nature communications

Article

https://doi.org/10.1038/s41467-023-39828-0

Detecting diagnostic features in MS/MS spectra of post-translationally modified peptides

Received: 4 September 2022

Accepted: 23 June 2023

Published online: 12 July 2023

Check for updates

Daniel J. Geiszler[1], Daniel A. Polasky[2], Fengchao Yu[2] & Alexey I. Nesvizhskii[1,2]

Post-translational modifications are an area of great interest in mass spectrometry-based proteomics, with a surge in methods to detect them in recent years. However, post-translational modifications can introduce complexity into proteomics searches by fragmenting in unexpected ways, ultimately hindering the detection of modified peptides. To address these deficiencies, we present a fully automated method to find diagnostic spectral features for any modification. The features can be incorporated into proteomics search engines to improve modified peptide recovery and localization. We show the utility of this approach by interrogating fragmentation patterns for a cysteine-reactive chemoproteomic probe, RNA-crosslinked peptides, sialic acid-containing glycopeptides, and ADP-ribosylated peptides. We also analyze the interactions between a diagnostic ion's intensity and its statistical properties. This method has been incorporated into the open-search annotation tool PTM-Shepherd and the FragPipe computational platform.

Post-translational modifications (PTMs) have long been of interest to proteomics researchers because of their central role in regulating cellular functions. Processes to maximize their recovery run the gamut of proteomics techniques, from sample preparation[1] to instrumental acquisition[2] and computational analysis[3–5]. At the computational level, proteomics search engines have grown in their capacity to identify PTMs. For PTMs with complex fragmentation patterns like glycosylation that exhibit multiple modes of fragmentation, entire search engines specific to the modification class have been developed[4,6,7]. Despite this work, many modifications continue to suffer from low recall in standard high-throughput workflows due to their behavior during tandem mass spectrometry (MS) analysis, producing unexpected or difficult fragmentation patterns that frustrate search engines[8]. Even small changes to workflows—such as the addition of isobaric labels—can alter fragmentation patterns and reduce or preclude identification of even the best-studied PTMs[9]. Recent work with synthetic peptides carrying less well-studied PTMs demonstrated that many diagnostic ions and neutral losses have yet to be identified[10].

With the proliferation of synthetic PTMs[11]—particularly ones that alter fragmentation patterns[9]—and new instrumental methods[2,12], keeping search engines up to date with knowledge of how an analyte will fragment in a particular setting is a large task. To overcome this, computational tools are being developed to identify modification fragmentation patterns without prior knowledge. The first such tools only identified diagnostic ions and were limited in their applications[13], but newer approaches have incorporated additional features. Synthetic peptides bearing modifications are generally seen as gold standards to study PTM fragmentation patterns and methods have been developed to extract them from spectra[14], but this approach adds additional benchwork to proteomics experiments. Furthermore, optimal search parameters are fragmentation-dependent and can change based on experimental settings, which requires reprocessing mass spectrometry data and reanalyzing fragmentation patterns for multiple experiment types. Zolg et al. developed a method to do this in a high-throughput manner[10], but it requires paired modified-unmodified peptides and cannot be easily reimplemented by other research groups. Their approach to identify neutral losses also

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [2]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. e-mail: nesvi@med.umich.edu

requires both the intact and fragmented peak to be present in the spectrum at consistent distances, precluding finding complete losses and many charged losses. Other approaches to score PSMs from modified peptides are trained for specific PTMs[15] or perform model refinement that focuses on distances between experimental peaks, discarding information about matched ions from the peptide backbone that would dramatically reduce the required training dataset size[16]. Chemoproteomics has a particular stake in this effort due to the diversity of probes employed[17–19]. However, existing tools for chemoproteomics require isotopic labeling signatures to be present at the MS1 and MS2 levels[20]. This limits their applications to chemical probes that are labeled non-isobarically, thus they cannot be used for some PTM probes[21], biological PTMs, or the development of isobaric mass tags[22]. In prior work, we have found that understanding PTM fragmentation patterns allowed us to maximize modified peptide recovery and localization. Thus, when studying a cysteine chemoproteomic probe, we developed a method to extract its diagnostic spectral features to improve coverage of the ligandable cysteineome[19]. Our approach did not require synthesizing standards or isotopically labeled peptides and facilitated the discovery of partial modification losses and diagnostic ions, ultimately leading to the identification of three diagnostic ions and two partial PTM fragmentation events that escaped manual inspection.

Here, we present an improved, fully automated, and empirically tuned implementation of our diagnostic feature extraction algorithm to study and score the fragmentation patterns of modifications. Our approach detects three separate types of diagnostic features—diagnostic ions, peptide remainder masses, and fragment remainder masses—and can be used in any experimental setting, including for the simultaneous characterization of multiple modifications and when only a handful of PSMs are present for a modification. We demonstrate the robustness of our technique by applying it at both massive and small scale, and across synthetic and biological PTMs. Finally, we perform a meta-analysis of diagnostic features and discuss how these can be used to further PTM discovery in diverse settings. Our method has been implemented within PTM-Shepherd[23] and is freely available as part of the FragPipe suite of tools (https://fragpipe.nesvilab.org/).

Results

Algorithm overview

The PTM-Shepherd diagnostic feature mining module aims to perform high throughput identification of spectral features that can be used to identify post-translational modifications (PTMs), facilitating the validation or discovery of PTM-specific signals. Probable modifications from an experiment are identified by passing the results of open or mass offset search to PTM-Shepherd. For each MS1 mass shift, PTM-Shepherd identifies enriched diagnostic features across three categories: diagnostic ions; mass shifts from the unmodified, intact peptide ions (peptide remainder masses); and mass shifts from unmodified fragment ions (fragment remainder masses). This module operates in three steps: calculating all possible spectral features for every peptide-spectrum match (PSM) with a particular mass shift, identifying the most abundant spectral features for every identified mass shift within each category, then finally performing statistical tests and filtering to see whether those features can be used to infer the presence of the modification via comparison to unmodified peptides. This module uses as input decharged and deisotoped MGF spectra produced by MSFragger[24], so the maximum charge state for all ions in MS/MS spectra is assumed to be one. Spectral ions are normalized to the base peak and only the top 150 peaks are considered (by default).

We illustrate our technique using a cysteine-specific chemical probe[19] that we previously analyzed with an early version of the algorithm, identifying all three ion types (Fig. 1a). The first step in our strategy is to calculate all possible diagnostic spectral features for each PSM within a mass shift identified by PTM-Shepherd. Any ions from experimental spectra that do not belong to the peptide are considered

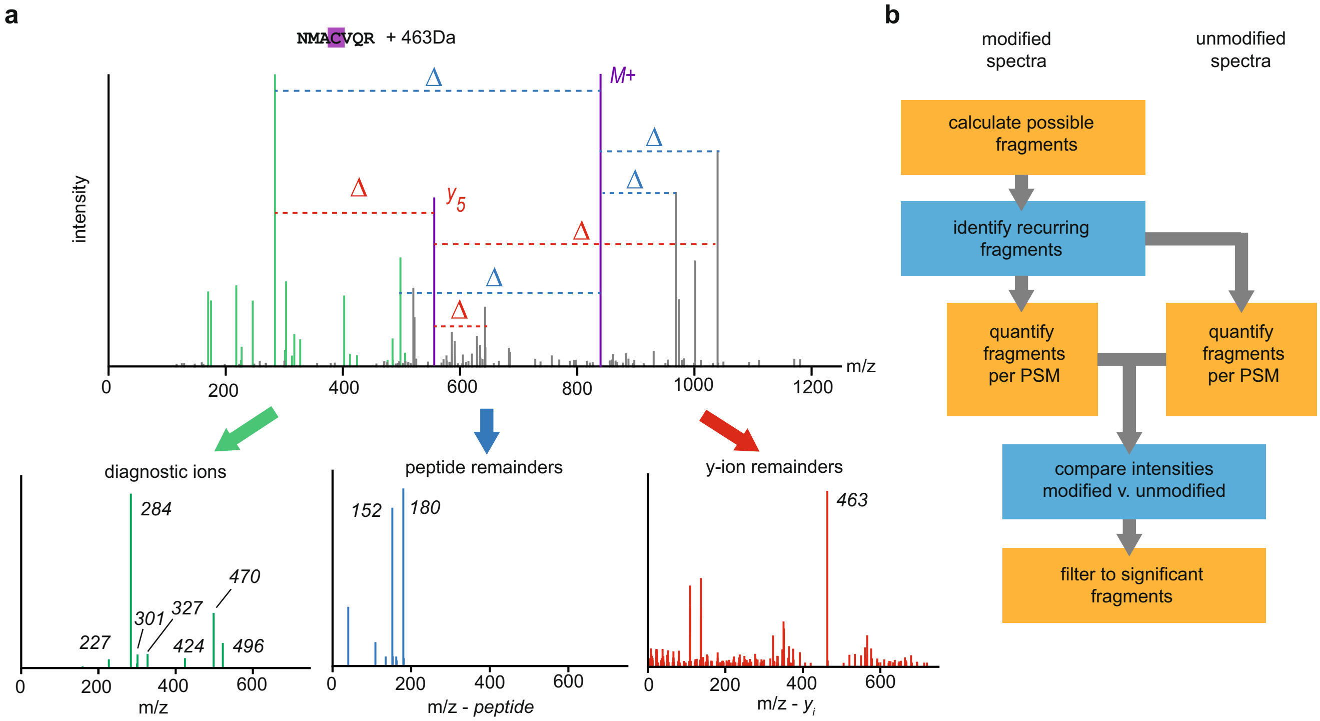

Fig. 1 | Implementation of the diagnostic feature mining algorithm. **a** Calculation of all possible diagnostic ions, peptide remainders, and fragment remainders, with the latter two calculated with respect to theoretical unmodified ions, for a synthetic Cys probe of mass 463 Da. Calculated features across spectra are subsequently pooled to find recurring features. Source data are provided as a Source Data file. **b** Workflow for diagnostic feature selection.

