Journal of proteome research

pubs.acs.org/jpr

Technical Note

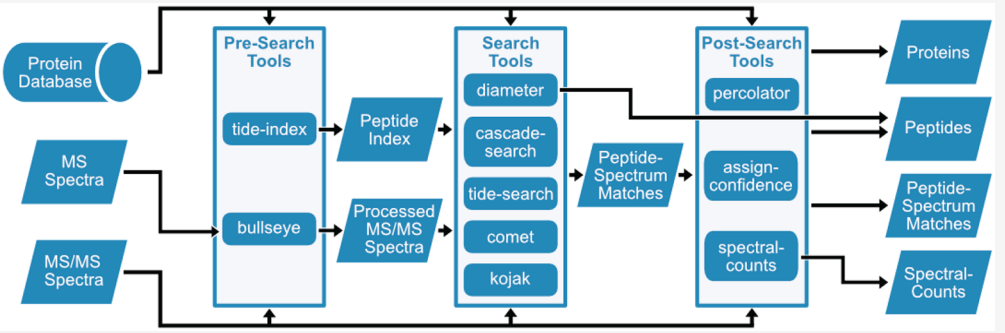
The Crux Toolkit for Analysis of Bottom-Up Tandem Mass Spectrometry Proteomics Data

Attila Kertesz-Farkas, Frank Lawrence Nii Adoquaye Acquaye, Kishankumar Bhimani, Jimmy K. Eng, William E. Fondrie, Charles Grant, Michael R. Hoopmann, Andy Lin, Yang Y. Lu, Robert L. Moritz, Michael J. MacCoss, and William Stafford Noble*

Cite This: *J. Proteome Res.* 2023, 22, 561−569

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The Crux tandem mass spectrometry data analysis toolkit provides a collection of algorithms for analyzing bottom-up proteomics tandem mass spectrometry data. Many publications have described various individual components of Crux, but a comprehensive summary has not been published since 2014. The goal of this work is to summarize the functionality of Crux, focusing on developments since 2014. We begin with empirical results demonstrating our recently implemented speedups to the Tide search engine. Other new features include a new score function in Tide, two new confidence estimation procedures, as well as three new tools: Param-medic for estimating search parameters directly from mass spectrometry data, Kojak for searching cross-linked mass spectra, and DIAmeter for searching data independent acquisition data against a sequence database.

KEYWORDS: *mass spectrometry, database search, false discovery rate control, open source software*

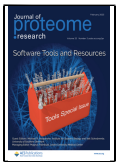
1. INTRODUCTION

Continual technological advances in mass spectrometry instrumentation, which yield higher throughput; increased data depth, accuracy, and precision; and innovative orthogonal modes of ion measurement require concomitant advances in analytical methods. Crux is an open source software project that implements a variety of state-of-the-art algorithms for interpreting bottom-up tandem mass spectrometry proteomics data. The algorithms implemented in Crux are described in 40 scientific papers, cited a total of 6413 times and with an H-index of 25 (https://scholar.google.com/citations?hl=en&user=Rw9S1HIAAAAJ, Sep 26, 2022). A typical Crux user is unlikely to read this large corpus of papers; hence, the goal of this paper is to provide an overview of Crux, with a focus on developments that have been introduced since our last overview paper in 2014.[1]

The field of computational mass spectrometry is broad, and Crux necessarily occupies a particular niche within that field. In particular, Crux focuses primarily on the initial stages of tandem mass spectrometry analysis: the assignment of peptides to spectra, with associated measures of statistical confidence at the level of spectra, peptides, and proteins. Crux includes four database search tools, two for standard search (Tide and Comet), one for searching against a database of cross-linked peptides (Kojak), and one for searching data-independent acquisition (DIA) data (DIAmeter) (Figure 1). Also included is the Bullseye tool for assigning high-resolution precursor masses to MS2 spectra, a machine learning postprocessor (Percolator), a separate tool for assigning confidence estimates to various types of discoveries (assign-confidence), and a spectral counting tool (spectral-counts). Practically speaking, Crux is a command line tool, written in C++. Source code is available, and we also provide precompiled binaries for use on Microsoft Windows, MacOS and Linux operating systems from http://crux.ms.

Special Issue: Software Tools and Resources 2023

Received: September 30, 2022
Published: January 4, 2023



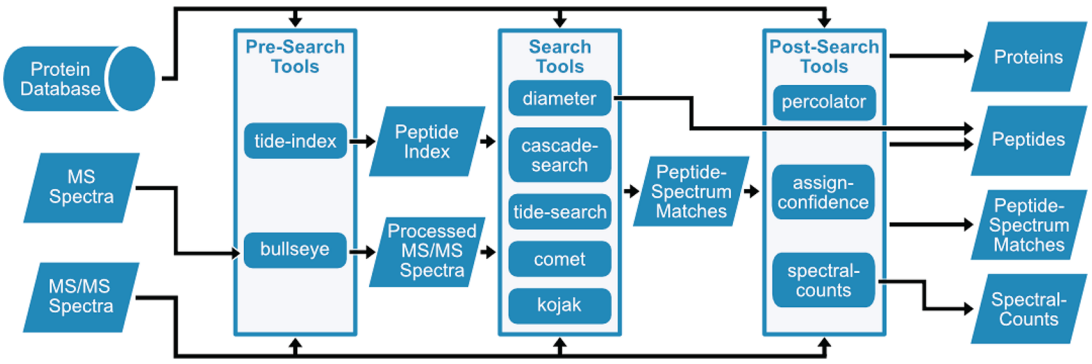


Figure 1. Overview of tools in Crux. Bullseye assigns high resolution precursor m/z values to tandem mass spectra. Crux includes two DDA search tools, Tide and Comet, plus a variant of Tide called cascade-search, described in Section 3.2. DIAmeter searches data-independent acquisition data, and Kojak searches cross-linked mass spectra. Percolator is a machine learning postprocessor, assign-confidence estimates statistical confidence estimates directly from search results, and spectral-counts computes several types of protein abundance measures using spectral counting.

