, 88% were assigned the same precursor charge across all three engines. Assigning a precursor mass (using the same criterion as TopPIC: 15 ppm), however, sees far lower agreement, with only 49% of the precursors being given the same mass despite having unanimous agreement on precursor charge (with many likely representing off-by-one monoisotope errors). Please note that a spectrum with an incorrectly determined precursor charge or inaccurate mass may still be identified to a partially correct proteoform, particularly when modifications of any mass are permitted.

Deconvolution discrepancies can have a substantial but sometimes subtle effect on identification, and the lower part of Figure 5 seeks to characterize these effects. Again, it is possible that the TopPIC search engine makes no report about a particular scan number because the PrSM does not pass the expectation value filters. In the *E. coli* data, 66% of the MS/MS scans identified by TopPIC after any type of deconvolution were identified after all three deconvolution engines. Of the spectra identified after all three deconvolution routes, 96% were matched to the same accession in the database (this would likely have been lower if we evaluated the data from *D. rerio* due to the presence of paralogs and isoform variants). When all three search engines identified an MS/MS to the same accession, however, the TopPIC-reported sequence truncation agreed across the three deconvolution routes only 87% of the time. The total mass of PTMs for these proteoforms was also inconsistent. The sequence truncation and total PTM mass disparities are likely a result of deconvolution routes supplying different intact masses or charges for MS/MS precursors. This type of deconvolution variation causes underestimation of FDR since the PrSMs are matched to partially correct sequences.

Figure 6 illustrates the effect of choosing a different deconvolution algorithm on the yield of distinct proteoforms. For Bruker maXis II experiments (Panels A and C), simply using TopFD on mzMLs produced from the raw data significantly impeded identification because each input MS/MS was reported separately. Approaches that can sum replicated MS/MS scans for deconvolution will lead to real gains in signal-to-noise for fragments. The result from the Fusion Lumos *E. coli* experiment (Panel B) was surprising because we expected that TopFD mass lists would be an ideal match with the TopPIC search engine; they were developed by the same team and are distributed together. Instead, we observed that supplying the peak lists from Xtract used by ProSight PD identified substantially more spectra. Though it struggled in the Thermo *E. coli* experiment, Mascot Distiller demonstrated its value for recovering identifications in the Bruker maXis II porcine heart (Panel A) and Bruker maXis I bovine milk (Panel D) experiments.

PTM Identification

Incorporating post-translational modifications in top-down identification can seem inconsistent among search engines. For example, is the acetylation of proteoform N-termini configured alongside other PTMs (pTop, MSPathFinderT), or is this modification handled as a special case (ProSight PD, TopPIC)? Because the cost of variable modifications (such as Met oxidation) can be substantial in top-down search, the way they are handled by the different software tools is important.

The example of Cys carbamidomethylation (+57 Da) in the *D. rerio* experiment gives an interesting effect. In bottom-up experiments, database searches are often configured to assume all Cys are shifted to 160 rather than 103 Da, implying that the alkylation reaction has run to completion. The TopPIC identifications from the original publication of PXD020342 employed a "fixed" shift of +57 Da for all Cys; correspondingly, when iodoacetamide failed to react with a Cys side chain, TopPIC often used its unanticipated modification feature to denote a shift of −57 Da (back to normal side chain mass) at these sites. Our searches incorporated +57 as a dynamic or variable modification, typically allowing up to four PTMs per PrSM. Of all Cys residues identified by the four algorithms in the complete database, an unexpectedly small fraction was found to bear the expected carbamidomethylation: MSPathFinderT = 30.8%, ProSight PD = 30.9%, pTop = 28.6%, and TopPIC = 20.9%. This low apparent reaction stoichiometry may be peculiar to this experiment, but it reinforces that one should not assume these mass shift-inducing reactions will mark all potential sites. Note that any incomplete reaction in sample handling would lead to an artificial increase in the number of proteoforms, entirely through artifacts of chemical processing, while also complicating identification.

Proteoform Phospho-Search in Bovine Milk

We created in this study a new milk data set (PXD031744-BOVIN), analyzed on three different instruments, that should be useful for improving the identification of phospho-proteoforms. To evaluate phosphorylation detection against a more challenging background of acetylated or unmodified proteins, we also evaluated ovarian cancer samples (PXD005420-HUMAN). We used two different strategies for reducing the database to a manageable size for phosphorylation searching. In the milk data, we used the TMT "bottom-up" proteome to reduce the complete bovine FASTA (37,883 sequences) to those matching peptides after parsimony (5388 sequences) (see Figure 7). We then performed an identification search in all four top-down engines to reduce to a 378 accession FASTA that we employed for PTM search. Our UniProt XML Subsetter script produced both FASTA and UniProtKB XML subsets for our list of accessions. In the human ovarian experiments, however, we did not have bottom-up data available. Instead, we used the two faster search engines (ProSight PD and pTop) to search the