Table 1 | Most prevalent mass shifts and their characteristics from an open search of a pRBS-ID experiment

peak apex	PSMs	% also in unmodified	mass annotation	ret. time shift	localized PSMs	n-term. localization rate	AA1	AA1 enrichment score	AA1 PSM count
0.0002	7751	100		−0.6	291				
226.0594	4241	34.81	226.0594 mass-shift	−449.3	1343		R	2.0	160
1.0032	2378	88.62	+1 isotopic peak	25.6	658		F	1.6	65
57.0216	945	80	Iodoacetamide derivative	−63.8	806	38.94	H	11.9	184
94.0168	830	50.81	94.0168 mass-shift	−477.6	799	7.47	H	9.2	140
−48.0032	568	35.35	Homoserine lactone	−1238.7	559	26.23	M	53.0	373

Two unannotated mass shifts were determined to be the most likely candidates for RNA moieties. This was confirmed by their similar effects on retention time.

potential diagnostic features for the mass shift. To identify recurring features for the mass shift, calculated features for every spectrum from the mass shift are sent to a common histogram. Peaks are identified from here and shuttled to downstream analysis. For diagnostic ions, the unannotated ions from the experimental spectrum are sent to their histogram as they are (Fig. 1a, green). Peptide remainder masses are calculated by computing mass differences between the theoretical, unshifted peptide ion (purple) and all ions in the spectrum (blue). Fragment remainder masses are calculated by iteratively computing mass differences between every theoretical ion from the peptide backbone (purple) and all ions in the spectrum.

Finding recurring ions does not mean that they are useful for identifying a mass shift. Our ion set contains features that might be abundant across the entire dataset, so it is necessary to remove baseline noise. We do this by comparing the recovered features from all spectra bearing the mass shift to those of unmodified peptides in bulk as a proxy for dataset background (Fig. 1b). For every feature detected in the prior step, it is quantified across modified and unmodified PSMs, with missing ions or offsets encoded as zeroes. The result is two lists of intensities, from which we can perform statistical tests. Encoding missing ions as zeroes is necessary for this step, but it can also produce a range of non-normal distributions, calling for the non-parametric Mann-Whitney-U test. Features that are significantly different between the modified and unmodified lists are then filtered for sensitivity criteria (minimum prevalence in the modified bin) and mean intensity fold-change between the two bins. Fragment remainder ions undergo an additional layer of filtering for ion formation propensity, where they are required to represent a minimum percentage of the number of ions in their series. True fragment remainder ions can also create "echoes" of their masses that are combinations of the original mass and adjacent amino acids, multiple of which can pass filtering for a mass shift. We correct these by checking for enrichment of adjacent amino acids from the residues the remainder mass is derived from and adjusting the mass accordingly. Because the adjacent residues are pseudo-random in most cases, we also reasoned that any fragment remainder mass less intense than the first corrected mass is likely to be noise. These are also filtered from the result. Additional details about this process can be found in the *Methods* section.

Re-analyzing the cysteine probe data used for illustration above, we identified all eight diagnostic features—five diagnostic ions (Fig. 1a, green), two peptide remainder masses (Fig. 1a, blue), and one fragment remainder mass (Fig. 1a, red)—that were annotated in the prior study and are high confidence identifications. Furthermore, we also identified two additional diagnostic ions, suggesting improved sensitivity for the empirically tuned and automated algorithm (the full attribute list can be found at Supplementary Table 1).

As has been acknowledged elsewhere[20], thorough statistical evaluations of fragmentation pattern detection algorithms are difficult to compute as there are no extant datasets conducive to this analysis. As such, we attempted to evaluate the specificity of the method using a method familiar to proteomics: decoys. We reasoned that datasets with more unique PTMs with uncorrelated diagnostic ions would produce a set of decoy fragmentation patterns that best approximates the feature null distributions. We used a subset of the human proteome published by Wang et al. [25], constructing a shuffled decoy dataset by randomizing the mass shifts of individual PSMs (unshuffled PSM list at Supplementary Data 1a, shuffled PSM list at Supplementary Data 1b). The original and shuffled PSM lists were then searched in parallel. PTM-Shepherd identified 341 diagnostic features from the standard dataset (Supplementary Data 1c), but no features from the shuffled dataset (Supplementary Data 1d), for an $\widehat{FDP}$ of 0. Thus, we demonstrate that our method is robust as well as sensitive.

New protocol characterization

RNA-crosslinking studies also feature labile modifications. Repeating sugar molecules can fragment in myriad ways, frustrating attempts to localize or even identify RNA moieties. Bae et al. recently developed pRBS-ID, an RNA crosslinking workflow utilizing photoactivatable nucleotides and chemical RNA cleavage to overcome these challenges[26]. Alongside the development of their bench technique, they needed to develop a bespoke computational workflow to identify RNA fragment remainder masses and identify and quantify their host peptides. We believed that this process could be recapitulated by PTM-Shepherd without the need for time-intensive custom workflows, and as such we struck a course to replicate their results for the commonly used 4-thiouridine (4SU) nucleotide analog[27].

First, we performed an open search using the default diagnostic ion mining setting available in FragPipe. As expected in any open search, PTM-Shepherd identified many mass shifts for biological and chemical PTMs, but also two unannotated mass shifts of 226 Da and 94 Da at high amounts likely corresponding to the modification of interest (Table 1, full mass shift profile can be found at Supplementary Data 2a). These mass shifts matched those identified by Bae et al. Notably, the fragment remainder masses PTM-Shepherd identified for both mass shifts were nearly identical (Table 2), indicating with a high degree of likelihood that they had the same source. In this case, fragment remainder masses of 94 Da were identified from both mass shifts' b- and y-ion series, and an additional fragment remainder mass of 77 Da (the prior remainder with a loss of ammonia) was identified from both mass shifts' b-ions (Supplementary Data 2b). This remainder mass appeared to be diagnostic for RNA-crosslinked peptides, which we further confirmed by performing an additional search incorporating this new fragment remainder mass alongside the 94 Da mass into a labile mode search. Including this feature resulted in an additional 3.8% RNA-crosslinked PSMs over the 94 Da mass alone (Supplementary Fig. 1).

After a more targeted search using these fragment remainder masses, we also wondered whether any additional diagnostic features might appear for the RNA-crosslinked peptides and performed a second pass at diagnostic feature mining (Supplementary Table 2). For fragment remainder masses, we recovered the two masses described above from both b- and y-ions from the intact 226 Da mass shift corresponding to the loss of five-member sugar ring (Fig. 2a) as well as an a-ion associated mass shift at 66 Da (94 Da minus 28 Da) from b-ions. Diagnostic ions can be of particular interest for future analyses, such as in ion-triggered instrument routines, even if they are left unused at the present. We found two easily explicable diagnostic ions for the intact nucleoside (Fig. 2b): an ion at 133 m/z corresponding to a dissociated ribose, the other half of the 94 Da fragment remainder mass, and an associated neutral loss of water. The partial modification loss on fragment ions was also observed from the 94 Da mass shift (Fig. 2c). However, the diagnostic ions were not diagnostic for the MS1 mass shift corresponding the nucleoside without the ribose (Fig. 2d), as with the ribose already dissociated there is nothing left to form the diagnostic ion.

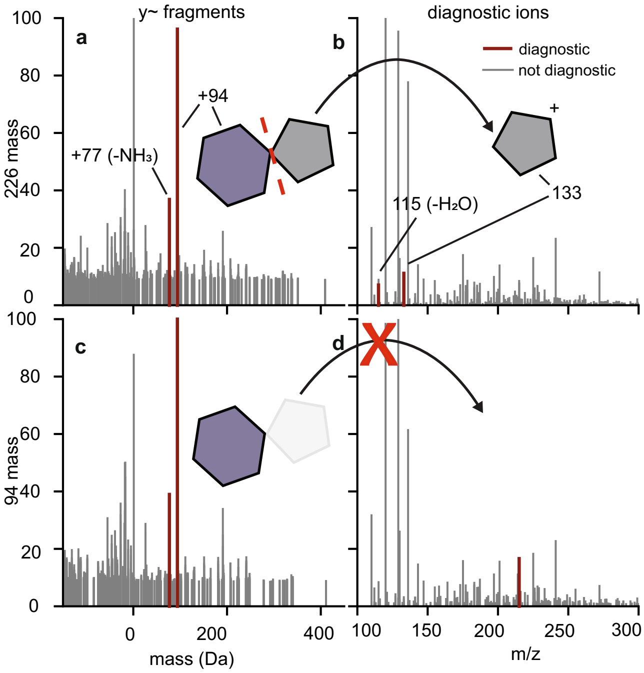

Fig. 2 | Characterization of 4SU fragmentation from a pRBS-ID experiment. **a** Fragment remainder masses for y-ions from the 226 Da mass shift. Remainder masses that passed PTM-Shepherd's filtering correspond to the retention of the 94 Da fragment on the peptide. **b** Diagnostic ions derived from the fragmentation of the nucleoside analog from the 226 Da mass shift. **c** Fragment remainder masses for y-ions from the 94 Da mass shift. **d** Diagnostic ions from the 94 Da mass shift. Source data are provided as a Source Data file.

