Computational Methods for Characterizing Post-translational and Chemical Modifications Found in Open Searches

by

Daniel Geiszler

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2022

Doctoral Committee:

Professor Alexey I. Nesvizhskii, Chair
Professor Kristina Hakansson
Associate Professor Alla Karnovsky
Associate Professor Ryan Mills
Associate Professor Arvind Rao

Daniel J. Geiszler

geiszler@umich.edu

ORCID ID: 0000-0002-7691-8534

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Alexey Nesvizhskii. His guidance and advice have made me a better scientist in every way, just like his unparalleled ability to find bugs has done for my software. I would also like to thank my committee members, Professors Kristina Hackansson, Ryan Mills, Arvind Rao, and Alla Karnovsky for being with me throughout this process.

Alexey wasn't alone in the lab. My colleagues that have put up with my incessant yammering and off-topic conversations have made the last five years some of the best of my life. From Chinese history to rocket ships and climate, these were what kept me motivated to go to work every day. I would especially like to thank three people who have disproportionately helped me in my scientific development: Dr. Andy Kong, who instilled in me a love for proteomics; Dr. Daniel Polasky, who instructed me in the intricacies of instrumentation and fragmentation; and Dr. Fengchao Yu, who always knew of a paper showing that someone had stolen my latest great idea and published it forty years ago. The whole lab fostered an environment of collaboration, and everyone was always ready to discuss exciting new ideas. For that, thank you.

On collaboration, I would also like to mention the collaborators from outside the lab I've had the pleasure of working with. Professor Keriann Backus from the University of California Los Angeles—and her lab members Nikolas Burton and Tianyang Yan—have provided us with data and partnerships and inspired in part one of the chapters of this dissertation.

Finally, I'd like to thank my family. I would never have made it to graduate school without my parents fostering my curiosity, and I would never have made it through graduate school without my incredible wife's support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Post-translational modifications (PTMs) govern many processes within cells and understanding their function is critical to both the basic and biomedical sciences. However, identifying modified peptides, particularly unexpected and rare modifications, remains a challenge to proteomics researchers. Recent advances in proteomics search tools have expanded the capacity to identify the entire modification landscape in an unbiased way, but the modifications identified in this manner—called "open searching"—require extensive post-processing to elucidate their identities. In this dissertation, I develop computational methods to characterize and identify modifications derived from open searches.

In Chapter Two, I develop a method for comprehensively characterizing open search results, PTM-Shepherd, enabling new applications for unbiased PTM discovery. PTM-Shepherd automates characterization of PTM profiles detected in open searches based on attributes such as amino acid localization, fragmentation spectra similarity, retention time shifts, and relative modification rates. I show how open searches can be used to profile experimental artifacts by identifying a set of PTMs common across several formalin-fixed paraffin-embedded datasets that researchers can include in future analyses, identifying a range of Cys-specific artifacts in a commonly used high-quality dataset, finding two previously undescribed PTMs in synthetic peptide data and TMT data, and tracing major site-specific PTM batch effects in a multi-university consortium's proteomics data back to sample processing. In Chapter Three, I extend the algorithm developed in Chapter Two, introducing additional metrics that allow researchers to peer into the spectra of PTMs and extract PTM-specific fragmentation patterns and diagnostic ions. I find new diagnostic for multiple common PTMs, including identifying new fragmentation patterns for glycopeptides under high energy fragmentation, new diagnostic ions for sialic acid under high energy fragmentation, and new diagnostic ions and peptide remainder masses for ADP-ribosylation, as well as examining general trends in the utility of PTM-specific diagnostic features such as the inverse relationship between an ion's average intensity and its specificity to

the modification. In Chapter Four, I expound my methods' utility by applying it in multiple settings to characterize synthetic and chemical PTMs. In each case, I show how these methods aid in interpretation of results or increase coverage of the proteome by recovering additional modified peptides. For chemoproteomics probes, I demonstrate how expensive isotopic labeling to identify fragmentation patterns can be avoided, finding multiple novel diagnostic ions for a Cys-specific triazole biotin probe. For RNA crosslinked data, I show how the number of recovered identifications increases by up to 50% over existing state of the art methods when incorporating fragmentation information discovered by PTM-Shepherd. Finally, for protein-protein crosslinking, I show how PTM-Shepherd can derive fragmentation patterns for non-cleavable crosslinkers without computationally expensive or custom workflows, discovering that auto crosslinks can be used to identify fragment remainder masses for that can reduce computational complexity during searching.