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

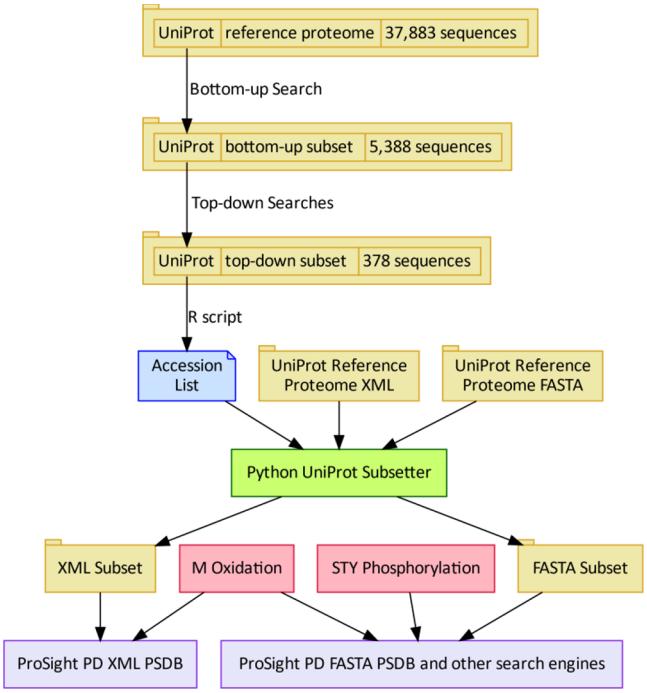


Figure 7. A complex process prepared a sequence database for bovine milk phosphorylation searches. First we created a subset database based on bottom-up experiments, and second we filtered that subset database to include only the proteins detected in top-down searches that incorporated minimal numbers of PTMs per PrSM. The very compact 378 protein databases could then be drawn from FASTA or XML-formatted UniProtKB databases, using the list of accessions as input. Met oxidation was always added as a variable modification for searches, but Ser, Thr, and Tyr phosphorylations were only added as a variable modification when the sequences were provided in FASTA format.

complete human proteome FASTA (101,014 sequences) to generate a 558 protein subset. The PTM search for the ovarian data employed this compact collection of sequences.

We hypothesized the PTM search might show a greater disparity between FASTA and UniProtKB XML databases in ProSight PD, so both searches are included in our comparison. The TopMG tool in TopPIC Suite is intended to detect multiply-modified proteoforms, and so it was included alongside TopPIC (as before, any PrSMs containing unknown PTMs from TopPIC were excluded from consideration). Taken as a whole, the sensitivity of identification for all proteoforms from the milk experiment seemed very consistent among the six searches, from a low of 21,765 PrSMs for TopPIC to a high of 25,922 PrSMs in ProSight PD with XML input. (Note that this search is one where the ProSight PD benefits greatly from the 1.1 Da precursor mass tolerance.) If the PrSMs are separated by the three instruments or the four different samples from which they derived, however (see Supporting Information Figure S7), a surprising difference between pTop and the other search algorithms emerged: the number of PrSMs identified from the Eclipse instrument data via pTop exceeded by 49% the PrSMs produced by ProSight PD (XML input), the second most sensitive algorithm for this instrument. From an experimental point of view, there was no reason to expect that the Eclipse data would be particularly fertile ground for pTop. The other five algorithms showed heightened performance on the "C13" sample (containing more cellular proteins than the other milks), with 41−51% of their identifications coming from the three experiments representing that sample, but pTop struggled with those data files, yielding only 23% of its PrSMs from "C13" experiments. Algorithm-specific performance variation may be accentuated in PTM searches.

Variation is particularly evident in the PTM content of these PrSMs. The phosphorylation-rich caseins should make the identification of singly-, doubly-, triply-, and even quadruply-phosphorylated proteoforms feasible in these experiments. Because the scripts for processing the PrSMs table recorded a nominal mass of PTMs added to each PrSM, simply counting the number of times "80" appears in this column expresses the number of proteoforms that carry exactly one phosphorylation and no other PTMs. Singly phosphorylated PrSMs comprised the following percentages of PrSMs by search engine: 17.9% (TopMG), 17.8% (pTop), 13.0% (MSPathFinderT), 12.9% (TopPIC), 1.4% (ProSight PD with XML database), and 1.2% (ProSight PD with FASTA database). ProSight PD and TopPIC also identified very small numbers of doubly-phosphorylated PrSMs but no PrSMs with more than two phosphorylations. Of all the search engines, only MSPathFinderT was able to identify many PrSMs for doubly-(3.8%), triply-(5.0%), and quadruply-phosphorylated (3.4%) proteoforms with no other modifications. The milk proteome highlights MSPathFinderT, pTop, and TopMG for phosphorylation searching.

Supporting Information Text S6 evaluates the ProSight PD searches from FASTA and from XML for detecting phosphorylations. Anecdotally, when ProSight PD was configured for annotated proteoform search rather than the truncation search, the phosphorylated proteoforms were a larger proportion of a smaller number of PrSMs.

A part of the pTop divergence for this set is explained by plotting the number of PrSMs for a given PTM mass sum (see Figure 8). pTop identifications included substantial numbers of singly- and multiply-oxidized PrSMs in the Orbitrap Eclipse experiments. It is unclear why the other search algorithms did not observe this pattern, given that they were also configured to allow for multiple oxidation of PrSMs. The distribution of PrSMs by PTM mass for ProSight PD using UniProtKB XML input reflects the great diversity of PTM sums that come into play when UniProtKB PTM annotation is available to modify proteoform mass ladders, including mass losses as low as −18 (no other search engine was configured for PTMs with negative masses). TopMG was the only tool allowed up to five PTMs per PrSM (its default setting), allowing it to reach a maximum mass gain of 400 Da (five phosphorylations).