Table 1. Parameter Settings for Comet and Tide

Tide		Comet	
parameter	value	parameter	value
enzyme	trypsin	search_enzyme_number	1
digestion	full-digest	num_enzyme_termini	2
missed-cleavages	2	allowed_missed_cleavage	2
min-peaks	10	minimum_peaks	10
precursor-window	10	peptide_mass_tolerance	10
precursor-window-type	ppm	peptide_mass_units	2
fragment-mass	mono	mass_type_fragment	1
decoy-format	peptide-reverse	N/A	
keep-terminal-aminos	C	N/A	
concat	T	decoy_search	1
top-match	1	num_results, num_output_lines	2, 1
remove-precursor-peak	T	remove_precursor_peak	1
remove-precursor-tolerance	15	remove_precursor_tolerance	15
use-flanking-peaks	F	theoretical_fragment_ions	1
use-neutral-loss-peaks	F	use_NL_ions	0
mz-bin-width	0.02	fragment_bin_tol	0.02
mz-bin-offset	0.4	fragment_bin_offset	0.4
min-mass, max-mass	200, 7200	digest_mass_range	200, 7200
N/A		max_fragment_charge	2
min-length max-length	6 40	peptide_length_range	6 40
mods-spec	2M+15.99,2STY+79.96	variable_mod01	15.99 M 0 2−1 0 0 0.0
N/A		variable_mod02	79.96 STY 0 2−1 0 0 0.0
nterm-protein-mods-spec	1K+42.01	variable_mod03	42.01 n 0 1 0 0 0 0.0
max-mods	2	max_variable_mods_in_peptide	2

In this paper, we provide an overview of new features in Crux (summarized in Supporting Information S1), beginning with empirical results demonstrating our recently implemented speedups to the Tide search engine. Other new features include a variety of new score functions in Tide, several enhancements to the Comet search engine, two new confidence estimation procedures, as well as three new tools: Param-medic,[2,3] Kojak,[4] and DIAmeter.[5]

2. METHODS

2.1. Data Sets

For the benchmarking in Sections 3.1 and 3.2, we selected at random one raw file (20190601_QX6_JoMu_SA_uPac200cm_HepG2_f4.raw) from a human sample in a recent large-scale study[6] (PRIDE accession PXD014877). The file contains 178 024 spectra. For the Param-Medic analyses in Section 3.4.1 we analyzed all 26 RAW files associated with PRIDE project PXD004424.

Crux is capable of analyzing RAW files directly, but only on a Windows machine. Because our analyses were performed on Linux systems, all RAW files were first converted to an open format using ThermoRawFileParser.[7] Supporting Information S2 summarizes all of the file formats used by Crux, both for input and output.

2.2. Protein Databases

Searches were conducted against the human reference proteome file (uniprot-proteome_UP000005640.fasta) downloaded from Uniprot on Feb 3, 2022. The fasta file contains canonical and isoform protein sequences.

https://doi.org/10.1021/acs.jproteome.2c00615
J. Proteome Res. 2023, 22, 561−569

2.3. Search Engines

In the comparison of search engines, we tried to ensure that comparable settings were employed between Comet and Tide (Table 1). Note that when switching to the exact p-value score function in Tide, we were obliged to set `mz-bin-width` to 1.0005079, and for the combined p-value score function, we used `--mz-bin-width` 1.0005079 and `--fragment-tolerance` 0.02. The database search was carried out on a Linux server equipped with an Intel Xeon CPU E5−2640 v4 2.40 GHz processor with 20 cores and 1TB SSD storage. Although both Comet and Tide allow multiple threads, the searches performed here use a single thread.

3. RESULTS

3.1. Tide Speedups and New Score Functions

We begin our analysis with a timing comparison of various score functions, as implemented in Crux's two DDA database search tools, Tide and Comet. In its initial implementation, Tide was markedly faster than competing search engines.[8] However, subsequent modifications to the code to implement new features and new score functions led to a decrease in Tide's efficiency. Consequently, we recently overhauled the Tide code with a focus on speeding it up, yielding a 3-fold increase in speed relative to the previous version of Tide (Tables 2 and 3). As a result, Tide is now quite efficient (Figure 2A), capable of searching the tryptic human proteome at ∼750 spectra/s. In particular, in its fastest mode, Tide searching is around 4.5 times faster than Comet searching. In addition, in the previous version of Crux, a bug occasionally prevented tide-search from running successfully with multiple threads. This bug has been fixed, and now tide-search runs stably on multithreaded systems. The search time comparisons using 8 threads can be found in Supporting Information S3.

Table 2. Running Time Comparison of Two Versions of Tide[a]

search	Old Crux	New Crux
Tide XCorr	1230	365
Tide Tailor	1250	284
Tide p-value	3640	813
Tide combined	16 300	6910
Comet	1140	1670

[a]Running time in seconds of Tide with four different score functions (XCorr, Tailor, exact p-value, and combined p-value) and Comet in the old (v3.2) versus the new version (v4.1-36) of Crux. The search was performed with data described in Sections 2.1−2.3.

Tide recently introduced a new scoring scheme, called Tailor calibration, which calibrates the top PSM score relative to the full distribution of scores generated during the database search. In this sense, it is similar to the E-value calibration implemented in Comet.[9] Specifically, Tailor considers the PSM scores $s_1, s_2, \ldots, s_N$, (in decreasing order) when matching one experimental spectrum to a set of N candidate peptides. Tailor calibration identifies the 99th quantile of this distribution by selecting the PSM score at the position $i^* = [N/100]$, where $[.]$ denotes the standard rounding operation. The Tailor method calibrates the top PSM score s_1 by $\tilde{s}_1 = \frac{s_1}{s_{i^*}}$. Tailor is thus a simple and quick method for score calibration.