Data-driven discovery of diagnostic features

Glycopeptides contain labile modifications that produce rich sequences of diagnostic ions and peptide and fragment remainder masses[28]. We reasoned that detecting known glycopeptide fragmentation patterns would be a good way to validate our algorithm's performance given the extensive literature characterizing glycopeptide fragmentation. To this end, we searched for glycopeptides in a large IMAC-enriched, TMT-labeled clear cell Renal Cell Carcinoma (CCRCC) dataset[29]. Phosphorylation enrichment by IMAC, the method employed in this publication, has been shown to simultaneously enrich glycopeptides, particularly those bearing sialic acids[30,31], so the data should be rich in glycan signals. This dataset also presents two challenges: TMT-labeling is known to affect PTM fragmentation patterns due to reduced proton mobility[9] and the relatively high collision energies used in this experiment cause extensive fragmentation of glycans, reducing the signal strength of typical glycan fragment ions. Because TMT is searched as fixed or variable modifications (i.e., it does not produce an open search-derived mass shift), peptides containing only TMT and no other mass shift are considered “unmodified”. Effectively, this controls for TMT-related fragmentation when determining glycopeptide fragmentation patterns, as TMT-related fragmentation patterns are present in both the glycopeptide spectra and the unmodified spectra.

We first wanted to verify that we could detect the commonly used glycopeptide-associated diagnostic ions from the MSFragger-Glyco[7] search and annotation that are most likely to be present[32]. After discarding any mass shifts less than 50 Da we were left with 493 likely glycan mass shifts from 967,264 glyco PSMs of 9623 unique glyco peptides, each of which should be enriched for diagnostic ions

Table 2 | Diagnostic features for the most abundant mass shifts detected in an open search of a pRBS-ID experiment

peak apex	ion type	mass	remainder propensity	delta mod. mass	percent PSMs (mod)	percent PSMs (unmod)	avg. intensity (mod)	avg. intensity (unmod)	intensity fold change
226.0594	diagnostic	241.0626			37.76	26.3	17.55	8.35	3.02
226.0594	diagnostic	133.0500			27.28	15.1	7.89	4.44	3.21
226.0594	b	94.0344	26.86	−132.025	29.96	0.8	46.16	36.40	47.49
226.0594	b	77.0092	15.15	−149.050	15.96	1.4	24.54	17.53	15.96
226.0594	y	94.0362	14.48	−132.023	27.53	0.9	43.49	32.78	40.58
57.0216	b	57.0216	46.25	0	22.07	1.7	52.66	12.89	53.05
57.0216	y	57.0216	23.03	0	20.16	1.6	53.24	13.47	49.82
94.0168	diagnostic	164.0832			25.70	10.0	10.86	5.20	5.37
94.0168	diagnostic	215.0582			32.53	15.3	12.71	6.64	4.07
94.0168	diagnostic	241.0626			44.98	26.3	15.37	8.35	3.15
94.0168	b	94.0168	32.01	0	37.35	0.9	46.32	42.79	44.93
94.0168	b	−19.0638	19.08	−113.081	16.47	0.4	36.95	18.82	80.82
94.0168	b	77.0042	18.13	−17.0126	18.07	1.7	37.81	13.33	30.15
94.0168	y	94.0168	26.76	0	34.54	0.9	44.15	42.79	39.60
−48.0032	diagnostic	181.0986			27.63	16.3	9.95	4.63	3.64
−48.0032	b	−48.0032	41.80	0	42.8	1.3	77.96	15.23	100.00
−48.0032	y	−48.0032	24.37	0	41.25	1.3	64.48	20.08	100.00

Remainder propensity scores are present only for b- and y- remainder masses. Features corresponding to the set of modified or unmodified PSMs used in comparisons are labeled as (mod) and (unmod), respectively. The difference between the MS1 mass shift and the observed fragment remainder masses (i.e., the lost mass) is enumerated in the “delta mod mass” column.

associated with the N-glycan core structure[12] and other monosaccharide(s) present, including sialic acid. Indeed, PTM-Shepherd successfully identifies many of the expected diagnostic ions used in glycopeptide searches and glycan identification[7,32], including three known sialic-acid related oxonium ions at 274, 292, and 657 m/z (Fig. 3a, Supplementary Data 3). In addition to these, we found 12 additional ions that were diagnostic for more than 50% of glycan mass shifts. We hypothesized that these might be diagnostic ions specific to a high-collision energy environment and attempted to identify them in a data-driven manner. We used PTM-Shepherd's diagnostic feature extraction module, which extracts intensities for user-specified ions of interest, to quantify these alongside the set of common diagnostic ions used in the MSFragger-Glyco, identifying clusters of highly correlated ions (Fig. 3b, see Methods). Known ions clustered together meaningfully, with annotated GalNAc, Hexose, HexNAc, and Phospho-Hexose ions being highly correlated with others from the same residue, lending credence to this approach's validity. Perhaps unsurprisingly given the nature of the enrichment method, most unannotated diagnostic ions formed a large cluster with the two monomeric sialic acid oxonium ions found at 274 and 292 m/z. We selected the diagnostic ions from a subcluster (Fig. 3b, cluster 5) that was highly correlated with both oxonium ions (Supplementary Data 4) to validate individually. These ions formed a potential neutral loss series from the annotated 292 and 274 m/z oxonium ions, with successive losses of 42, 17, 18, and 30 Da. To our knowledge, manuscripts covering sialic acid fragmentation make no mention of these as diagnostic ions[12,33,34], so their presence in spectra acquired at high collision energies may be of interest to other researchers when assigning sialic acids to glycan composition.

Aside from diagnostic ions, glycopeptides also produce an intense series of peptide remainder ions, called Y-ions in glycopeptide fragmentation nomenclature, where the peptide is intact while the modification has fragmented[12]. Mammalian N-glycans have a common core structure. When the core structure fragments, it produces a pattern of Y-ions with peptide remainder masses that are identical irrespective of the peptide's or glycan's mass and can even be used to diagnose the presence of glycopeptides[6]. Like the diagnostic ions discussed above, we find an expected pattern of peptide reminder masses corresponding to the N-glycan's core's Y-ion series (Fig. 3c). Aside from these, two peptide remainder masses that are not considered in the MSFragger-Glyco search recurred across mass shifts: +83 Da and −17 Da. The smallest glycan mass from the N-glycan core, corresponding to a single GlcNAc retained on the peptide, is 203 Da, so seeing masses smaller than that being as diagnostic for glycopeptides as the complete loss of glycan (+ 0 Da) or a single GlcNAc (+ 203 Da) was unexpected. This pattern—consisting of a cross-ring fragmentation event at the core GlcNAc and a loss of ammonia, respectively—has previously been identified as a conserved fragmentation pattern for glycopeptides[35], but appears not to be used in current state-of-the-art tools[6,7,36]. This indicates that even for very well characterized