PTMs in Human Ovarian and *S. islandicus* Proteomes

The ovarian cancer samples (PXD005420-HUMAN) lead to very different results. Singly phosphorylated proteoforms comprise a smaller fraction of all spectra in the set; instead, N-terminal acetylation (+42) is the most dominant PTM. For this set, pTop held the advantage in the number of PrSMs identified, apparently through improved sensitivity from the "PB" and "PT" triplicates (see Supporting Information Figure S9). Five of the six searches concurred that between 36.9 and 44.2% of identified PrSMs represented acetylated proteoforms, but for TopMG only 17.2% of PrSMs were acetylated (the PTM distributions for all six identification pathways in the ovarian cancer samples are shown in Supporting Information Figure S10). Recovery of +80 PrSMs was generally consistent among search engines: TopMG (8.7%), ProSight PD (5.4 or 5.3%, depending on database format), MSPathFinderT (3.2%), and pTop (3.1%). For TopPIC, however, only 0.4% of PrSMs were singly phosphory-

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

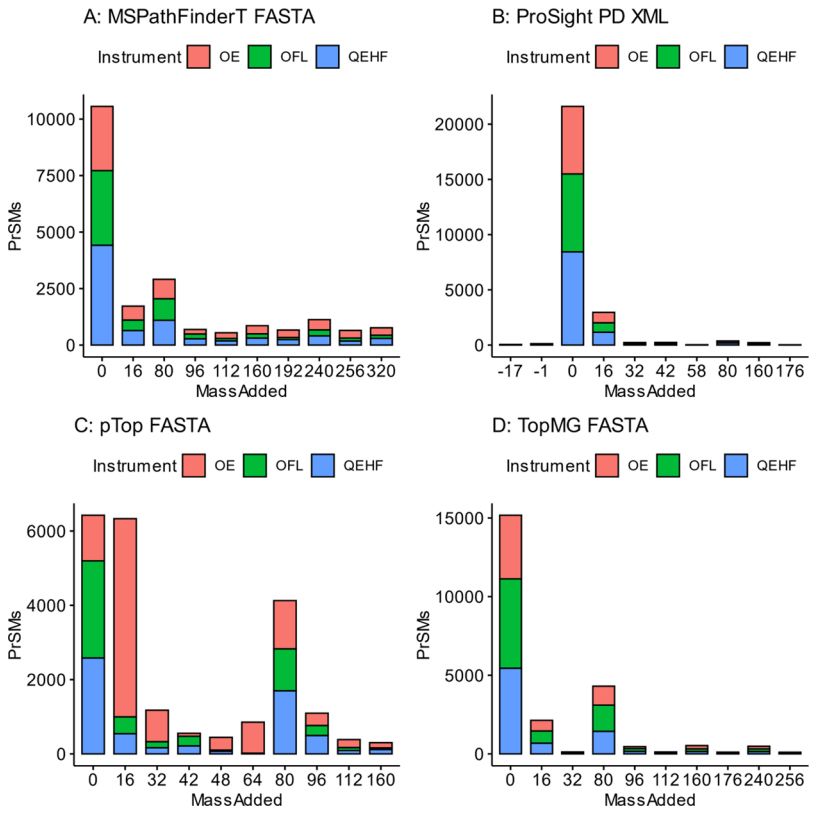


Figure 8. Number of PrSMs identified by four search algorithms, split by the amount of mass contributed by PTMs, for the bovine milk data set in three ThermoFisher instruments. Only the top ten most frequent PTM masses for each algorithm were visualized. "0" implies a proteoform identified without any PTMs. See Supporting Information Figure S8 for complete set of plots for all PTM masses.

lated. While the milk data suggested that ProSight PD identified phosphorylated PrSMs with poor sensitivity, the ovarian cancer set ranks it as the second-best at detecting this class of PTMs. Both sets demonstrate the relative strength of TopMG over TopPIC for recognizing multiple-PTM PrSMs. While the ovarian cancer data might have been expected to reveal glycosylated proteoforms, as well, TopPIC analysis of unanticipated mass shifts revealed calcium and sodium adducts were far more prominent in the set (see Supporting Information Text S7).

The SULIS data set examined in the identification yield section above also contains interesting challenges for PTM hunting. The *S. islandicus* proteome is extensively Lys-methylated, and this variable PTM was included in all searches of that set. The 7029 sequences of its combined ECOLI/SULIS database considerably outnumbered the hundreds used for the final PTM searches in bovine milk and human ovarian sets. For each of the four algorithms used to identify PrSMs in this data set, we sought to determine the percentage that represented singly methylated (+14 Da), doubly methylated (+28 Da), and triply methylated or acetylated (+42 Da) proteoforms. All four algorithms detected many proteoforms that bore a single methylation (ranging from 10.5 to 16.4% in HCD experiments and from 10.8 to 19.3% in ETD experiments). The detection of acetylation or three methylations was also substantial (from 7.2 to 18.7% in HCD and from 6.8 to 14.4% in ETD). ProSight PD produced an outlier for the case of dimethylated proteoforms. It identified zero PrSMs with two added methyl groups in HCD and in ETD sets, while the other algorithms identified a minimum of 2.6% and a maximum of 7.3%. In combination with the bovine milk result, the absence of dimethylated PrSMs for SULIS suggests that the scoring or filtering for ProSight PD favors unmodified or singly-modified PrSMs rather than combining multiple PTMs in a single proteoform. By contrast, pTop2 produced a surprising number of PrSMs containing a total of 70 or 86 Da in PTMs (28% of all PrSMs in HCD and 15.1% in ETD); these mass shift values could be interpreted as dimethylation plus acetylation (70 Da) or dimethylation, acetylation, and oxidation (86 Da). If top-down search algorithms vary in their ability to identify multiple PTMs, interpreting the results of PTM searches requires critical consideration.

CONCLUSIONS

Having analyzed several different datasets, this study may offer some insight on the question "which instruments can be used for top-down proteomics?" The results above indicate that hundreds of proteoforms can be identified even with a ThermoFisher Q-Exactive mass spectrometer (released in 2011). Features that boost mass range, acquire spectra from precursor ions representing different neutral masses, and accelerate the overall rate of MS/MS acquisition can all play a role in improving the quality of a data set representing a top-down proteome.