From the user's perspective, speed is only useful in conjunction with accurate results. Accordingly, we compared the statistical power of various search strategies by counting the number of peptide-spectrum matches (PSMs) accepted at a 1% false discovery rate (FDR) threshold, as estimated using target-decoy competition. The results show several expected trends (Figure 2B). First, the raw XCorr score, as implemented in either Comet or Tide, does not perform as well as the corresponding calibrated score (the Comet E-value or Tide's Tailor score[10]). Tide also includes an alternative calibrated score, the "exact p-value," that is estimated using a dynamic programming procedure.[11] However, the exact p-value is designed to work with data that is generated using low-resolution fragment scans, so it actually yields decreased statistical power on the high-resolution data we used. Tide's "combined p-value" score is designed to combat this problem by combining the exact p-value with another dynamic programming procedure that operates on pairs of amino acids.[12] This score yields the best overall performance but is markedly slower to compute.

3.2. Confidence Estimation Procedures

The Tide search engine now supports two new procedures to improve statistical confidence estimation. The first procedure, known as cascade search,[13] aims to boost statistical power—that is, the number of peptides detected at a specified FDR threshold. Cascade search is applicable when the peptide database can be divided into groups a priori, and the groups can be ordered from more likely peptides toward more rare peptides. Cascade search works by sequestering at each stage any spectrum that is identified with a specified statistical confidence and then searching the remaining spectra against the next database in the list. For instance, such a cascade of databases could include fully tryptic, semitryptic, and non-enzymatic peptides or peptides with increasing numbers of modifications.

To demonstrate the empirical benefit of cascade search, we analyzed a sample data set in two ways: using a single peptide database followed by FDR control with target-decoy competition (TDC), and using cascade search with respect to a series of databases created using fully tryptic, semitryptic and nonenzymatic digestion. In Crux, cascade search is implemented as a separate command (`cascade-search`)

Table 3. Running Time Comparison for Tide and Comet[a]

number of peptides	Tide XCorr	Tide Tailor	Tide p-value	Tide combined	Comet E-value
60 896 400	241	251	646	7140	1100
40 484 062	211	219	581	5570	807
24 777 903	188	193	566	4450	694
13 635 673	171	173	551	3660	631
7 461 453	159	161	531	3070	599

[a]All times are reported in seconds. The data corresponds to Figure 2A.

https://doi.org/10.1021/acs.jproteome.2c00615
J. Proteome Res. 2023, 22, 561−569

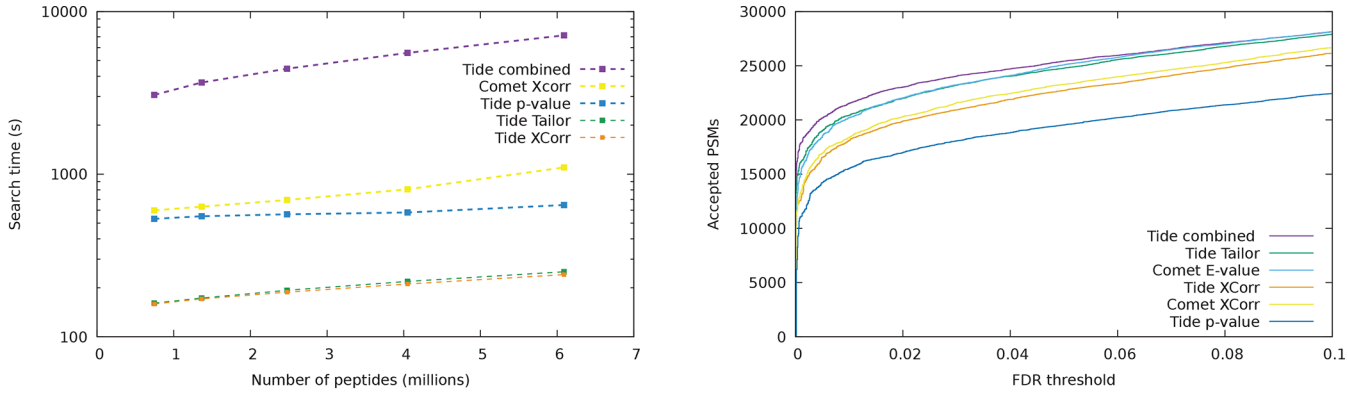


Figure 2. Comparisons of search tools. (A) Total running time of Tide and Comet, as a function of database size. The series correspond to Comet and Tide with four different score functions (XCorr, Tailor, exact p-value, and combined p-value). The search was performed with data described in Sections 2.1−2.3. The proteome was randomly downsampled to contain the specified number of peptides. Detailed timing information is provided in Table 3. (B) Plot of the number of accepted PSMs as a function of q-value threshold. The series correspond to two different Comet scores (XCorr and E-value) and Tide with four different score functions (XCorr, Tailor, exact p-value, and combined p-value). The search was performed with data described in Sections 2.1−2.3. All q-values are assigned using target-decoy competition, as implemented in `assign-confidence` in Crux.

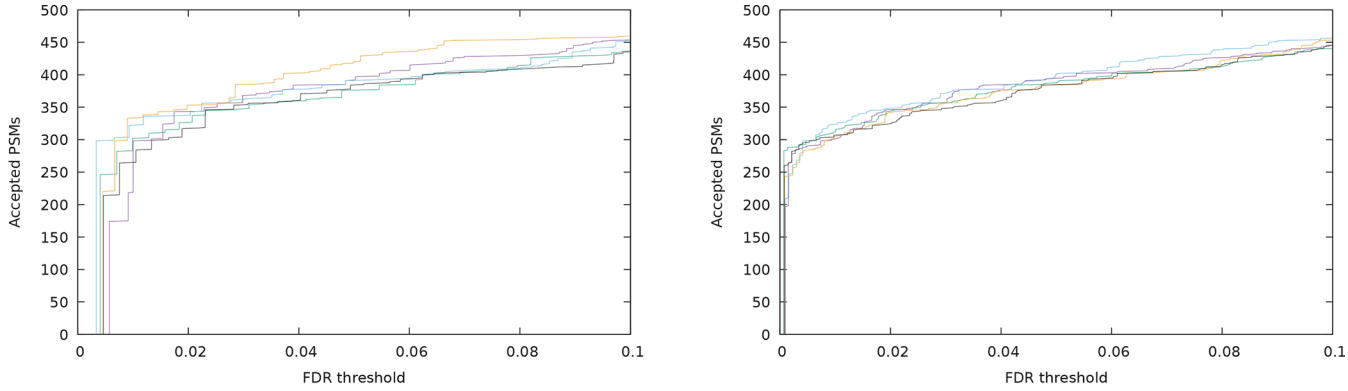


Figure 3. Average target-decoy competition reduces decoy-induced variance. (A) Plots of the number of accepted PSMs (*y*-axis) as a function of FDR threshold (*x*-axis), for searches against databases of varying size. Each series is generated by searching a different, randomly shuffled decoy database. (B) Similar to panels (A), except that each of the five series in the plot corresponds to FDR estimates from aTDC, using five decoys per target.