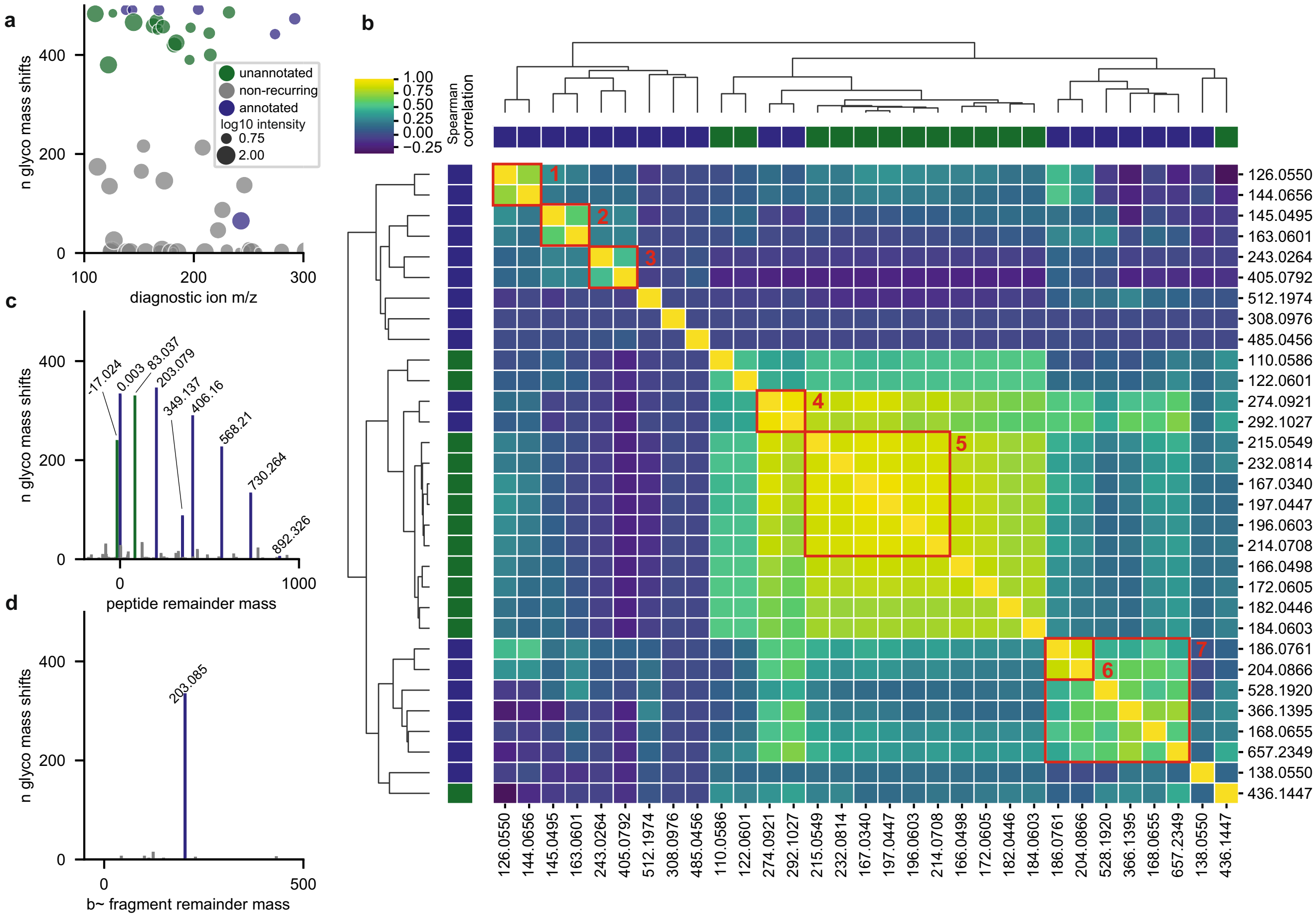

Fig. 3 | Diagnostic features of IMAC-enriched glycopeptides under high energy conditions. a Scatterplot of recovered diagnostic ions across glyco mass shifts. Ions occurring in >50% of mass shifts are considered recurring and are included in **b**. **b** Spearman correlation clustering between diagnostic ions across all glyco spectra. Identifiable clusters are as follows: 1: GalNAc, 2: Hex, 3: Phospho-Hexose, 4: NeuAc, 5: sialic acid, 6: HexNAc monomers; 7: HexNAc including non-monomers. **c** Histogram of peptide remainder ions across glyco mass shifts. **d.** Histogram of fragment remainder ions across glyco mass shifts. Source data are provided as a Source Data file.

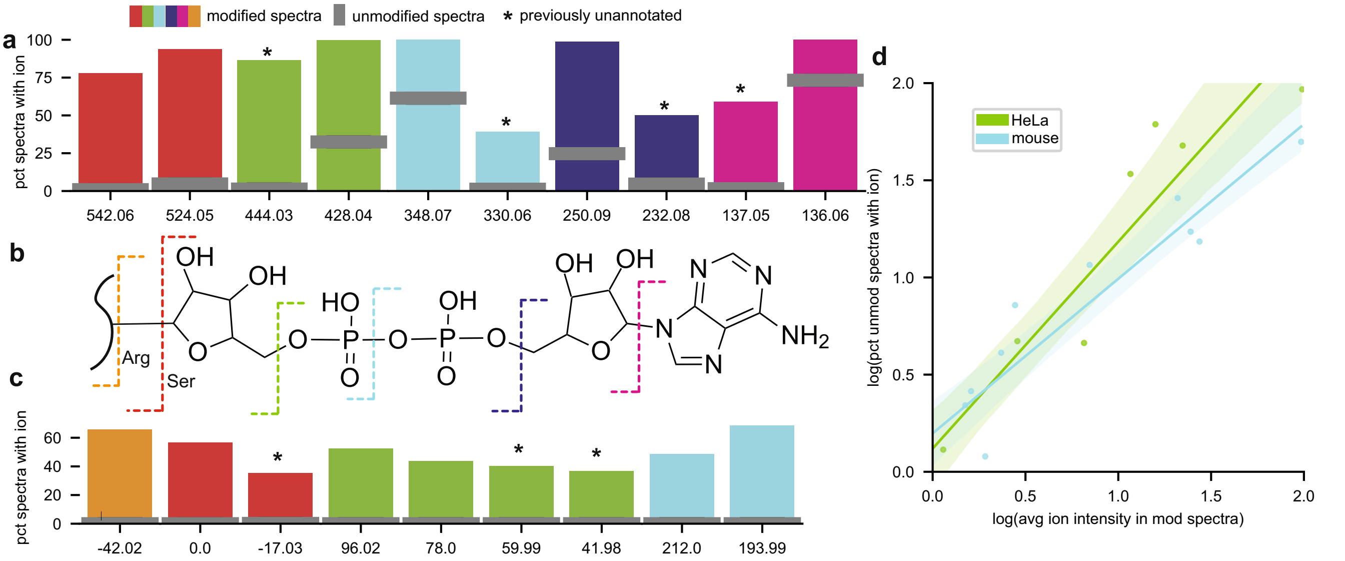

Fig. 4 | Analysis of ADP-ribosylation fragmentation patterns. a ADP-ribosylation diagnostic ion presence in modified and unmodified spectra. Features detected in both datasets were averaged across all values. **b** Structure of ADP-ribosylation. Dashed lines correspond to breakpoints in the molecule. **c** ADP-ribosylation peptide remainder mass presence in modified and unmodified spectra. **d** Correlation between average intensity of diagnostic ions and their presence in unmodified spectra across both datasets. 95% confidence intervals for regression estimates are highlighted. Source data are provided as a Source Data file.

modifications, gaps can exist between knowledge of fragmentation patterns and their use in computational tools, a disconnect that PTM-Shepherd's automated fragmentation analysis can correct.

The final diagnostic feature we assessed for this glycan dataset is shifted fragment ion series. When the peptide and glycan have both fragmented, the glycan can leave a signature +203 fragment remainder mass on the peptide ion series[12]. PTM-Shepherd recovered this fragment remainder mass exactly (Supplementary Fig. 2) and with little interference from artefactual mass shifts despite the noisy nature of pairwise ion differences.

Some of the identified ions, particularly the Y-ion series of peptide remainder masses, appeared to taper off very quickly at larger masses, which is a known issue when identifying labile modifications at relatively high collision energies. We reasoned that using these extra ions in our search when they can be low-abundance or absent injects additional noise into the search results and suppresses real glycopeptide identifications. To test this, we used the fragmentation information provided by PTM-Shepherd and reduced our fragment and peptide remainder masses to only the four Y-ions appearing in >50% of glycan mass shifts. Though more careful analysis would surely yield better results, even the incorporation of a crude cutoff from a subset of the data resulted in a 4.5% increase in glyco-PSMs, proving that the fragmentation information provided by PTM-Shepherd enables researchers to tune search parameters to best suit their individual experiments.

We showed that PTM-Shepherd was sensitive to known diagnostic features for glycopeptides. New features detected by PTM-Shepherd also had chemical meaning relevant to the experimental setting, and PTM-Shepherd was able to identify unannotated sialic acid diagnostic ions for high-energy TMT experiments in a data-driven manner. Additionally, we proved that the information provided by PTM-Shepherd can be incorporated into subsequent searches to fine-tune parameters for different experimental settings.

Automated fragmentation analysis

ADP-ribosylation (ADPR) has seen a surge of interest in recent years, with many enrichment methods[37,38] and instrumental techniques[39] developed over the last decade to aid in its study. Despite this, specialized computation techniques have lagged behind. Fragmentation studies—necessary to design tools or workflows for the analysis of PTMs—require careful analysis and examination of individual spectra[40]. We believed that PTM-Shepherd's diagnostic feature mining module could expedite fragmentation studies and reveal useful insights to their behavior. To demonstrate this, we reanalyzed ADPR-enriched data from Martello et al. [39]. from peroxide-treated HeLa cells, rich in Ser-directed ADPR, and mouse liver, rich in Arg-directed ADPR.

To validate the fragmentation patterns we detected, we first cross-checked them against published ones[40]. As expected, we found previously annotated diagnostic ions (Fig. 4a, Supplementary Data 5a,b) corresponding to almost every expected breakpoint on the ADPR side chain (Fig. 4b). These were all found at relatively high levels among ADPRylated spectra (78–100%). Interestingly, the most intense of these ions—e.g., the adenine-derived ion at 136—was also abundant in unmodified spectra (73%) as a result of co-fragmentation of ADPR-containing peptides (Supplementary Note 1). This speaks to the robustness of PTM-Shepherd's algorithm; even features whose presence alone is not specific to a particular mass shift can be recovered because our scoring and filtering utilizes intensity information. We also recovered additional ions that correspond to derivatives of annotated ions: an oxidized 428 m/z ion (+16 Da), a 348 m/z ion that has undergone a loss of water (−18 Da), and a 250 m/z ion that has undergone a loss of water (−18 Da). These ions were all far more specific to the ADPR PTM than their annotated counterparts and thus may be of interest to others studying ADPR. A final diagnostic ion of interest did not correspond to a common mass offset from an annotated ion. At 137.0458 m/z, we could not identify this ion as being a secondary product of any annotated ions. Its exact mass is strongly suggestive of a deamidation event occurring on the adenine ion at 136.0618 m/z (+0.9840 Da).