Many top-down proteomics practitioners need answers to the question "what algorithm should I use to identify proteoforms from my data?" A case can be made for the use of any of these four identification pipelines, depending upon the priorities of the researcher. *TopPIC* frequently lagged behind pTop and MSPathFinderT in PrSM counts, largely because our ProForma reformatter suppressed all PrSMs that contain unanticipated modifications. The software comes with substantial benefits,

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

though, such as PTM localization ambiguity reporting, the ability of all its component tools to run in the Linux operating system, and the greater explanatory power afforded by the unanticipated modifications.[66] Although this study routinely disabled the HTML reporting of TopPIC, users get considerable value from its voluminous output, visualizing PrSMs and proteoforms in a portable report that can be browsed without specialized software. *ProSight PD* will continue to be a popular option for users familiar with the Proteome Discoverer framework, and its conservative estimation of FDR[58] protects against falsely identified proteoforms. The search speed of ProSight PD is also impressive, with deconvolution often requiring more time than matching sequences to fragment masses. The cost for a license to this software provides technical support and continuous improvement in the software suite. *MSPathFinderT* delivered very impressive sensitivity throughout and because it exports identifications in the mzIdentML format, it is the easiest search engine for producing a "complete" entry in ProteomeXchange. On the other hand, because MSPathFinderT is only being maintained rather than actively developed, it will likely remain the slowest of the four search engines. The *pTop* algorithm is in a transitional state, with pTop 1 completed and well-characterized while pTop 2 receives final touches. The hybrid of pParseTD from version 1.2 and pTop from version 2.0 delivered first-rate sensitivity on par with that of MSPathFinderT at a speed comparable to that of ProSight PD, and pTop is more likely to incorporate multiple PTMs in its matches than other algorithms. All these pipelines have features that favor them in different biological research scenarios. We recommend the use of ProSight PD or pTop for rapid screening of experiments, with more thorough assessment through MSPathFinderT or TopPIC. As the UpSet diagrams of Figure 4 reveal, however, users must be aware that the algorithm they use will strongly impact the information they produce from top-down experiments.

Making it easy to integrate different top-down identification workflows will require a wide variety of efforts. A likely route would include these elements:

- At present, the file formats used to store deconvolved mass spectra and tandem mass spectra differ by pipeline (PF and MS for pTop, PBF for Informed Proteomics, msAlign for TopPIC Suite, and MSF for ProSight PD). Mass spectrometry bioinformatics needs an open, rigorously described, standardized format for storing deconvolved data sets. The widely used HUPO-PSI mzML format (or one of its more compact variants) is an obvious candidate for storing deconvolutions if the top-down field can agree to common conventions for storing neutral or singly charged mass lists. When deconvolutions are recorded in a standard way, two key benefits will result. Any compliant deconvolution engine can be paired with any compliant search engine (thus interoperability), and it will become far easier to determine whether deconvolutions agree for a given MS/MS and its precursor ion information.
- Recording PrSMs in a standard format to aid ProteomeXchange import would greatly aid the reproducibility of proteoform identification. The Informed Proteomics software suite, featuring MSPathFinderT, stands out for recording its identifications in HUPO-PSI mzIdentML format. HUPO-PSI mzTab format[67] is a lighter-weight alternative that may represent an easier option for software developers.
- At present, search engines all employ different formats for reporting truncated sequence and PTM information for each PrSM. The HUPO-PSI ProForma version 2 format is intended to resolve this problem, but at the time of this writing, none of the search engines format their proteoform descriptions by this standard. In particular, the communication of PTM localization ambiguity needs attention; only TopPIC, of all four pipelines analyzed here, records to its text reports when a PTM location is unambiguous.
- As Cesnik et al. observe,[68] relating proteoforms to protein accessions is a first step, but associating proteoforms to genes will be a necessary step for proteogenomics. As a first step, all top-down search engines should enumerate all FASTA accessions that match a particular proteoform, not just the first in the database.
- As noted above, simply "taking the union" of all PrSMs from multiple search engines will escalate FDR in the accepted identifications. Similarly, accepting only the PrSMs on which multiple search algorithms agree ("the intersection") is far too conservative. Models are needed to compare multiple deconvolution assessments on the same data and integrate multiple attempts at identification in a way that emphasizes agreements and flags disagreements.

In short, the top-down identification field contains excellent opportunities for bioinformatics advances. The state-of-the art for proteoform identification already produces solid results, and yet more will become possible with new advances in the field.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00673.

Text S1 (PDF): Detailed sample handling details for PXD031744 bovine milk samples. Text S2 (PDF): Impact of "suppressing" PrSMs for unanticipated PTMs in TopPIC search. Text S3 (PDF): Why would a search identify multiple PrSMs for an individual MS/MS? Text S4 (PDF): How do identification pathways compare in duration. Text S5 (PDF): How much time does using a subset sequence database save. Text S6 (PDF): Testing ProSight PD for detecting phosphorylation in PXD031744 bovine milk. Text S7 (PDF): Are glycosylated proteoforms prominent in PXD005420 ovarian tumor data. Figure S1 (PDF): Configuration in Mascot Distiller 2.8. Figure S2 (PDF): Configuration in Bruker DataAnalysis 5.3. Figure S3 (PDF): Overlap of identified proteoforms and scan numbers in PXD003074 SULIS. Figure S4 (PDF): Deconvolution may produce very different numbers of MS/MS mass lists. Figure S5 (PDF): Deconvolution impacts precursor charge distribution. Figure S6 (PDF): Deconvolution engines export MS/MS mass lists of different length. Figure S7 (PDF): PXD031744 bovine milk PrSMs by search engine. Figure S8 (PDF): PTM distributions for the PXD031744 bovine milk identifications. Figure S9 (PDF): PXD005420 ovarian tumor PrSMs by search engine. Figure S10 (PDF): PTM distributions for the PXD005420 ovarian tumor sample identifications (PDF)