that takes as input one or more spectrum files plus a comma-separated list of Tide indices. For this experiment, we used the same human data set as before (described in Sections 2.1−2.3). We observe that at 1% FDR, cascade search accepts 27 400 PSMs, whereas a single-database Tide search accepts only 20 448, 25 325, or 23 046 PSMs, depending on whether the database is tryptic, semitryptic, or nonenzymatic. Thus, cascade-search leads to an increase in the number of accepted PSMs between 8−34% at 1% FDR.

Note that the cascade search procedure, in this case, is somewhat inefficient because the three databases are supersets of one another; for example, all tryptic peptides are also included in the semitryptic database. To avoid this inefficiency, Crux provides an auxiliary command, `subtract-index`, that will remove from one Tide index all peptides that occur in a second index.

The second new procedure aims to reduce the variance in FDR estimates that is intrinsic to any decoy-based confidence estimation method. The procedure, called "average target-decoy competition" (aTDC),[14,15] works by searching a given set of spectra against a collection of peptide databases: one databases containing target peptides and multiple database containing shuffled decoy peptides. In Crux, aTDC is implemented via the `num-decoys-per-target` parameter of `tide-index`. Setting this parameter to any integer >1 will cause Tide to carry out aTDC.

We demonstrated the utility of aTDC using the same human data set as before (described in Section 2.1−2.3). In practice, averaging is most useful when the total number of discoveries is small, because in this setting the decoy-induced variance in the estimated FDR can have a substantial impact on the results. Accordingly, to simulate such a scenario, we searched a database containing 100 proteins selected at random from the human proteome. In this setting, the variability that we observe in the FDR estimates from standard TDC is substantially reduced when we use aTDC with five decoys per target (Figure 3). For example, at a 1% FDR threshold, the standard deviation in the number of accepted PSMs decreases by 83%, from 42 to 7.

https://doi.org/10.1021/acs.jproteome.2c00615
J. Proteome Res. 2023, 22, 561−569

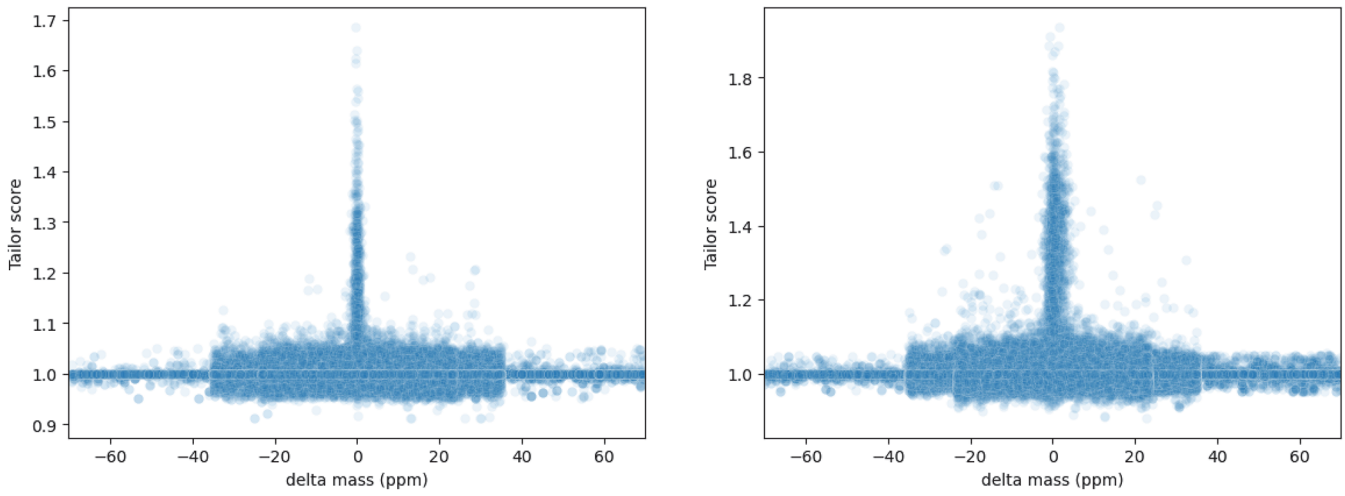


Figure 4. Comparison of precursor acquisition in two different runs. (A) Plots for each PSM produced by searching sample 151009_exo3_5 against the human proteome, the Tailor score (*y*-axis) as a function of the difference between the observed precursor mass and the peptide mass (*x*-axis). To show a broad range of values, the search was performed with a precursor window size of 70 ppm. For this data, Param-Medic infers a precursor window size of 16.79 ppm. (B) Same as panel (A), but for 151218_exo4_4. The inferred precursor window size is 68.48 ppm.

3.3. Comet Updates

Since the last Crux overview paper in 2014, the Comet search tool has incorporated many updates and bug fixes.

One feature that has been extended for analysis flexibility, based on requests by various researchers attempting to optimize specific analysis, is the control of how variable modifications are applied. This includes distance constraints of modifications from peptide or protein termini, forcing the requirement of a modification to be present in a peptide, including the ability to specify the minimum and maximum number of each variable modification, controlling whether or not a variable modification can appear on the C-terminal residue, and consideration of neutral loss peaks on those fragment ions that contain a variable modification.