The ability to survey the entire post-translational modification landscape has major implications across proteomics subdisciplines. In total, the work described herein represents a major milestone in the interpretation of open search results and opens the door to better understandings of cellular processes and disease by facilitating new modes of analysis.

CHAPTER I

Introduction to the Analysis of Post-translational Modifications in Proteomics

1.1 The importance of post-translational modifications

The central dogma of biology presents a triumvirate of biological molecules that govern all aspects of cellular function: DNA is transcribed to RNA which is then translated to proteins[1]. But the regulation of cellular processes is much more complicated than this model suggests. Proteins, the workhorses of cells that perform nearly all cellular functions, can undergo extensive co- and post-translational modification (PTM) to regulate their activity. Hundreds of biological PTMs govern many dynamic interactions between proteins and genomes, transcriptomes, and proteomes[2], resulting in a vast number of potential proteoforms[3]. In other words, they play an important role at every stage of the protein lifecycle[4,5].

For example, N-linked glycosylation, one of the most abundant covalent PTMs in eukaryotic organisms, has be found to modify nascent proteins as they are translated. facilitating the fundamental process of protein folding[6]. Most eukaryotic proteins are immediately processed to remove their initiator Met[7], a PTM correlated with protein half-life[8], effectively starting the clock on their lifespans. After translation some proteins will sit in an inactive form waiting to be triggered--such as by a phosphorylation event--causing events ranging from being transported to a different area of the cell[9] to starting an oncogenic signal cascade[10]. Their purpose fulfilled, they can then be recycled for parts. The time on the clock that was begun by PTM-induced cleavage of the initiator Met, runs out via PTM-induced degradation. The transferring of ubiquitin, a protein in its own right, as a PTM can induce a protein's passage to the proteasome for its degradation[11]. But even as a peptide degrades its modifications play a role; remaining peptides from its degradation are sent to cell surface markers, where they interact with immune cells. In some cases, the immune system will specifically recognize peptides containing PTMs as markers of disease, targeting the host cell for death [12,13].