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

Table S1 (XLSX): QC, ID, and Deconvolution Metrics for Identification Section. Table S2 (XLSX): QC, ID, and Deconvolution Metrics for Deconvolution Section. Table S3 (XLSX): QC, ID, and Deconvolution Metrics for PTM Section. Table S4 (XLSX): Agreement statistics for spectra identified by multiple search engines (XLSX)

File S1 (ZIP): Search configurations, output files, and ProForma exports for each search (ZIP)

AUTHOR INFORMATION

Corresponding Author

Julia Chamot-Rooke − *Université Paris Cité, Institut Pasteur, CNRS UAR 2024, Mass Spectrometry for Biology Unit, Paris 75015, France;* orcid.org/0000-0002-9427-543X; Email: julia.chamot-rooke@pasteur.fr

Authors

David L. Tabb − *Université Paris Cité, Institut Pasteur, CNRS UAR 2024, Mass Spectrometry for Biology Unit, Paris 75015, France;* orcid.org/0000-0001-7223-578X

Kyowon Jeong − *Applied Bioinformatics, Computer Science Department, University of Tübingen, Tübingen 72076, Germany;* orcid.org/0000-0003-3776-3098

Karen Druart − *Université Paris Cité, Institut Pasteur, CNRS UAR 2024, Mass Spectrometry for Biology Unit, Paris 75015, France;* orcid.org/0000-0002-9572-7741

Megan S. Gant − *Université Paris Cité, Institut Pasteur, CNRS UAR 2024, Mass Spectrometry for Biology Unit, Paris 75015, France*

Kyle A. Brown − *School of Medicine and Public Health, University of Wisconsin, Madison, Wisconsin 53705, United States;* orcid.org/0000-0003-1255-9146

Carrie Nicora − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Mowei Zhou − *Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0003-3575-3224

Sneha Couvillion − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Ernesto Nakayasu − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Janet E. Williams − *Department of Animal, Veterinary, and Food Sciences, University of Idaho, Moscow, Idaho 83844, United States*

Haley K. Peterson − *Department of Animal, Veterinary, and Food Sciences, University of Idaho, Moscow, Idaho 83844, United States*

Michelle K. McGuire − *Margaret Ritchie School of Family and Consumer Sciences, University of Idaho, Moscow, Idaho 83844, United States*

Mark A. McGuire − *Department of Animal, Veterinary, and Food Sciences, University of Idaho, Moscow, Idaho 83844, United States*

Thomas O. Metz − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00673

Author Contributions

○These authors constitute the Milk Microbiome and Metabolome Team.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study has been supported by EPIC-XS, project number 823839, funded by the Horizon 2020 programme of the European Union. This project has received funding from the European Horizon 2020 research and innovation programme under grant agreement number 829157. DLT would like to thank Mick Greer, Xiaowen Liu, In Kwon Choi, Matt Monroe, Chi Hao, Michael Marty, and Rachel Miller for their expert assistance in configuring and interpreting these bioinformatic workflows. The peer-reviewers of this study contributed many useful insights, and we acknowledge the effort each invested. Bovine milk experiments were performed in the Environmental Molecular Sciences Laboratory (under project 10.46936/reso.proj.2020.51433/60000202), a national scientific user facility sponsored by the U.S. OBER and located at PNNL in Richland, Washington. PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830. Funding was provided by the National Institutes of Health grant number 1R01HD092297-01A1 to Mark A. McGuire and Michelle K. McGuire.

ABBREVIATIONS

FDR, false discovery rate; MS/MS, tandem mass spectra; PrSM, proteoform-spectrum match; PTM, post-translational modification

REFERENCES

(1) Smith, L. M.; Kelleher, N. L. Consortium for Top Down Proteomics. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10*, 186−187.

(2) Harshman, S. W.; Young, N. L.; Parthun, M. R.; Freitas, M. A. H1 Histones: Current Perspectives and Challenges. *Nucleic Acids Res.* **2013**, *41*, 9593−9609.

(3) Fischer, F.; Wolters, D.; Rögner, M.; Poetsch, A. Toward the Complete Membrane Proteome: High Coverage of Integral Membrane Proteins through Transmembrane Peptide Detection. *Mol. Cell. Proteomics* **2006**, *5*, 444−453.

(4) Peng, Y.; Chen, X.; Zhang, H.; Xu, Q.; Hacker, T. A.; Ge, Y. Top-down Targeted Proteomics for Deep Sequencing of Tropomyosin Isoforms. *J. Proteome Res.* **2013**, *12*, 187−198.

(5) Zhou, M.; Malhan, N.; Ahkami, A. H.; Engbrecht, K.; Myers, G.; Dahlberg, J.; Hollingsworth, J.; Sievert, J. A.; Hutmacher, R.; Madera, M.; Lemaux, P. G.; Hixson, K. K.; Jansson, C.; Paša-Tolić, L. Top-down Mass Spectrometry of Histone Modifications in Sorghum Reveals Potential Epigenetic Markers for Drought Acclimation. *Methods* **2020**, *184*, 29−39.

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

(7) Frottin, F.; Martinez, A.; Peynot, P.; Mitra, S.; Holz, R. C.; Giglione, C.; Meinnel, T. The Proteomics of N-Terminal Methionine Cleavage. *Mol. Cell. Proteomics* **2006**, *5*, 2336−2349.

(8) Antelmann, H.; Tjalsma, H.; Voigt, B.; Ohlmeier, S.; Bron, S.; van Dijl, J. M.; Hecker, M. A Proteomic View on Genome-Based Signal Peptide Predictions. *Genome Res.* **2001**, *11*, 1484−1502.