Comet was also one of the first search tools to support the Proteomics Standards Initiative's Extended Fasta Format (PEFF).[16] Comet's initial published PEFF support included the ability to search PEFF database files to analyze the annotated modifications and single amino acid substitutions.[17] More recently, Comet's PEFF support has been extended to include the ability to analyze "VariantComplex" annotations which encode sequence variations that are more complex than a single amino acid substitution. VariantComplex annotations can encode deletions, insertions, and combinations of the two, which allows the PEFF database to encapsulate sequence variations such as protein isoforms within a single sequence entry.

Comet was also extended to support the real-time search application that was initially implemented in the Schweppe lab's Orbiter platform for real-time instrument control.[18] Subsequently, Comet's real-time search application has been adopted by Thermo Scientific and is now available for real-time analysis on their Tribrid mass spectrometers, typically for support of tandem mass tag workflows to increase unique data depth

3.4. New Tools

3.4.1. Param-Medic. The Param-Medic command automatically infers several key characteristics—precursor window size, fragment ion tolerance, and the presence of several common types of post-translational modifications—of a given MS/MS data set by examining the MS1 and MS2 spectra. The primary goal is to facilitate automated processing of public data sets, when metadata such as instrument settings may be hard to come by. Param-Medic can also be useful to identify problems with a data set, for example, when the nominal mass accuracy of the data disagrees with the mass accuracy inferred by the program.

To demonstrate Param-Medic's utility, we downloaded all 26 RAW files associated with PRIDE identifier PXD004424 and subjected them to Param-Medic analysis. Notably, the results suggested a fairly broad range of precursor window sizes, ranging from 16.79 ppm up to 68.48 ppm, whereas the authors of the original study used a 20 ppm window for all of the analyses.[19] To follow up on this assessment, we selected two specific RAW files, one with the minimum inferred window size of 16.79 ppm (151009_exo3_5) and one with the maximum inferred window size of 68.48 ppm (151218_exo4_4). The relationship between the search engine score and delta mass shows a notably broader distribution for the second file, including a handful of outlier points with high Tailor scores (Figure 4), potentially indicative of problematic acquisition.

Note that Param-Medic can be called automatically from within Tide or Comet by using the `auto-modifications-spectra`, `auto-precursor-window`, and `auto-bin-width` options.

3.4.2. Kojak

Kojak performs database search on mass spectra from cross-linked samples.[4] Similar to other cross-linked database search algorithms such as plink2,[20] XLinkX,[21] and XiSearch,[22] Kojak identifies the amino acid sequences of peptides that have been covalently linked together using chemical cross-linkers, a common technique in proteomics for studying protein structure and interactions.[23] Cross-link peptide sequence identification occurs by matching observed fragment ions from MS2 spectra following collisional dissociation and considering unique ion masses that occur due to the tethering

https://doi.org/10.1021/acs.jproteome.2c00615
J. Proteome Res. 2023, 22, 561−569

of two peptides. Kojak also supports analysis of cleavable cross-linkers, a feature shared with cross-linking tools such as those mentioned previously, as well as MS Annika[24] and MeroX,[25] and is capable of searching whole proteomes.

Here, we describe how to run Kojak on a cross-linked sample from PRIDE project PXD014337[26] and upload the results into the web-based platform ProXL[27] for visualization. Kojak takes as input mzML spectra data files and a fasta protein sequence file. For this analysis, we analyze the three DSS-linked replicate files. Spectral peaks should be transformed to centroid representation during the conversion from raw spectra to mzML format. Then, it is necessary to tailor a few Kojak parameters to the data:

```
fragment_bin_offset = 0.0
fragment_bin_size = 0.01
decoy_filter = DECOY 1
max_miscleavages = 2
min_spectrum_peaks = 25
spectrum_processing = true
top_count = 5
min_peptide_score = 0.25
```

These parameters can be specified on the command line or in the Crux parameter file. To run the Kojak analysis on all three data files at once, execute the following command:

```
crux kojak-parameter-file kojak.params.txt *.mzML Cas9_plus10.fasta
```

This analysis produces a series of files containing cross-linked spectrum matches (CSMs). The files contain the suggested peptide or peptides matched to each spectrum, but these matches must then be validated using a target-decoy approach with Percolator. CSMs are divided into several categories, and we want to validate the intraprotein and interprotein CSMs. To do this, rename the .txt extensions for *.perc.intra.* *.perc.inter.* files to .pin (e.g., XLpeplib_Beveridge_QEx-HFX_DSS_R1.perc.intra.txt becomes XLpeplib_Beveridge_QEx-HFX_DSS_R1.perc.intra.pin) so that Percolator can read them. Then execute the following command:

```
crux percolator --only psms T --tdc T *.pin
```

This command will combine all the Kojak intraprotein and interprotein CSMs into a single set for Percolator analysis and produce estimated error rates at the CSM-level. Using a q-value threshold of 0.01 to estimate a 1% error rate, 1944 CSMs are returned. Because we know the ground truth in this data set, we can compare the CSMs to the set of correct results, and find that 1919 are correct, and 25 are incorrect, for an error rate of 1.3%, or approximately the estimated error rate at the chosen threshold.

Visualization of the spectra and CSM annotations is done with ProXL. Instructions to convert and upload CSMs to ProXL are provided in Supporting Information S4.

3.4.3. DIAmeter

DIAmeter is a library-free database search tool for DIA data.[5] DIA data analysis tools can be loosely categorized into two types: (1) library-free methods such as Pecan,[28] DIA-Umpire,[29] and directDIA,[30] and (2) spectral library-based methods such as OpenSWATH,[31] DIA-NN,[32] and MaxDIA.[33] DIAmeter falls into the former category of library-free methods; therefore, DIAmeter does not rely on real or in silico spectral libraries, which can be expensive to produce or may not capture properties specific to a particular instrument or set of acquisition parameters. Some of the library-free methods work by first extracting pseudospectra and then searching with methods developed for conventional DDA data. However, the extraction of pseudospectra depends heavily on the quality of the precursor signals in the precursor scans; hence, pseudospectrum-based methods by design cannot detect peptides with undetectable precursor signals, which commonly arise due to limitations of intrascan dynamic range. DIAmeter, by contrast, operates directly on the DIA spectra.