We also observed a strikingly strong relationship between an ion's average intensity and its presence in unmodified spectra across both ADPR datasets analyzed (Fig. 4d, Spearman's R:^{2} mouse = 0.857; HeLa = 0.884). We have previously commented on this phenomenon when looking at biotin-derived Cysteine probes[19]. In that case, reducing the isolation window and employing ion mobility gave a modest boost to diagnostic ion specificity, an effect that was presumed to be caused by reduced co-fragmentation of peptides. Additionally, we see a strong relationship between the purity of a given unmodified spectrum and

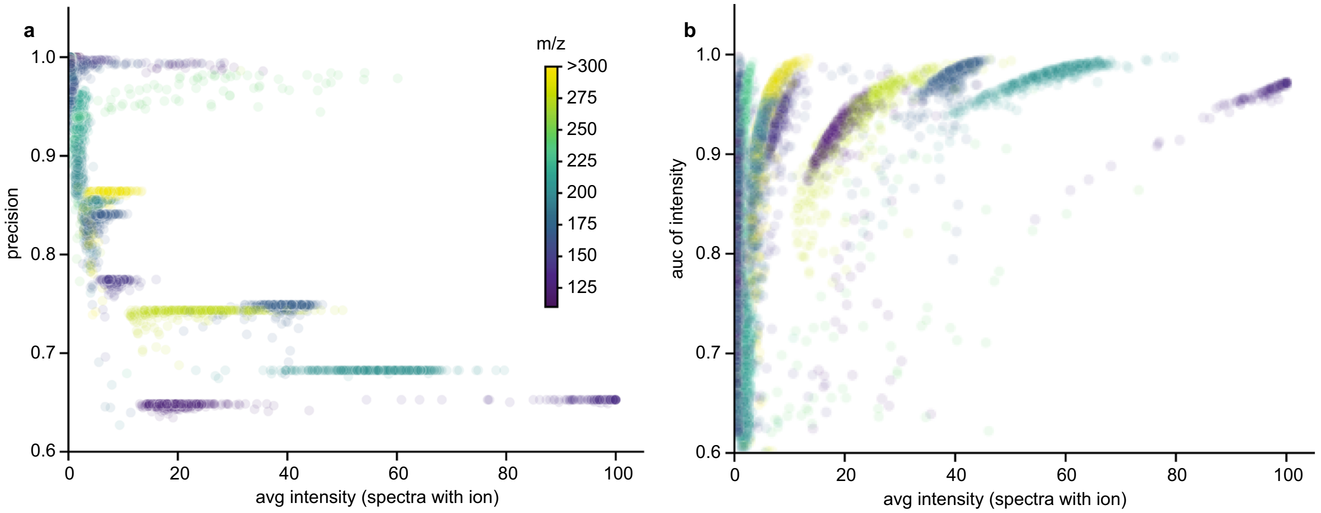

Fig. 5 | Trends in diagnostic remainder ions. a Relationship between diagnostic ions' observed intensity and the precision of their presence. **b** Relationship between diagnostic ions' observed intensity and the classification strength of their intensity as measured by AUC. Source data are provided as a Source Data file.

the intensity of PTM-specific diagnostic ions in this dataset (Supplementary Fig. 3), lending further credence to the co-fragmentation hypothesis. It is worth noting that the issue of co-fragmented ions has been well-studied in the context of isobaric tandem mass tags[41]. But, to our knowledge, this is an understudied issue for biological PTMs.

PTM-Shepherd also identified both types of remainder ions in this dataset, peptide (Fig. 4c) and fragment. Of note was PTM-Shepherd's recovery of a −42 Da peptide remainder mass from the Arg-directed ADPR dataset (Supplementary Data 5b). When Arg-linked ADPR dissociates from the peptide, it appears to frequently take a portion of the Arg side chain with it. The result is a negative peptide remainder mass corresponding to the loss of the Arg reactive group that is both prevalent (66% of PSMs) and distinguishes ADPR on Arg from other residues. This is also reflected in the fragment remainder masses. The b- and y-ion series were found to consist of 40 and 26% ions shifted by −42, respectively (Supplementary Data 5b). Since only ions downstream of the modification site are expected to be shifted, we only expect to find half of all ions containing PTM-related mass shifts. The abundance of the Arg-specific fragment ions indicates that the modification itself should be easily localizable. We also found a noteworthy number of neutral loss-associated peptide remainder ions. When ADPR fragments after the primary ribose (Fig. 4b, green), we would expect a peptide remainder mass of 114 Da if it were to remain intact. We do not find that mass, but instead find four sequential neutral losses of water from that mass. Equally of interest is that the neutral loss peaks—despite neutral losses not being unique to ADPRylated peptides—are not found in unmodified spectra. Though counterintuitive, even common losses can produce PTM-specific peaks. By thinking of them as losses of almost the entire modification and a common neutral loss, it is easier to reconcile their uniqueness to specific modifications. In other words, a −17 peptide remainder mass (Fig. 4c, red) will appear at the precursor m/z − 17 for unmodified peptides, but at precursor m/z − 558 for modified peptides.

Use cases and applicability of diagnostic features

To investigate the extent to which co-fragmentation affects diagnostic feature characteristics, we leveraged our ability to identify large numbers of glycan diagnostic ions from the CPTAC IMAC-enriched dataset. This dataset represents 117 unique diagnostic ions, each found to be diagnostic for between 1 and 493 mass shifts, for a total of 13707 data points (Supplementary Data 1). Every diagnostic ion was evaluated individually for its ability to separate glyco and unmodified spectra based on its precision and AUC (Supplementary Note 2). This was repeated for the 64 unique peptide remainder masses observed between 1 and 344 times, totaling 2261 data points.

Here, precision can be interpreted as the probability *y* that a spectrum is a glyco spectrum given that the diagnostic ion is present in the spectrum at intensity *x* (Fig. 5a). Diagnostic ion precision attenuates rapidly as the intensity increases, losing more than a third of its usefulness when it becomes the spectral base peak (average intensity 100.0). In enriched datasets such as this, that can be explained by co-fragmentation. As mentioned above, for co-fragmented spectra the probability of seeing ions from the minor product is inversely proportional to the spectral purity. For all ions in the spectrum, their probability of passing the limit of detection is proportional to their intensity. Thus, if you have a very intense ion, it can appear in a spectrum even if is coming from a minor product in a relatively high purity spectrum. In enriched datasets such as this, co-fragmented minor peptides are disproportionately likely to be PTM-bearing and produce diagnostic ions. If those ions are also very intense by nature, they will be present in many unmodified spectra. This is a trend that can be reversed by taking intensity information into account rather than only checking for the presence or absence of the ion (Fig. 5b). The AUC statistic here can be directly interpreted as the probability *y* that a diagnostic ion of intensity *x* drawn from a random modified PSM will be greater than the intensity of the same diagnostic ion drawn from a random unmodified PSM. After including intensity information, an ion's ability to separate glyco and non-glyco spectra increases with intensity. Incorporating this feature into PTM-Shepherd allows us to detect diagnostic ions that are as ubiquitous as ADPR's adenine ion, in 92.9% of off-target spectra in the HeLa dataset (Supplementary Data 5a). It also shows that researchers can effectively use intense diagnostic ions for scoring PTMs, but only if they empirically learn the distribution of intensities among unmodified PSMs beforehand.

For peptide remainder masses, unlike diagnostic ions, precision does not attenuate with intensity (Supplementary Fig. 4). As mentioned above, peptide remainder masses are mass- (although not sequence) specific. Co-fragmented peptides can only share peptide remainder masses if they share a mass that is indistinguishable at MS/MS mass accuracy, which is not guaranteed even for co-fragmented peptides with the same charge state. Excluding noise peaks that happen to fall within the tolerance of a theoretical peptide remainder ion, there should be few erroneously matched peptide remainder masses. The result is a very specific feature that does not attenuate as it gets more intense. Accordingly, peptide remainder ions discovered by PTM-Shepherd have many applications. Experiments performed with data-independent acquisition (DIA) have many co-fragmented peptides by design and present a prime opportunity for their use. Plus, with the advent of real-time searching, peptide remainder ions can also be used for instrumental enrichment[42].

Discussion

Our analyses show that PTM-Shepherd can be used to reliably identify diagnostic features for any modification of interest. In high-energy glycopeptide fragmentation, we showed that diagnostic ions for sialic acid could be identified without prior knowledge in a data-driven way, as well as finding two peptide remainder masses that had been described by experimentalists but neglected by cutting-edge glycopeptide search tools. In our discussion of a novel RNA-crosslinking workflow, we showed that we can easily automate experimental characterization in the FragPipe/PTM-Shepherd environment. Finally, our discussion of ADPR fragmentation demonstrated that fragmentation studies—traditionally done by hand with manual annotation of spectra or using custom tools[10] and synthetic peptides[14]—could be automated and democratized to reach a broader audience and study PTMs without additional benchwork. We even found meaningful fragmentation patterns that would have been missed by annotation focused on modification structure alone. Although our analysis focused on demonstrating PTM-Shepherd's capabilities, we also used our ability to generate diagnostic features in large numbers to better understand their nature. We showed that co-fragmentation of peptides presents a major issue for the precision of diagnostic ions in PTM analysis and explored ways to overcome it, as well as interrogating the utility of peptide and fragment remainder masses.

Automated diagnostic feature detection has wide-ranging applications across proteomics fields. Chemical probes can be characterized instantly, facilitating their development[19]. It could be advantageous to use these features to develop custom modification scores for localization-by-proxy strategies[43] or to perform rescoring in Percolator[44]. Furthermore, for enriched datasets or DIA-studies, the remainder masses identified by PTM-Shepherd might be the only reliable way to definitively identify labile modifications. There are myriad ways in which understanding modification behavior aids researchers, and thus we believe that the diagnostic feature detection enabled by PTM-Shepherd will be an invaluable tool in the analysis of proteomics data.