(9) Doucet, A.; Overall, C. M. Protease Proteomics: Revealing Protease in Vivo Functions Using Systems Biology Approaches. *Mol. Aspects Med.* **2008**, *29*, 339−358.
(10) Kim, J.-S.; Monroe, M. E.; Camp, D. G.; Smith, R. D.; Qian, W.-J. In-Source Fragmentation and the Sources of Partially Tryptic Peptides in Shotgun Proteomics. *J. Proteome Res.* **2013**, *12*, 910−916.
(11) Janssen, E. M.; Dy, S. M.; Meara, A. S.; Kneuertz, P. J.; Presley, C. J.; Bridges, J. F. P. Analysis of Patient Preferences in Lung Cancer - Estimating Acceptable Tradeoffs Between Treatment Benefit and Side Effects. *Patient Prefer Adherence* **2020**, *14*, 927−937.
(12) Taylor, G. K.; Kim, Y.-B.; Forbes, A. J.; Meng, F.; McCarthy, R.; Kelleher, N. L. Web and Database Software for Identification of Intact Proteins Using "Top down" Mass Spectrometry. *Anal. Chem.* **2003**, *75*, 4081−4086.
(13) Meng, F.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L. Informatics and Multiplexing of Intact Protein Identification in Bacteria and the Archaea. *Nat. Biotechnol.* **2001**, *19*, 952−957.
(14) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced Global Post-Translational Modification Discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17*, 1844−1851.
(15) Frank, A. M.; Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L.; Pevzner, P. A. Interpreting Top-down Mass Spectra Using Spectral Alignment. *Anal. Chem.* **2008**, *80*, 2499−2505.
(16) Liu, X.; Sirotkin, Y.; Shen, Y.; Anderson, G.; Tsai, Y. S.; Ting, Y. S.; Goodlett, D. R.; Smith, R. D.; Bafna, V.; Pevzner, P. A. Protein Identification Using Top-Down. *Mol. Cell. Proteomics* **2012**, *11*, No. M111.008524.
(17) Kou, Q.; Xun, L.; Liu, X. TopPIC: A Software Tool for Top-down Mass Spectrometry-Based Proteoform Identification and Characterization. *Bioinformatics* **2016**, *32*, 3495−3497.
(18) Sun, R.-X.; Luo, L.; Wu, L.; Wang, R.-M.; Zeng, W.-F.; Chi, H.; Liu, C.; He, S.-M. PTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal. Chem.* **2016**, *88*, 3082−3090.
(19) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolić, N.; Paša-Tolić, L.; Smith, R. D.; Payne, S. H.; Kim, S. Informed-Proteomics: Open-Source Software Package for Top-down Proteomics. *Nat. Methods* **2017**, *14*, 909−914.
(20) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. MzML–a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics* **2011**, *10*, No. R110.000133.
(21) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; Julian, R.; Binz, P.-A.; Deutsch, E. W.; Hermjakob, H.; Reisinger, F.; Griss, J.; Vizcaíno, J. A.; Chambers, M.; Pizarro, A.; Creasy, D. The MzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Mol. Cell. Proteomics* **2012**, *11*, No. M111.014381.
(22) Kou, Q.; Wu, S.; Tolic, N.; Paša-Tolic, L.; Liu, Y.; Liu, X. A Mass Graph-Based Approach for the Identification of Modified Proteoforms Using Top-down Tandem Mass Spectra. *Bioinformatics* **2017**, *33*, 1309−1316.
(23) Cai, W.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X.; Ge, Y. MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. *Mol. Cell. Proteomics* **2016**, *15*, 703−714.
(24) Vorontsov, E. A.; Rensen, E.; Prangishvili, D.; Krupovic, M.; Chamot-Rooke, J. Abundant Lysine Methylation and N-Terminal Acetylation in Sulfolobus Islandicus Revealed by Bottom-Up and Top-Down Proteomics. *Mol. Cell. Proteomics* **2016**, *15*, 3388−3404.
(25) Dupré, M.; Duchateau, M.; Malosse, C.; Borges-Lima, D.; Calvaresi, V.; Podglajen, I.; Clermont, D.; Rey, M.; Chamot-Rooke, J. Optimization of a Top-Down Proteomics Platform for Closely Related Pathogenic Bacterial Discrimination. *J. Proteome Res.* **2021**, *20*, 202−211.
(26) Xu, T.; Shen, X.; Yang, Z.; Chen, D.; Lubeckyj, R. A.; McCool, E. N.; Sun, L. Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative and Quantitative Top-Down Proteomics. *Anal. Chem.* **2020**, *92*, 15890−15898.