The `diameter` command in Crux takes as input the DIA data and a user-specified database of proteins, which must first be indexed by the `tide-index` command. DIAmeter computes a series of scores for each candidate peptide and then calls Percolator internally to produce a ranked list of peptides, with associated confidence estimates (q-values).

A comparative evaluation of DIAmeter appears in the original publication describing the method.[5] Here, we demonstrate how to run the software and show that it gives consistent results on several DIA runs from a recently published study. In this analysis, we use data from a large-scale Alzheimer's study,[34] selecting three runs at random from the hippocampus brain region, batch 1. To search a file "HZR03.mzml" against the Uniprot human proteome ("human.fa") requires two steps:

1. Create a Tide index from the human reference proteome using the command `crux tide-index human.fa human`.
2. Search the mzML file against the index using the command `diameter --diameter-instrument orbitrap HZR03.mzml human`

For this particular file, DIAmeter detects 12 037 peptides. We also analyzed files from two other samples (HZR07 and HZR10) and detected similar numbers of peptides, with >8000 peptides detected in all three runs (Figure 5).

4. DISCUSSION

Crux provides a rich set of software tools for analyzing proteomics mass spectrometry data. In this paper, we have emphasized the newer aspects of the toolkit, focusing on improvements to our two standard DDA search engines, Comet and Tide, as well as the introduction of several new

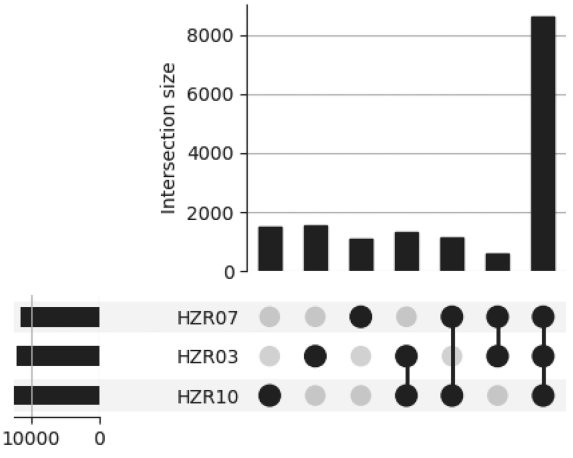


Figure 5. DIAmeter analysis of three Alzheimer's samples. Three samples from a recent Alzheimer's study[34] were searched against the Uniprot human reference proteome. The number of peptides that were detected at a 1% FDR threshold in all three runs, in any combination of two runs, and in single runs.

https://doi.org/10.1021/acs.jproteome.2c00615
J. Proteome Res. 2023, 22, 561−569

tools, Kojak, Param-Medic, and DIAmeter. Our aim is to ensure that the Crux software can be easily applied to many standard workflows, while also producing accurate results with high statistical power. Crux supports a variety of standard input and output formats, including mzIdentML output that can be directly uploaded to ProteomeXChange. Sample command lines for the new features described in this paper are presented in Supporting Information S5.

Of course, Crux is not the only software toolkit in bottom-up proteomics. Some of the most popular competing toolkits include MaxQuant,[35] Proteome Discoverer,[36] FragPipe,[37,38] and pFind Studio.[39] All of these toolkits provide a core search engine and thus have overlapping functionality with Crux. However, one thing that sets Crux apart from the tools listed above is that the Crux source code is publicly available. This is important, because reproducible science requires full access to source code.[40]

A common question for Crux users is which of the two primary search engines, Comet or Tide, should be used for a given analysis. The short answer is that, for many tasks, either search engine will work well. Both Comet and Tide are reimplementations of the original SEQUEST search engine, but they do differ somewhat in their functionality. First, as shown in Section 3.1, Tide is often markedly faster than Comet, especially when the Tailor score is employed. Second, the two search engines differ somewhat in the range of available options. For example, some options are available only in Comet—including the ability to read PEFF, the recently added flexibility in handling PTMs, and options related to different types of theoretical fragment ions, the maximum fragment charge state, and nucleotide reading frame—whereas other options are only available in Tide, including the various score functions described in Section 3.2, the ability to search with multiple decoys per target, and several options related to decoy peptide generation.

As mass spectrometry instrumentation and data collection technology advances, so too do the software tools used to make sense of mass spectrometry data. Accordingly, Crux is under constant development as we work with collaborators and other users of the software to ensure that it addresses their needs. We have a variety of tools planned for future releases, including labeled and label-free quantification tools akin to Libra[41] and FlashLFQ,[42] respectively, as well as a mass calibration tool similar to the procedures in MetaMorpheus[43] or MSFragger.[44] Crux users who have specific needs—including new tools to suggest, desired new functionality, or bugs to report—are encouraged to submit an issue to our Github issue tracker, which is linked from the main Crux web page, http://crux.ms.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00615.

Supporting Information S1: Table describing new features in Crux; Supporting Information S2: Table of input and output file formats in Crux.; Supporting Information S3: Running time results with 8 threads; Supporting Information S4: Instructions for visualizing Kojak results in ProXL; and Supporting Information S5: Sample command lines (PDF)

Supplementary Data File S1: `commands.zip` — Zipped sample data and bash scripts for Supporting Information S5 including `Cas9_plus10.fasta` fasta file for Kojak analysis, `HZR03.mzml` mass spectrometry data, `uniprot-proteome_UP000005640+reviewed_yes.fasta` fasta file, `UPS1.mzML.gz` mass spectrometry data, and `script.sh` sample linux commands to run crux (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

William Stafford Noble − *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States; Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-2350, United States;* orcid.org/0000-0001-7283-4715; Email: E-mail

Authors

Attila Kertesz-Farkas − *Department of Data Analysis and Artificial Intelligence and Laboratory on AI for Computational Biology, Faculty of Computer Science, HSE University, Moscow 101000, Russia*

Frank Lawrence Nii Adoquaye Acquaye − *Department of Data Analysis and Artificial Intelligence and Laboratory on AI for Computational Biology, Faculty of Computer Science, HSE University, Moscow 101000, Russia*