Methods

Algorithm overview

Here we provide an overview of the algorithm used for diagnostic feature detection. Additional details are provided in the sections below. The goal for this algorithm is to identify spectral features for the PTMs identified in an analysis. We used open (or mass offset) searches to identify peptides bearing PTMs. PSMs are grouped based on their mass shift, using PTM-Shepherd's peakpicking algorithm[23], into a "mass shift bin." To do this, MS1 mass shifts for the experiment are allocated into a histogram with 0.0002 Da width bins. The histogram is then smoothed by distributing each bin's mass into itself and the two adjacent bins on either side using a normal distribution, and peak apexes are detected by identifying regions of the histogram with a prominence greater than 0.3 and the 500 highest signal-to-noise peaks (calculated by summing PSM counts within a 0.004 Da window of the apex and subtracting PSM counts for 0.01 Da windows on either side) are reported. Redundant PSMs are not considered for this analysis, with the highest E-Value PSM for each peptide ion within a mass shift bin being retained as a representative PSM. For each representative PSM within a mass shift bin, we calculate three features: potential diagnostic ions in the spectrum, potential peptide remainder masses in the spectrum, and potential fragment remainder masses in the spectrum (see: Spectral feature calculation). We then check to see which of these features recurs across representative PSMs within the mass shift bin (see: Identifying recurring features). Ideally, one could use these patterns to describe how the PTM in the mass shift bin fragments. However, detecting a fragmentation pattern this way does not mean that it necessarily describes the PTM within the mass shift bin. For example, immonium ions from individual unmodified amino acids will be detected as recurring features but are related to peptide fragmentation rather than PTM fragmentation. Similarly, a-ions will be detected as fragment remainder masses because they produce consistent mass shifts from b-ions. To limit the list of recovered features to only those that describe the relevant PTM, we thus need to see whether they are more abundant within the mass shift bin than in some background dataset. We reasoned that unmodified peptides (contained within the zero bin) are the best representative of a dataset's background. Following this logic, we then test whether each recurring feature from the mass shift bin is more abundant among representative PSMs in the mass shift bin than the representative PSMs of unmodified peptides (see: Identifying significant features). Features that pass statistical and abundance filtering are reported.

Spectral feature calculation

Rather than using every PSM for what is inherently a noisy process, we select only those that are most likely to have the highest quality spectra. To do this, PSMs within each mass shift bin are first grouped by their peptide ion (sequence, modification state, and precursor charge state), then each group of PSMs has its lowest E-value representative selected for all downstream processing.

The first MS/MS spectral feature we analyze is raw spectral ions, such as immonium and oxonium ions, which we will refer to simply as diagnostic ions. All spectra from PSMs containing a given delta mass are stripped of matched a-ions, b-ions, and y-ions (by default). Spectra are also stripped of a-, b-, and y-ions that are found to be shifted by the PSM's delta mass, preventing backbone fragments containing the modification from being counted as diagnostic ions. At this point, a spectrum can be thought of as a vector composed of m ions, where each ion$_i$ has a corresponding mz$_i$ and int$_i$ corresponding to the ion's mass at charge state one and its intensity. All remaining ions are considered potential diagnostic ions and stored in a vector $\mathbf{U}$ of length m. This can be represented as

$$\mathbf{U} = [(\mathrm{mz}_1, \mathrm{int}_1), \cdots, (\mathrm{mz}_m, \mathrm{int}_m)] \tag{1}$$

The second MS/MS spectral feature we analyze in the MS/MS spectra is peptide remainder masses. All spectra from PSMs containing a given delta mass are stripped of shifted and unshifted a-, b-, and y-ions, as described above, before precursor remainder mass calculation. A theoretical peptide mass P of charge state one is calculated based on the peptide sequence and variable modifications identified for the PSM during spectral searching but excluding any MS1 mass shift. Then, the pairwise distance d between each remaining ion in the MS/MS spectrum and the theoretical peptide mass P is calculated and stored in a vector $\mathbf{V}$ of length m, where m is the number of ions remaining in the spectrum after filtering. Each component $\mathbf{V}_i$ contains the pairwise distance between P and mz$_i$ as well as the intensity int$_i$. This can be represented as

$$\mathbf{V} = [\mathbf{V}_1, \cdots, \mathbf{V}_k] \tag{2}$$

with each component

$$\mathbf{V}_i = (\mathrm{mz}_i - P, \mathrm{int}_i) \tag{3}$$

Intuitively, each component can be interpreted as what the precursor remainder mass and intensity would be if the ith ion were a shifted precursor in the spectrum.

The third MS/MS spectral feature we analyze is fragment remainder masses. All spectra from PSMs containing a given delta mass are stripped of unshifted a-, b-, and y-ions only, allowing us to identify instances where the entire delta mass remains on the fragment ions. We reasoned that understanding how modifications affect individual ion series would provide insight into fragmentation patterns, so fragment remainder masses for b- and y-ions are calculated independently. Our procedure is similar to the procedure described by Dancik et al. [45]. and reiterated by He et al. [20]. For each fragment ion series, the peptide's theoretical fragment ions of charge state one are calculated based on the peptide sequence and modifications identified for the PSM during spectral searching; the vector $\mathbf{F}$ holds each of n theoretical fragment ions, where n is the length of the peptide minus one and $\mathbf{F}_j$ corresponds to the jth fragment ion. Then, the pairwise distance between each remaining ion in the MS/MS spectrum and each theoretical fragment ion $\mathbf{F}_j$ is calculated and stored in a matrix $\mathbf{W}$ of size m by n, where m is the number of ions remaining in the spectrum. This can be represented as

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{m1} & \cdots & \mathbf{W}_{mn} \end{bmatrix} \quad (4)$$

with each component

$$\mathbf{W}_{ij} = (\mathrm{mz}_i - \mathbf{F}_j, int_i) \quad (5)$$

Intuitively, each matrix component can be interpreted as what the jth fragment's remainder mass and intensity would be if the ith ion in the spectrum were the jth theoretical fragment's shifted counterpart.

Identifying recurring features

We then determine which features represent the most intensity and are thus worthy of undergoing testing for enrichment. To do this, we place every value in a histogram with a bin width of 0.2 mDa spanning the range of possible features. Each feature (diagnostic ion, peptide remainder mass, and fragment remainder masses), is given its own histogram. For peptide and fragment remainder masses, the left tail of the histogram is truncated at −250 Da because values smaller than that would necessitate the losses of multiple residues. Because peaks in these histograms are generally too jagged to cleanly identify peaks, they are smoothed. This is done by calculating a rolling mean across the histogram, but one that increases in spread for heavier ions to account for their larger uncertainty in Da terms. The spread of the peak in Da is determined by

$$\mathrm{tol}^* \frac{(\mathrm{mass}_i + \mathrm{mass}_x)}{1000000} \quad (6)$$

where the MS/MS spectrum tolerance is tol in ppm terms, the mass of the ion being inserted is mass_i, and the average mass of fragment ions or peptide ions from the mass bin is mass_x. The mass_x term allows us to properly account for the uncertainty of remainder masses. Distributing the mass without considering the peptide it came from would lead to improperly small spreads. To illustrate the issue, a peptide remainder mass of 100 Da derived from a 1000 Da peptide should have an MS/MS error calculated from 1100 Da, not 100 Da. For diagnostic ions, a mass of 150 Da is used as mass_x.

Peaks in the histogram are defined by descending each side of a local maximum bin until a bin with either zero intensity is reached or the next bin's value increases. Manual calibration found that a bin-to-bin tolerance of 1% was enough to prevent noisy bins from splitting peaks in two. Peaks are then integrated by summing histogram bins within the MS/MS tolerance without regard for adjacent peak boundaries. Any peak with an integrated area greater than 0.1% (by default), representing an average intensity greater than 0.1% of the base peak, is selected for downstream analysis. A final check is performed to remove redundant peaks where the least intense of any two histogram peaks that cannot be resolved under the provided MS/MS tolerance is removed.

Identifying significant features

To find features specific to a particular mass shift, the full feature set—every major peak from the feature histograms above—needs to have features pruned from it that are not specific to the mass shift. We reasoned that peptides without mass shifts would be a good representative of a dataset's noise, and as such testing whether features are more likely to appear among peptides with a particular mass shift than those without any mass shift would filter out non-modification-specific features.

Representative PSMs for every peptide ion with a particular mass shift and representative PSMs with no mass shift are first assembled, then every feature from the list of diagnostic ions, precursor remainder masses, and fragment remainder masses is quantified for each PSM in both lists. For spectra that do not have the diagnostic feature, the intensity is coded as a zero. Fragment remainder masses are likely to appear by chance solely based on the number of theoretical-to-experimental ion offsets calculated, so PSMs are considered to be missing a fragment remainder mass if there are fewer than two shifted ions of the ion type in the spectrum, i.e., fewer than two matching fragment remainder masses within feature matrix $\mathbf{W}$ for any ion type. A maximum of 1000 representative PSMs are selected per group, with a seed provided internally for reproducibility.