(27) Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. Capillary Zone Electrophoresis-Mass Spectrometry for Top-down Proteomics. *TrAC, Trends Anal. Chem.* **2019**, *120*, No. 115644.
(28) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods* **2017**, *14*, 513−520.
(29) Holman, J. D.; Ma, Z.-Q.; Tabb, D. L. Identifying Proteomic LC-MS/MS Data Sets with Bumbershoot and IDPicker. *Curr. Protoc. Bioinf.* **2012**, *13*, Unit13.17.
(30) Delcourt, V.; Franck, J.; Leblanc, E.; Narducci, F.; Robin, Y.-M.; Gimeno, J.-P.; Quanico, J.; Wisztorski, M.; Kobeissy, F.; Jacques, J.-F.; Roucou, X.; Salzet, M.; Fournier, I. Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine* **2017**, *21*, 55−64.
(31) Brown, K. A.; Chen, B.; Guardado-Alvarez, T. M.; Lin, Z.; Hwang, L.; Ayaz-Guner, S.; Jin, S.; Ge, Y. A Photocleavable Surfactant for Top-down Proteomics. *Nat. Methods* **2019**, *16*, 417−420.
(32) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, *24*, 2534−2536.
(33) Yang, L.; Gregorich, Z. R.; Cai, W.; Zhang, P.; Young, B.; Gu, Y.; Zhang, J.; Ge, Y. Quantitative Proteomics and Immunohistochemistry Reveal Insights into Cellular and Molecular Processes in the Infarct Border Zone One Month after Myocardial Infarction. *J. Proteome Res.* **2017**, *16*, 2101−2112.
(34) Brown, K. A.; Tucholski, T.; Alpert, A. J.; Eken, C.; Wesemann, L.; Kyrvasilis, A.; Jin, S.; Ge, Y. Top-Down Proteomics of Endogenous Membrane Proteins Enabled by Cloud Point Enrichment and Multidimensional Liquid Chromatography-Mass Spectrometry. *Anal. Chem.* **2020**, *92*, 15726−15735.
(35) An, H.; Ordureau, A.; Körner, M.; Paulo, J. A.; Harper, J. W. Systematic Quantitative Analysis of Ribosome Inventory during Nutrient Stress. *Nature* **2020**, *583*, 303−309.
(36) Vincent, D.; Mertens, D.; Rochfort, S. Optimisation of Milk Protein Top-Down Sequencing Using In-Source Collision-Induced Dissociation in the Maxis Quadrupole Time-of-Flight Mass Spectrometer. *Molecules* **2018**, *23*, E2777.
(37) Vincent, D.; Ezernieks, V.; Elkins, A.; Nguyen, N.; Moate, P. J.; Cocks, B. G.; Rochfort, S. Milk Bottom-Up Proteomics: Method Optimization. *Front. Genet.* **2016**, *6*, 360.
(38) Khalid, M. F.; Iman, K.; Ghafoor, A.; Saboor, M.; Ali, A.; Muaz, U.; Basharat, A. R.; Tahir, T.; Abubakar, M.; Akhter, M. A.; Nabi, W.; Vanderbauwhede, W.; Ahmad, F.; Wajid, B.; Chaudhary, S. U. PERCEPTRON: An Open-Source GPU-Accelerated Proteoform Identification Pipeline for Top-down Proteomics. *Nucleic Acids Res.* **2021**, *49*, W510−W515.
(39) Karabacak, N. M.; Li, L.; Tiwari, A.; Hayward, L. J.; Hong, P.; Easterling, M. L.; Agar, J. N. Sensitive and Specific Identification of Wild Type and Variant Proteins from 8 to 669 KDa Using Top-down Mass Spectrometry. *Mol. Cell. Proteomics* **2009**, *8*, 846−856.
(40) Tsai, Y. S.; Scherl, A.; Shaw, J. L.; MacKay, C. L.; Shaffer, S. A.; Langridge-Smith, P. R. R.; Goodlett, D. R. Precursor Ion Independent Algorithm for Top-down Shotgun Proteomics. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2154−2166.
(41) LeDuc, R. D.; Schwämmle, V.; Shortreed, M. R.; Cesnik, A. J.; Solntsev, S. K.; Shaw, J. B.; Martin, M. J.; Vizcaino, J. A.; Alpi, E.; Danis, P.; Kelleher, N. L.; Smith, L. M.; Ge, Y.; Agar, J. N.; Chamot-Rooke, J.; Loo, J. A.; Pasa-Tolic, L.; Tsybin, Y. O. ProForma: A Standard Proteoform Notation. *J. Proteome Res.* **2018**, *17*, 1321−1325.
(42) LeDuc, R. D.; Deutsch, E. W.; Binz, P.-A.; Fellers, R. T.; Cesnik, A. J.; Klein, J. A.; Van Den Bossche, T.; Gabriels, R.; Yalavarthi, A.; Perez-Riverol, Y.; Carver, J.; Bittremieux, W.; Kawano, S.; Pullman, B.; Bandeira, N.; Kelleher, N. L.; Thomas, P. M.; Vizcaíno, J. A. Proteomics