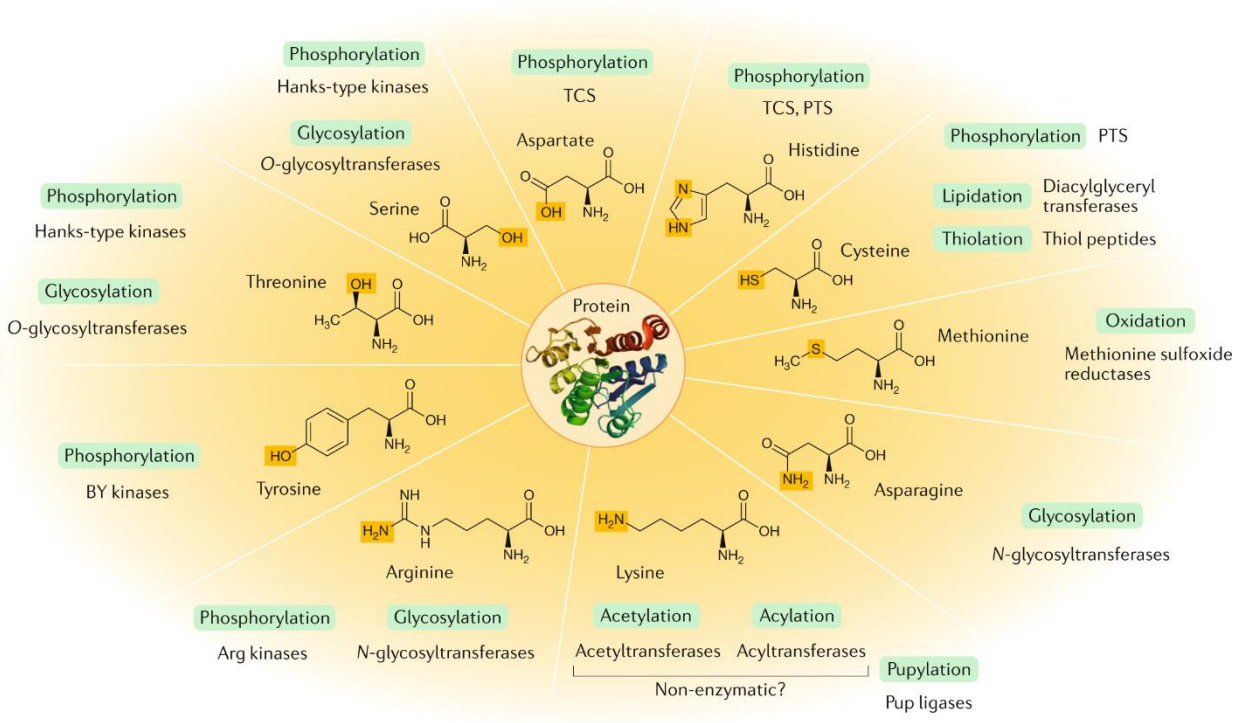

Figure 1-1: Examples of PTMs occurring in bacteria. The diversity of PTMs found in bacteria is mirrored in many other organisms. Residues can be subjected to many types of PTMs that radically alter their chemistry. Similarly, the same modifications can appear on a variety of amino acids. This phenomenon motivates localization algorithms for PTMs. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Nature, Nature Microbiology, Macek et al. (2019)[20], Copyright © 2019, Springer Nature Limited.

PTMs are a fundamental and integral part of the protein lifecycle that participate in protein synthesis, folding, signaling, trafficking, degradation and processing, and immune response. Characterizing the state of PTMs in proteins across the proteome is thus essential to a wide range of basic and translational research. To this end, recent years have seen large-scale efforts to augment our ability to identify modified peptides and proteins[14-19]. Despite this study of PTMs still presents challenges to proteomics researchers, with incorporating PTMs into proteome-wide analyses remaining a significant obstacle in understanding cell function and disease. In particular, the diversity of PTMs (Fig 1-1). that participate in cellular regulation adds complications to proteomics analyses.

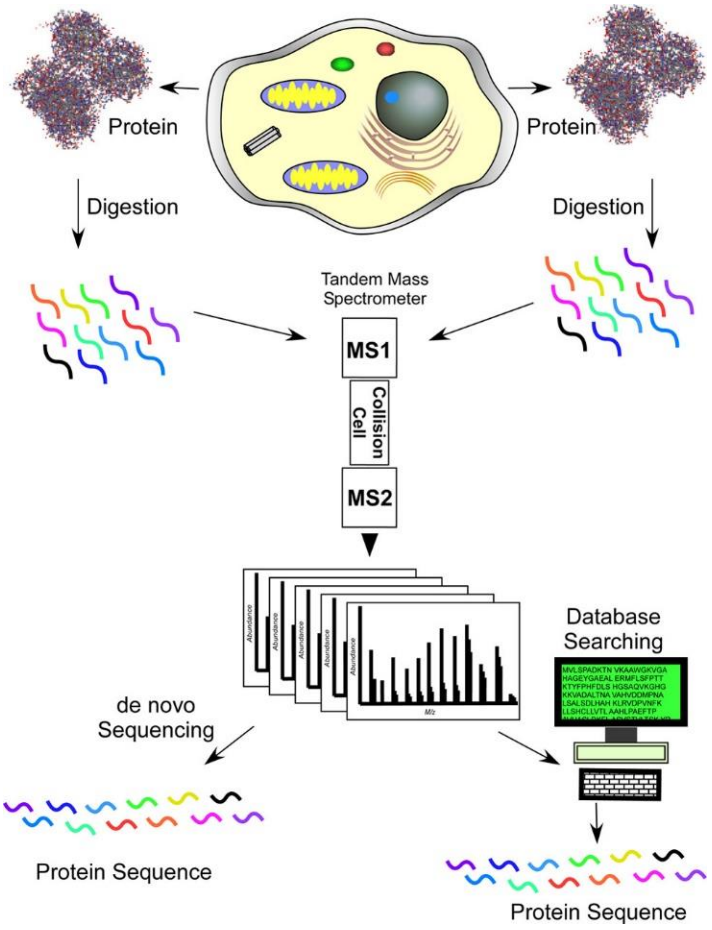

Figure 1-2: Outline of a proteomics experiment. Proteomics experiments begin with the extraction of proteins from the sample of interest, most commonly a biological tissue. Proteins are digested into peptides that generally have basic terminal residues to facilitate ionization. Following this, they might be fractionated before being separated using an LC-MS instrument. Ionization and entrance into the MS occur after being eluted from the LC. Two layers of mass analysis are done: MS1 on the full peptide and MS2 on its characteristic fragments after it has gone through the collision cell. Resultant data can be searched with a variety of algorithms to extract peptide sequence, but these methods generally fall under either database search or *de novo* search algorithms. Reprinted with permission from Yates (2013)[22]. Copyright © 2013 American Chemical Society.