Kishankumar Bhimani − *Department of Data Analysis and Artificial Intelligence and Laboratory on AI for Computational Biology, Faculty of Computer Science, HSE University, Moscow 101000, Russia*

Jimmy K. Eng − *Proteomics Resource, University of Washington, Seattle, Washington 98109-4725, United States;* orcid.org/0000-0001-6352-6737

William E. Fondrie − *Talus Bioscience, Seattle, Washington 98122, United States;* orcid.org/0000-0002-1554-3716

Charles Grant − *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States*

Michael R. Hoopmann − *Insititute for Systems Biology, Seattle, Washington 98109, United States*

Andy Lin − *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0003-0072-612X

Yang Y. Lu − *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States*

Robert L. Moritz − *Insititute for Systems Biology, Seattle, Washington 98109, United States*

Michael J. MacCoss − *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0003-1853-0256

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00615

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded in part by National Institutes of Health grants from the National Institute General Medical Sciences R01GM087221, the National Heart, Lung, and Blood Institute R01HL133135, the Office of the Director S10OD026936, and

the National Institute on Aging U19AG023122, and by the National Science Foundation award 1920268.

REFERENCES

(1) McIlwain, S.; Tamura, K.; Kertesz-Farkas, A.; Grant, C. E.; Diament, B.; Frewen, B.; Howbert, J. J.; Hoopmann, M. R.; Käll, L.; Eng, J. K.; MacCoss, M. J.; Noble, W. S. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **2014**, *13*, 4488−4491.

(2) May, D. H.; Tamura, K.; Noble, W. S. Param-Medic: A tool for improving MS/MS database search yield by optimizing parameter settings. *J. Proteome Res.* **2017**, *16*, 1817−1824.

(3) May, D. H.; Tamura, K.; Noble, W. S. Detecting modifications in proteomics experiments with Param-Medic. *J. Proteome Res.* **2019**, *18*, 1902−1906.

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

(5) Lu, Y. Y.; Bilmes, J.; Rodriguez-Mias, R. A.; Villén, J.; Noble, W. S.. "*DIAmeter: Matching Peptides to Data-Independent Acquisition Mass Spectrometry Data*; ISMB. 2021.

(6) Muller, J. B.; Geyer, P. E.; Colaco, A. R.; Treit, P. V.; Strauss, M. T.; Oroshi, M.; Doll, S.; Virreira Winter, S.; Bader, J. M.; Kohler, N.; Theis, F.; Santos, A.; Mann, M.; et al. The proteome landscape of the kingdoms of life. *Nature* **2020**, *582*, 592−596.

(7) Hulstaert, N.; Shofstahl, J.; Sachsenberg, T.; Walzer, M.; Barsnes, H.; Martens, L.; Perez-Riverol, Y. ThermoRaw-FileParser: modular, scalable and cross-platform RAW file conversion. *J. Proteome Res.* **2020**, *19*, 537−542.

(8) Diament, B.; Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.* **2011**, *10*, 3871−3879.

(9) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open source tandem mass spectrometry sequence database search tool. *Proteomics* **2013**, *13*, 22−24.

(10) Sulimov, P.; Kertész-Farkas, A. Tailor: A Nonparametric and Rapid Score Calibration Method for Database Search-Based Peptide Identification in Shotgun Proteomics. *J. Proteome Res.* **2020**, *19*, 1481−1490.

(11) Howbert, J. J.; Noble, W. S. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Mol. Cell. Proteomics* **2014**, *13*, 2467−2479.

(12) Lin, A.; Howbert, J. J.; Noble, W. S. Combining High-Resolution and Exact Calibration To Boost Statistical Power: A Well-Calibrated Score Function for High-Resolution MS2 Data. *J. Proteome Res.* **2018**, *17* (11), 3644−3656.

(13) Kertesz-Farkas, A.; Keich, U.; Noble, W. S. Tandem mass spectrum identification via cascaded search. *J. Proteome Res.* **2015**, *14*, 3027−3038.

(14) Keich, U.; Noble, W. S.. "Progressive calibration and averaging for tandem mass spectrometry statistical confidence estimation: Why settle for a single decoy" Lecture Notes in Computer Science. In *Proceedings of the International Conference on Research in Computational Biology (RECOMB)*; Sahinalp, S., Ed.; Springer, 2017; Vol. *10229*; pp 99−116.

(15) Keich, U.; Tamura, K.; Noble, W. S. Averaging strategy to reduce variability in target-decoy estimates of false discovery rate. *J. Proteome Res.* **2019**, *18*, 585−593.

(16) Binz, P.-A.; Shofstahl, J.; Vizcaino, J. A.; Barsnes, H.; Chalkley, R. J.; Menschaert, G.; Alpi, E.; Clauser, K.; Eng, J. K.; Lane, L.; Seymour, S. L.; Sanchez, L. F. H.; Mayer, G.; Eisenacher, M.; Perez-Riverol, Y.; Kapp, E. A.; Mendoza, L.; Baker, P. R.; Collins, A.; Van Den Bossche, T.; Deutsch, E. W.; et al. Proteomics standards initiative extended FASTA format. *J. Proteome Res.* **2019**, *18*, 2686−2692.

(17) Eng, J. K.; Deutsch, E. W. Extending Comet for global amino acid variant and post-translational modification analysis using the PSI extended FASTA format. *Proteomics* **2020**, *20*, 1900362.

(18) Schweppe, D. K.; Eng, J. K.; Yu, Q.; Bailey, D.; Rad, R.; Navarrete-Perea, J.; Huttlin, E. L.; Erickson, B. K.; Paulo, J. A.; Gygi, S. P. Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *J. Proteome Res.* **2020**, *19*, 2026−2034.

(19) Cypryk, W.; Lorey, M.; Puustinen, A.; Nyman, T. A.; Matikainen, S. Proteomic and bioinformatic characterization of extracellular vesicles released from human macrophages upon influenza A virus infection. *J. Proteome Res.* **2017**, *16*, 217−227.