For every diagnostic feature tested, a series of metrics are produced for filtering noise peaks from real peaks. First, the lists of feature intensities from the unmodified and mass shifted PSMs are compared via a two-sided Mann-Whitney-U test with tie and continuity correction (adapted from the Hipparchus statistics library for Java, v1.8). E-values for each diagnostic feature are calculated by multiplying by the number of tests performed within the feature class for the current mass shift. By default, any feature with an E-value less than 0.05 is filtered out. A second metric to quantitatively assess the strength of the feature is included in PTM-Shepherd's output: Area Under the Curve (AUC). This is commonly used as a measure of effect size for the Mann-Whitney U test and can be directly interpreted as the probability that a mass shifted PSM will have a higher intensity for this feature than an unmodified PSM. Second, we calculate a feature's fold change of average intensity across all PSMs. Any features with fold change of less than 3.0 is filtered out by default. This metric primarily helps to identify diagnostic ions and non-specific but increased neutral losses for peptide and fragment remainders. Third, we filter out any features that are not sensitive for the modification, occurring in less than 25% of representative PSMs for diagnostic ions and peptide remainder masses. Owing to the multiple ion requirement for fragment remainder masses, this filter is reduced to 15% but is accompanied by an ion propensity filter requiring at least 12.5% of the identified ions within that series having the mass shift.

Fragment ions undergo an additional post-filtering processing step. Because a theoretical-experimental peak offset $\mathbf{W}_{ij}$ is created for n theoretical ions in the theoretical ion series, a single peak in the experimental MS/MS spectrum produces a sequence specific pattern. For example, if the jth residue produces an offset with fragment $\mathbf{F}_j$ from the experimental ion i, the same experimental ion responsible for that offset will also produce an "echo" offset from fragment $\mathbf{F}_{j-1}$ equal to the original offset plus the mass of the residue at position j. Similarly, it will produce an "echo" offset from fragment $\mathbf{F}_{j+1}$ equal to the mass of residue $j+1$ minus the original offset. Depending on the fragment ions containing the mass shift, some modifications can produce very weak signals for their primary mass shift but strong signals from shifted fragment ions upstream or downstream of the modification site. To correct for this, we check for residue enrichment both on and adjacent to the peptide site responsible for producing the mass shift. If any residue is found at position *j + 1* for a modification more than 50% of the time, the fragment remainder mass is adjusted by subtracting that residue's weight from the fragment remainder mass. If any residue is found at position *j* more than 50% of the time, the mass of residue *j* is added to the fragment remainder mass. With all fragments downstream of a peptide's modification site carrying the mass shift, the residues responsible for these shifts should be roughly uniformly distributed across all 20 amino acids. Thus, any mass shift that is less prevalent than one of these adjusted offsets is unlikely to be a real peak, and reporting for fragment remainder masses is truncated after the first adjustment.

Data processing

Four datasets were used throughout this manuscript. The first consists of a single Thermo Fisher Raw file "2021-2-23_EA_296_1A_Final.raw" from ProteomeXchange repository the PXD028853. This contains a cysteine chemoproteomic probe from Yan et al. (2022)[19] that was used to demonstrate the algorithm. Data was collected on a Thermo Scientific Orbitrap Eclipse Tribrid in DDA mode and processed directly in FragPipe (v18.0) without conversion to mzML. Data was searched against the Uniprot reviewed protein sequences database retrieved on 13 June 2021 with decoys and common contaminants appended. An offset search was performed in MSFragger (v3.5)[5] by loading the "Mass-Offset-Common-PTMs" workflow, replacing the offset list with 0 and 463.236554, and replacing the fixed cysteine carbamidomethylation with a variable one. "Write calibrated MGF"[24] was turned on for the PTM-Shepherd[23] diagnostic feature mining module, and "Diagnostic Feature Discovery" in PTM-Shepherd (v2.0.0) was enabled with default parameters. Filtering to 1% PSM, peptide, and protein levels was performed by Philosopher (v4.2.2)[46].

The second dataset consists of the Clinical Proteomics Tumor Analysis Consortium (CPTAC) IMAC-enriched clear cell renal cell carcinoma (ccRCC) samples[29] from the CPTAC data portal[47]. These 299 files represent TMT-labeled solid tumor or adjacent normal tissue from 110 human ccRCC patients. Samples were acquired on a Thermo Fisher Fusion Lumos in data-dependent acquisition (DDA) mode using high-collision dissociation (HCD). Thermo Fisher raw files were converted to mzML format using Proteowizard v3.0.11392[48] with vendor peakpicking enabled. The 23 TMT-plexes were separated into separate experiment folders and processed using FragPipe v18.0. For the primary analysis, the default "glyco-N-TMT" workflow was used with minor changes to account for the goals of the analysis and experimental setup. Data was searched against the Uniprot reviewed protein sequences database retrieved on 13 June 2021 with decoys and common contaminants appended. During the MSFragger[5] search, two variable phosphorylation modifications were allowed on the residues STY due to the expected enrichment of phosphorylated peptides and "Write calibrated MGF"[24] was turned on for the PTM-Shepherd[23] diagnostic feature mining module. In PTM-Shepherd, "Assign Glycans with FDR" was disabled, and "Diagnostic Feature Discovery" was enabled with default parameters. Finally, "Isobaric Labeling-Based Quantification" with TMT-Integrator was disabled. Filtering to 1% PSM, peptide, and protein levels was performed by Philosopher[46]. PTM-Shepherd was then run via command line to enable the reporting of isotopic peaks.

For the secondary analysis wherein known and discovered diagnostic ions were quantified, PTM-Shepherd's "Diagnostic Feature Extraction" module was used with the ion list presented in Fig. 2b. This was performed using the mzMLs rather than the deneutrallossed and deisotoped[24] mgf files from MSFragger to prevent neutral losses that would be correlated under normal conditions from being anticorrelated in the analysis. For the tertiary analysis wherein the landscape of diagnostic features was explored, PTM-Shepherd was rerun, but with the filtering parameters for diagnostic ions and peptide ions set to 0 for "Min. % of spectra with ion" and 1 for "Min. intensity fold change." PSM-level correlations for all PSMs with mass shifts >50 Da were computed for each diagnostic ion present in more than 50% of glycan mass shifts. Spectral ions are normalized to the base peak by default, creating nonlinear relationships between some ions and necessitating the use of Spearman's rank correlation. Correlation was calculated using the Pandas package in Python.

The third dataset consists of a novel protocol for photoactivatable ribonucleoside-crosslinking from the ProteomeXchange repository PXD023401[26]. Only the two 4SU nucleotide-specific raw files from this repository were used. Samples were acquired on a Thermo Fisher Orbitrap Fusion Lumos using HCD fragmentation. Only the two 4SU-specific raw files from the repository were using in this analysis, and both samples were processing using FragPipe v18.0 directly without conversion to mzML. Samples were processed three times. The first, to find diagnostic features, was a standard open search using the FragPipe default "Open" workflow but with "Write calibrated MGF" and PTM-Shepherd's "Diagnostic Feature Discovery" enabled with default settings. The second, to validate fragment remainder masses, was adapted from the default "Mass-Offset-CommonPTMs" workflow but with the mass offsets limited to 0, 226.0594, and 94.0168; "Labile modification search mode" enabled; "Y ion masses" and "Diagnostic fragment masses" removed; "Remainder masses" set to 94.0168 and 76.9903; "Write calibrated MGF" enabled; and PTM-Shepherd's "Diagnostic Feature Discovery" enabled with default settings. The settings for the third analysis to validate an ammonium loss were identical to the second but without the 76.9903 fragment remainder mass. All analyses were run against the Uniprot database described above. Crystal-C[49] was used to clean up open search results. Filtering to 1% PSM, peptide, and protein levels was performed by Philosopher[46].

The fourth dataset consists of two samples from the ProteomeXchange repository PXD004245 corresponding to ADPR -enriched samples of mouse and HeLa origin[39]. The former is derived from mouse liver, processed in triplicate, and was acquired on a Thermo Fisher Orbitrap Q-Exactive Plus instrument in DDA mode using HCD. The latter was treated with H_2O_2 to induce oxidative stress, then collected in the same manner described above. Raw files were converted to mzML using Proteowizard v3.0.19296 with vendor peakpicking enabled. Both datasets were searched against their respective Uniprot reviewed sequence databased with decoys and common contaminants appended, with the mouse database retrieved on 27 September 2021 and the human database described above. Both datasets were searched separately in FragPipe v18.0 using the default "Labile_ADPR-ribosylation workflow with a few changes. During the MSFragger search, "Report mass shift as variable mod" was set to "No" so that PTM-Shepherd would register these ADPRs as mass shifts and "Write calibrated MGF" was enabled for the PTM-Shepherd diagnostic feature mining module. PeptideProphet[50] and ProteinProphet defaults for "Offset search" were loaded, then PTM-Shepherd and its "Diagnostic Feature Discovery" Module were enabled.

One additional dataset was used to evaluate PTM-Shepherd specificity. This dataset consists of a subset of the human proteome dataset obtained from PXD010154[25]. Specifically, the 36 raw files corresponding to the brain tissue labeled "V102" tissue were processed. This was acquired on a Q Exactive Plus in DDA mode using HCD fragmentation. Raw files were searching directly in FragPipe v18.0 as described above using the default "Open search" workflow with PTM-Shepherd's "Diagnostic Feature Discovery" Module enabled.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in this manuscript is publicly available and previously published. The following datasets were downloaded from ProteomeXchange: the cysteine probes data are available under the accession code PXD028853, RNA crosslinking data are available under accession code PXD023401, ADP-ribosylation data are available under accession code PXD004245 and human brain proteome data are available under accession code PXD010154. IMAC-enriched glycan datasets were downloaded from the CPTAC data portal (https://proteomic.datacommons.cancer.gov/pdc/) and are available under study identifier PDC000471. All data was searched against the Uniprot reviewed protein sequences database retrieved on 13 June 2021. Source data are provided with this paper.