1.2 LC-MS/MS in proteomics

Mass spectrometry (MS) has emerged as the most popular way to identify proteins at whole-proteome scales in complex mixtures, such as biological fluids or tissues[21]. Its workflow is laid out in Figure 1-2. The basic principle of MS is that the masses of analytes can be deduced by

manipulating ions using electric and/or magnetic fields in a vacuum to determine their mass. Analytes' masses can be measured in many ways, including for example, by observing the oscillating image current induced by ions orbiting in an electrostatic trap (for Orbitrap-type analyzers) or the time taken for an ion to travel a precisely known distance (for time-of-flight mass analyzers), values which depend only on the mass-to-charge ratio (m/z) of the ion[23]. An analyte's m/z is not always specific enough to identify it, however. Tandem mass spectrometry (MS/MS or MS2) solves this by performing multiple rounds of mass analysis to obtain additional information about the analyte. After analyzing an intact analyte, the analyte is fragmented, most commonly by deliberately colliding the analyte with inert gas molecules, and the resulting fragments of the molecule are then mass analyzed themselves. The MS/MS spectrum is a molecular fingerprint that provides information about a molecule's substructures[24].

Tandem MS facilitates peptide identification by providing sequence information in the MS/MS spectrum. Peptides primarily fragment along their backbone, producing fragment ions that can indicate the sequence of amino acids in the peptide[25]. Breakage at the peptide backbone produces two complementary molecules corresponding to each terminus of a peptide. When at least one of the molecules is charged, it appears as an ion. The repeating structure of the peptide backbone means that this process takes place at multiple points in the peptide, with differences between the resultant ions corresponding to amino acid masses. Predictable fragmentation pathways therefore allow peptide sequence deconvolution, as copies of the peptide will break at different locations[26]. These sequence-informative pathways compete with other fragmentation pathways that do not produce sequence information, and peptide sequencing errors can occur due to resulting gaps in the sequence ladder or false ions existing at the same (within a tolerance) m/z as predicted ones. Ions from secondary fragmentation pathways and co-fragmented peptides can both contribute peaks that can be mistakenly interpreted as coming from the peptide sequence, requiring advanced processing methods for automated interpretation of peptide tandem mass spectra, which will be discussed below.

MS-based proteomics begins with the digestion of proteins into peptides[21]. Despite losing information about a peptide's proteoform-of-origin, this provides two critical advantages. First, smaller analytes are typically easier to ionize and, by choosing proteases that cleave the peptide at Lys and Arg, peptides are left with basic C-termini that further enhance ionization[21,27]. Second, it reduces the number of possible peptides to only those with termini corresponding to the enzyme used. After digestion, other sample preparation procedures can be done to facilitate downstream analysis. These include peptide barcoding such as the addition of isobaric tandem mass tags (TMT)[28] or chemical proteomics reagents[29].

Following this, peptides are typically separated by physicochemical characteristics to reduce sample complexity[30]. In some cases, multiple rounds of separation are employed. A second type of separation is done online, connected to the mass spectrometer. Liquid chromatography (LC) separates peptides[21] based on polarity, with non-polar peptides having the longest retention times. Peptides are eluted from the LC directly into the MS via electrospray ionization (ESI), giving rise to the term LC-MS. Front-end separation of peptides, and especially multi-step fractionation and separation, reduces the number of peptides entering the instrument at once, decreasing ion suppression[31] and co-fragmentation and increasing sequencing depth.

As peptides are eluted from the LC column, ESI aerosolizes them into charged droplets. The droplets enter a vacuum chamber where they evaporate, concentrating the charge into a smaller space until the droplet bursts. This continues happening until peptide ions are in the gas phase devoid of any solvent. ESI is a "soft ionization" technique, so peptides can take on a range of charge states without fragmenting[32]. Since multiple peptides are often eluted into the mass spectrometer at a time, the MS separates ions first based on their m/z and measures the intact peptide (MS1). Many copies of the same peptide are also eluted into the MS simultaneously.