(20) Chen, Z.-L.; Meng, J.-M.; Cao, Y.; Yin, J.-L.; Fang, R.-Q.; Fan, S.-B.; Liu, C.; Zeng, W.-F.; Ding, Y.-H.; Tan, D.; Wu, L.; Zhou, W.-J.; Chi, H.; Sun, R.-X.; Dong, M.-Q.; He, S.-M.; et al. A high-speed search engine pLink2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **2019**, *10*, 1−12.

(21) Liu, F.; Lössl, P.; Scheltema, R.; Viner, R.; Heck, A. J. "Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification" In *Nature Communications*; 2017, XlinkX; Vol. *8.1*, pp 1−8.

(22) Mendes, M. L; Fischer, L.; Chen, Z. A; Barbon, M.; O'Reilly, F. J; Giese, S. H; Bohlke-Schneider, M.; Belsom, A.; Dau, T.; Combe, C. W; Graham, M.; Eisele, M. R; Baumeister, W.; Speck, C.; Rappsilber, J.; et al. An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **2019**, *15*, No. e8994.

(23) Leitner, A.; Bonvin, A. M.J.J.; Borchers, C. H.; Chalkley, R. J.; Chamot-Rooke, J.; Combe, C. W.; Cox, J.; Dong, M.-Q.; Fischer, L.; Gotze, M.; Gozzo, F. C.; Heck, A. J.R.; Hoopmann, M. R.; Huang, L.; Ishihama, Y.; Jones, A. R.; Kalisman, N.; Kohlbacher, O.; Mechtler, K.; Moritz, R. L.; Netz, E.; Novak, P.; Petrotchenko, E.; Sali, A.; Scheltema, R. A.; Schmidt, C.; Schriemer, D.; Sinz, A.; Sobott, F.; Stengel, F.; Thalassinos, K.; Urlaub, H.; Viner, R.; Vizcaino, J. A.; Wilkins, M. R.; Rappsilber, J.; et al. Toward increased reliability, transparency, and accessibility in cross-linking mass spectrometry. *Structure* **2020**, *28*, 1259−1268.

(24) Pirklbauer, G. J.; Stieger, C. E.; Matzinger, M.; Winkler, S.; Mechtler, K.; Dorfer, V. MS Annika: A new cross-linking search engine. *J. Proteome Res.* **2021**, *20*, 2560−2569.

(25) Götze, M.; Pettelkau, J.; Fritzsche, R.; Ihling, C. H.; Schäfer, M.; Sinz, A. Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 83−97.

(26) Beveridge, R.; Stadlmann, J.; Penninger, J. M.; Mechtler, K. A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes. *Nat. Commun.* **2020**, *11*, 1−9.

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

(28) Ting, Y. S.; Egertson, J. D.; Bollinger, J. G.; Searle, B.; Payne, S. H.; Noble, W. S.; Mac-Coss, M. J. PECAN: a library free peptide detection tool for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **2017**, *14*, 903−908.

(29) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: a comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **2015**, *12*, 258−264.

(30) Mehta, D.; Scandola, S.; Uhrig, R. G. "Direct data-independent acquisition (direct DIA) enables substantially improved label-free quantitative proteomics in Arabidopsis" In *bioRxiv*; 2020, https://www.biorxiv.org/content/10.1101/2020.11.07.372276.

(31) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinovic, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmstrom, J.; Malmstrom, L.; Aebersold, R. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Methods* **2014**, *32*, 219−223.

(32) Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **2020**, *17*, 41−44.

https://doi.org/10.1021/acs.jproteome.2c00615
J. Proteome Res. 2023, 22, 561−569

(33) Sinitcyn, P.; Hamzeiy, H.; Salinas Soto, F.; Itzhak, D.; McCarthy, F.; Wichmann, C.; Steger, M.; Ohmayer, U.; Distler, U.; Kaspar-Schoenefeld, S.; Prianichnikov, N.; Yılmaz, S.; Rudolph, J. D.; Tenzer, S.; Perez-Riverol, Y.; Nagaraj, N.; Humphrey, S. J.; Cox, J.; et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* **2021**, *39*, 1563−1573.
(34) Hubbard, E. E.; Heil, L. R.; Merrihew, G. E.; Chhatwal, J. P.; Farlow, M. R.; McLean, C. A.; Ghetti, B.; Newell, K. L.; Frosch, M. P.; Bateman, R. J.; Larson, E. B.; Keene, C. D.; Perrin, R. J.; Montine, T. J.; MacCoss, M. J.; Julian, R. R.; et al. Does data-independent acquisition data contain hidden gems? A case study related to Alzheimer's disease. *J. Proteome Res.* **2022**, *21*, 118−131.
(35) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.
(36) Orsburn, B. C. ProteomeDiscoverer−A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **2021**, *9*, 15.
(37) 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.
(38) Teo, G. C.; Polasky, D. A.; Yu, F.; Nesvizhskii, A. I. Fast deisotoping algorithm and its implementation in the MSFragger search engine. *J. Proteome Res.* **2021**, *20*, 498−505.
(39) Li, D.; Fu, Y.; Sun, R.; Ling, C. X.; Wei, Y.; Zhou, H.; Zeng, R.; Yang, Q.; He, S.; Gao, W. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **2005**, *21*, 3049−3050.
(40) Heil, B. J.; Hoffman, M. M.; Markowetz, F.; Lee, S.-I.; Greene, C. S.; Hicks, S. C. Reproducibility standards for machine learning in the life sciences. *Nat. Methods* **2021**, *18*, 1132−1135.
(41) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150−1159.
(42) Millikin, R.; Solntsev, S.; Shortreed, M.; Smith, L. Ultrafast Peptide Label-Free Quantification with FlashLFQ. *J. Proteome Res.* **2018**, *17*, 386−391.
(43) 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.
(44) Yu, X.; Lin, J.; Zack, D. J.; Qian, J. Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors. *BMC Bioinformatics* **2007**, *8*, 437.

https://doi.org/10.1021/acs.jproteome.2c00615
J. Proteome Res. 2023, 22, 561−569