Code availability

PTM-Shepherd is open source under an Apache 2.0 license and can be found at https://github.com/Nesvilab/PTM-Shepherd. The version used in the manuscript has been publicly released (https://github.com/Nesvilab/PTM-Shepherd/releases/tag/v2.0.0). PTM-Shepherd is distributed as part of FragPipe [https://github.com/Nesvilab/FragPipe]. Custom secondary analysis and graphing scripts have been archived on Zenodo (https://doi.org/10.5281/zenodo.8056053).

References

1. Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9**, 4632–4641 (2009).
2. Reiding, K. R., Bondt, A., Franc, V. & Heck, A. J. The benefits of hybrid fragmentation methods for glycoproteomics. *TrAC - Trends Anal. Chem.* **108**, 260–268 (2018).
3. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
4. Lu, L., Riley, N. M., Shortreed, M. R., Bertozzi, C. R. & Smith, L. M. O-Pair Search with MetaMorpheus for O-glycopeptide characterization. *Nat. Methods* **17**, 1133–1138 (2020).
5. 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* **14**, 513–520 (2017).
6. Zeng, W.-F., Cao, W.-Q., Liu, M.-Q., He, S.-M. & Yang, P.-Y. Precise, fast and comprehensive analysis of intact glycopeptides and modified glycans with pGlyco3. *Nat. Methods* **18**, 1515–1523 (2021).
7. Polasky, D. A., Yu, F., Teo, G. C. & Nesvizhskii, A. I. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods* **17**, 1125–1132 (2020).
8. Chen, G., Zhang, Y., Trinidad, J. C. & Dann, C. III Distinguishing sulfotyrosine containing peptides from their phosphotyrosine counterparts using mass spectrometry. *J. Am. Soc. Mass Spectrom.* **29**, 455–462 (2018).
9. Everley, R. A., Huttlin, E. L., Erickson, A. R., Beausoleil, S. A. & Gygi, S. P. Neutral Loss Is a Very Common Occurrence in Phosphotyrosine-Containing Peptides Labeled with Isobaric Tags. *J. Proteome Res.* **16**, 1069–1076 (2017).
10. Zolg, D. P. et al. ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides. *Mol. Cell Proteom.* **17**, 1850–1863 (2018).
11. Drewes, G. & Knapp, S. Chemoproteomics and chemical probes for target discovery. *Trends Biotechnol.* **36**, 1275–1286 (2018).
12. Riley, N. M., Hebert, A. S., Westphall, M. S. & Coon, J. J. Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nat. Commun.* **10**, 1–13 (2019).
13. Kelstrup, C. D., Frese, C., Heck, A. J., Olsen, J. V. & Nielsen, M. L. Analytical utility of mass spectral binning in proteomic experiments by SPectral Immonium Ion Detection (SPIID). *Mol. Cell Proteom.* **13**, 1914–1924 (2014).
14. Wang, J. et al. A turn-key approach for large-scale identification of complex posttranslational modifications. *J. Proteome Res.* **13**, 1190–1199 (2014).
15. Dorl, S., Winkler, S., Mechtler, K. & Dorfer, V. PhoStar: identifying tandem mass spectra of phosphorylated peptides before database search. *J. Proteome Res.* **17**, 290–295 (2018).
16. Altenburg, T., Giese, S. H., Wang, S., Muth, T. & Renard, B. Y. Ad hoc learning of peptide fragmentation from mass spectra enables an interpretable detection of phosphorylated and cross-linked peptides. *Nat. Mach. Intell.* **4**, 378–388 (2022).
17. Zanon, P. R. et al. Profiling the proteome-wide selectivity of diverse electrophiles. *ChemRxiv* (2021).
18. Abbasov, M. E. et al. A proteome-wide atlas of lysine-reactive chemistry. *Nat. Chem.* **13**, 1081–1092 (2021).
19. Yan, T. et al. Enhancing Cysteine Chemoproteomic Coverage Through Systematic Assessment of Click Chemistry Product Fragmentation. *Anal. Chem.* **94**, 3800–3810 (2022).
20. He, J.-X. et al. pChem: a modification-centric assessment tool for the performance of chemoproteomic probes. *bioRxiv* (2021).
21. Storck, E. M. et al. Dual chemical probes enable quantitative system-wide analysis of protein prenylation and prenylation dynamics. *Nat. Chem.* **11**, 552–561 (2019).
22. Virreira Winter, S. et al. EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15**, 527–530 (2018).
23. Geiszler, D. J. et al. PTM-Shepherd: analysis and summarization of post-translational and chemical modifications from open search results. *Mol. Cell Proteom.* **20,** 100018 (2021).
24. 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.* **20**, 498–505 (2020).
25. Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).
26. Bae, J. W., Kim, S., Kim, V. N. & Kim, J.-S. Photoactivatable ribonucleosides mark base-specific RNA-binding sites. *Nat. Commun.* **12**, 1–10 (2021).
27. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
28. Vékey, K. et al. Fragmentation characteristics of glycopeptides. *Int. J. Mass Spectrom.* **345**, 71–79 (2013).
29. Clark, D. J. et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **179**, 964–983.e931 (2019).
30. Palmisano, G. et al. Selective enrichment of sialic acid–containing glycopeptides using titanium dioxide chromatography with analysis by HILIC and mass spectrometry. *Nat. Protoc.* **5**, 1974–1982 (2010).
31. Larsen, M. R., Jensen, S. S., Jakobsen, L. A. & Heegaard, N. H. Exploring the sialiome using titanium dioxide chromatography and mass spectrometry. *Mol. Cell Proteom.* **6**, 1778–1787 (2007).
32. Polasky, D. A., Geiszler, D. J., Yu, F. & Nesvizhskii, A. I. Multi-attribute Glycan Identification and FDR Control for Glycoproteomics. *Mol. Cell Proteom.* **21**, 100205 (2022).
33. Pett, C. et al. Effective assignment of α2, 3/α2, 6-sialic acid isomers by LC-MS/MS-based glycoproteomics. *Angew. Chem. Int. Ed.* **57**, 9320–9324 (2018).
34. Medzihradszky, K. F., Kaasik, K. & Chalkley, R. J. Characterizing sialic acid variants at the glycopeptide level. *Anal. Chem.* **87**, 3064–3071 (2015).
35. Hoffmann, M. et al. The fine art of destruction: a guide to in-depth glycoproteomic analyses—exploiting the diagnostic potential of fragment ions. *Proteomics* **18**, 1800282 (2018).

36. Bern, M., Kil, Y. J. & Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinform.* **40**, 20. 11–13.20. 14 (2012). 13.
37. Bonfiglio, J. J. et al. An HPF1/PARP1-based chemical biology strategy for exploring ADP-ribosylation. *Cell* **183**, 1086–1102. e1023 (2020).
38. Jungmichel, S. et al. Proteome-wide identification of poly (ADP-Ribosyl) ation targets in different genotoxic stress responses. *Mol. Cell* **52**, 272–285 (2013).
39. Martello, R. et al. Proteome-wide identification of the endogenous ADP-ribosylome of mammalian cells and tissue. *Nat. Commun.* **7**, 1–13 (2016).
40. Gehrig, P. M. et al. Gas-phase fragmentation of ADP-ribosylated peptides: arginine-specific side-chain losses and their implication in database searches. *J. Am. Soc. Mass Spectrom.* **32**, 157–168 (2020).
41. Savitski, M. M. et al. Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on Orbitrap-type mass spectrometers. *Anal. Chem.* **83**, 8959–8967 (2011).
42. Schweppe, D. K. et al. Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *J. Proteome Res.* **19**, 2026–2034 (2020).
43. Potel, C. M., Lemeer, S. & Heck, A. J. Phosphopeptide fragmentation and site localization by mass spectrometry: an update. *Anal. Chem.* **91**, 126–141 (2018).
44. MacCoss, M. J., Noble, W. S. & Käll, L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).
45. Dančík, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **6**, 327–342 (1999).
46. da Veiga Leprevost, F. et al. Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870 (2020).
47. Edwards, N. J. et al. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **14**, 2707–2713 (2015).
48. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).
49. Chang, H. Y. et al. Crystal-C: A Computational Tool for Refinement of Open Search Results. *J. Proteome Res.* **19**, 2511–2515 (2020).
50. Choi, H. & Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **7**, 254–265 (2008).

Acknowledgements

National Institutes of Health grants R01GM094231, 5R01GM135504, and U24CA271037 provided funding for AIN, F.Y., and D.A.P. D.J.G. was supported in part by the Proteogenomics of Cancer Training Program (5T32CA140044). We would like to thank Jong Woo Bae et al. for providing early access to ProteomeXchange repository PXD023401. We would also like to thank our users for providing feedback on our tools.

Author contributions

D.J.G. developed the algorithm and wrote the software, with D.A.P. and F.Y. assisting in the development of the algorithm and D.A.P. assisting in the development of the software; D.J.G. and A.I.N. jointly analyzed the data and conceived the project while A.I.N. supervised the project; D.J.G. and D.A.P. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-39828-0.

Correspondence and requests for materials should be addressed to Alexey I. Nesvizhskii.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023