At this point, typical proteomics workflows diverge into two groups: data-dependent acquisition (DDA) and data-independent acquisition. In the case of DDA, instrument control software will check to see if a peptide ion has been analyzed recently, selecting it for MS/MS if it has not[33].

Ion selection for MS/MS is typically performed by an intermediate, low resolution mass analyzer, and is less precise than typical MS1 measurements. This leads to multiple peptides being selected for fragmentation at the same time if their m/z values are close to each other. Depending on the density of peptides in the sample and separation and MS settings, it is possible for many, if not most, MS2 spectra to contain multiple co-fragmented peptides. Selecting and fragmenting individual peptide ions is a time-consuming process that leads to some peptide ions being skipped. DIA analysis remedies this by embracing co-fragmentation[34]. DIA workflows use wide isolation windows to allow many peptides to be fragmented simultaneously. By not selecting ions in a data-dependent manner and increasing the ion isolation width, the number of ionized peptides fragmented and detected in a DIA MS2 spectrum goes up. Theoretically, this allows the recovery of less abundant peptides, but at the cost of increased complexity in the MS2 spectrum that must be handled computationally.

1.3 Peptide sequencing algorithms identifying modified peptides

The core of peptide identification is the search algorithm which searches MS/MS spectra for peptide sequences. These can be segregated into two camps, although many algorithms incorporate elements from both: *de novo* search algorithms and database search algorithms. *De novo* algorithms, such as SHERENGA[35] and PEAKS[36], look at the peaks in the spectrum, calculate distances between peaks that may correspond to amino acid masses, and compute a score for how well a spectrum matches a theoretical peptide. Database search algorithms, such as Sequest[26] and Comet[37], rely on a reference database of protein sequences expected to be in the sample. Search engines produce similar outputs regardless of their algorithm[38]. This is primarily a score that measures how well the top peptide hit for a spectrum corresponds to a theoretical peptide spectrum derived from the digested reference database spectrum, and a calibrated version of that score that is used to compare peptide hits between spectra[39]. To score peptides from a theoretical set of proteins, anything done to the *in vitro* sample needs to be mirrored in the *in silico* sample. The most obvious of these is digestion of proteins into peptides to create an *in silico* peptide reference database from the protein database. But PTMs must also be included in

the search space in order for modified peptides to be recovered. How different algorithms deal with this challenge will be the focus of this section.

The most straightforward approach to incorporating PTMs into the search space is including that as either fixed or variable (also known as dynamic) modifications during a database search (Figure 1-3a). Fixed modifications amount to replacing any amino acid mass with a different mass, or adding a mass to any residue of a particular type in the database. Any peptide fragments containing the modification site are searched for in the spectrum at a mass equal to the fragment plus the modification mass. Only modifications which occur with very high fidelity are generally included in this manner—examples include isobaric tags or Cys alkylation—because any peptide containing unmodified residues will be missed. Variable modifications function the same way as fixed modifications, i.e, shifting peptide fragment ions by the mass of the modification. However, this is done in a way such that the reference database contains both modified and unmodified versions of the peptide. It has the unfortunate side effect of increasing the size of the reference peptide database in a combinatorial manner. When considering a peptide with 10 possible modification sites and a maximum of 3 modifications per peptide, the database will contain (10C3 + 10C2 +10C1 =) 175 copies of the original peptide. Inflating the search space like this increases the time complexity of the search and increases the likelihood of matching a spectrum by chance, which makes it harder to separate true from false peptide matches[40]. It is also unable to identify peptides bearing modifications that were not specified, creating a fundamental trade-off between identifying uncommon modifications and maintaining a reasonable search space.

De novo approaches work without a reference database. The MS1 mass is used to generate a list of all possible peptide sequences the spectrum could correspond to. They then identify candidate fragment ions from the spectrum by looking for common losses from peaks. Once fragment ions have been identified, the peptide is sequenced by looking at the distances between peaks in the spectrum and checking for corresponding amino acid masses. In some cases, modifications can be included in the amino acid list as well[41].