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX

Standards Initiative's ProForma 2.0: Unifying the Encoding of Proteoforms and Peptidoforms. *J. Proteome Res.* **2022**, *21*, 1189−1195.
(43) Sun, R.-X.; Wang, R.-M.; Luo, L.; Liu, C.; Chi, H.; Zeng, W.-F.; He, S.-M. Accurate Proteoform Identification and Quantitation Using PTop 2.0. In *Proteoform Identification*; Sun, L.; Liu, X., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2022; Vol. *2500*, pp 105−129.
(44) Kirchner, M.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. MGFp: An Open Mascot Generic Format Parser Library Implementation. *J. Proteome Res.* **2010**, *9*, 2762−2763.
(45) Jeong, K.; Kim, J.; Gaikwad, M.; Hidayah, S. N.; Heikaus, L.; Schlüter, H.; Kohlbacher, O. FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. *Cell Syst.* **2020**, *10*, 213−218.e6.
(46) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nat. Methods* **2016**, *13*, 741−748.
(47) Vyatkina, K.; Dekker, L. J. M.; Wu, S.; VanDuijn, M. M.; Liu, X.; Tolić, N.; Luider, T. M.; Paša-Tolić, L. De Novo Sequencing of Peptides from High-Resolution Bottom-Up Tandem Mass Spectra Using Top-Down Intended Methods. *Proteomics* **2017**, *17*, 23−24.
(48) Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K. A.; Benesch, J. L. P.; Robinson, C. V. Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **2015**, *87*, 4370−4376.
(49) Thermo Scientific. XCALI-98282: FreeStyle User Guide: Software Version 1.8, 2021. https://assets.thermofisher.com/TFS-Assets/CMD/manuals/xcali-98282-free-style-user-guide-xcali98282-en.pdf.
(50) Klimek, J.; Eddes, J. S.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P. R.; Katz, J. E.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J. K.; Aebersold, R.; Martin, D. B. The Standard Protein Mix Database: A Diverse Data Set to Assist in the Production of Improved Peptide and Protein Identification Software Tools. *J. Proteome Res.* **2008**, *7*, 96−103.
(51) Falkner, J. A.; Kachman, M.; Veine, D. M.; Walker, A.; Strahler, J. R.; Andrews, P. C. Validated MALDI-TOF/TOF Mass Spectra for Protein Standards. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 850−855.
(52) Dang, X.; Scotcher, J.; Wu, S.; Chu, R. K.; Tolić, N.; Ntai, I.; Thomas, P. M.; Fellers, R. T.; Early, B. P.; Zheng, Y.; Durbin, K. R.; Leduc, R. D.; Wolff, J. J.; Thompson, C. J.; Pan, J.; Han, J.; Shaw, J. B.; Salisbury, J. P.; Easterling, M.; Borchers, C. H.; Brodbelt, J. S.; Agar, J. N.; Paša-Tolić, L.; Kelleher, N. L.; Young, N. L. The First Pilot Project of the Consortium for Top-down Proteomics: A Status Report. *Proteomics* **2014**, *14*, 1130−1140.
(53) Shen, X.; Xu, T.; Hakkila, B.; Hare, M.; Wang, Q.; Wang, Q.; Beckman, J. S.; Sun, L. Capillary Zone Electrophoresis-Electron-Capture Collision-Induced Dissociation on a Quadrupole Time-of-Flight Mass Spectrometer for Top-Down Characterization of Intact Proteins. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 1361−1369.
(54) Bennett, K. L.; Wang, X.; Bystrom, C. E.; Chambers, M. C.; Andacht, T. M.; Dangott, L. J.; Elortza, F.; Leszyk, J.; Molina, H.; Moritz, R. L.; Phinney, B. S.; Thompson, J. W.; Bunger, M. K.; Tabb, D. L. The 2012/2013 ABRF Proteomic Research Group Study: Assessing Longitudinal Intralaboratory Variability in Routine Peptide Liquid Chromatography Tandem Mass Spectrometry Analyses. *Mol. Cell. Proteomics* **2015**, *14*, 3299−3309.
(55) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C. Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography-Tandem Mass Spectrometry. *J. Proteome Res.* **2010**, *9*, 761−776.
(56) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(57) Lechner, M.; Findeiss, S.; Steiner, L.; Marz, M.; Stadler, P. F.; Prohaska, S. J. Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis. *BMC Bioinf.* **2011**, *12*, 124.
(58) LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Shams, D. P.; Thomas, P. M.; Kelleher, N. L. Accurate Estimation of Context-Dependent False Discovery Rates in Top-Down Proteomics. *Mol. Cell. Proteomics* **2019**, *18*, 796−805.
(59) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111−1120.
(60) Frank, A. M. A Ranking-Based Scoring Function for Peptide-Spectrum Matches. *J. Proteome Res.* **2009**, *8*, 2241−2252.
(61) Zolg, D. P.; Gessulat, S.; Paschke, C.; Graber, M.; Rathke-Kuhnert, M.; Seefried, F.; Fitzemeier, K.; Berg, F.; Lopez-Ferrer, D.; Horn, D.; Henrich, C.; Huhmer, A.; Delanghe, B.; Frejno, M. INFERYS Rescoring: Boosting Peptide Identifications and Scoring Confidence of Database Search Results. *Rapid Commun. Mass Spectrom.* **2021**, No. e9128.
(62) Jones, A. R.; Siepen, J. A.; Hubbard, S. J.; Paton, N. W. Improving Sensitivity in Proteome Studies by Analysis of False Discovery Rates for Multiple Search Engines. *Proteomics* **2009**, *9*, 1220−1229.
(63) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining Results of Multiple Search Engines in Proteomics. *Mol. Cell. Proteomics* **2013**, *12*, 2383−2393.
(64) Smith, L. M.; Thomas, P. M.; Shortreed, M. R.; Schaffer, L. V.; Fellers, R. T.; LeDuc, R. D.; Tucholski, T.; Ge, Y.; Agar, J. N.; Anderson, L. C.; Chamot-Rooke, J.; Gault, J.; Loo, J. A.; Paša-Tolić, L.; Robinson, C. V.; Schlüter, H.; Tsybin, Y. O.; Vilaseca, M.; Vizcaíno, J. A.; Danis, P. O.; Kelleher, N. L. A Five-Level Classification System for Proteoform Identifications. *Nat. Methods* **2019**, *16*, 939−940.
(65) Gazis, P. R.; Horn, D. M. *Poster Note 64404: The Sliding Window Algorithm for the Analysis of LC/MS Intact Protein Data*; 2015. https://assets.thermofisher.com/TFS-Assets/CMD/posters/PN-64404-LC-MS-Sliding-Window-Intact-Protein-Data-ASMS2015-PN64404-EN.pdf.
(66) Wilson, J. W.; Zhou, M. Discovery of Unknown Posttranslational Modifications by Top-Down Mass Spectrometry. *Methods Mol. Biol.* **2022**, *2500*, 181−199.
(67) Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; Cox, J.; Neumann, S.; Fan, J.; Reisinger, F.; Xu, Q.-W.; Del Toro, N.; Pérez-Riverol, Y.; Ghali, F.; Bandeira, N.; Xenarios, I.; Kohlbacher, O.; Vizcaíno, J. A.; Hermjakob, H. The MzTab Data Exchange Format: Communicating Mass-Spectrometry-Based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Mol. Cell. Proteomics* **2014**, *13*, 2765−2775.
(68) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. *J. Proteome Res.* **2021**, *20*, 1826−1834.

https://doi.org/10.1021/acs.jproteome.2c00673
J. Proteome Res. XXXX, XXX, XXX−XXX