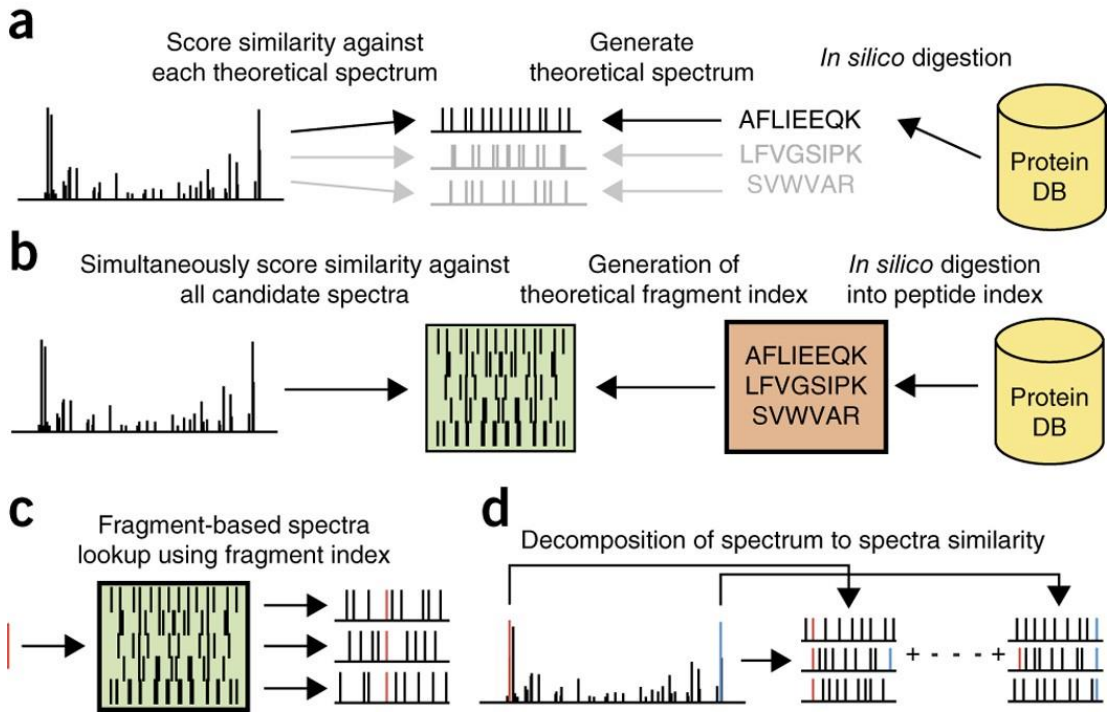

Figure 1-3: Database search strategies pre- and post-ion indexing. a) Traditional database search strategy. Reference proteins are digested *in silico*, then each spectrum is matched to a list of peptides with a potential matching mass before being score individually. b) Ion index search strategy. In this strategy, the reference peptide database is indexed for fast access. c) Spectra are queried ion-by-ion against the reference database, which is dramatically sped up by the indexing process. All peptides with matching precursor masses are queried and incremented simultaneously. d) Similarity scores are calculated for each potential PSM for a spectrum, then the top hit undergoes score calibration to enable comparisons between scans. Reprinted with permission Springer Nature Customer Service Centre GmbH: Nature, Nature Methods, Kong et al. (2017)[15], Copyright © 2017, Springer Nature Limited.

Tag-based approaches, such as TagRecon[42] and PIPI [43], and have also become increasingly popular. This approach straddles the line between *de novo* searches and database searches. Similar to the *de novo* approach, distances between peaks inform the search engine about the amino acids candidate peptides can contain. Rather than trying to directly extract sequence information from the spectrum itself, sequence tags spanning multiple residues are used to restrict the database search space to only peptides containing those tags. Plus, since the tags encode information about the positions of fragment ions relative to each other, they are not sensitive to shifts induced by upstream modification sites and are thus modification agnostic.

Open searches, or mass-tolerant searches, are a variant of the database search strategy that relaxes the error tolerances around peptide masses, producing a mass shift corresponding to the deviation between a peptide's experimental and theoretical masses[44]. That deviation amounts to the sum of all modifications and mutations contained by the peptide. Programs with the ability to perform open searches include MSFragger[15] and Sequest[26]. This strategy has increased in popularity thanks to improvements in algorithms that have dramatically reduced the computing time for database searches by orders of magnitude (Fig 1-3b-d). Because the search space in open searches includes "chemically unlikely" space, or spaces that are unlikely to contain any modifications, peptides are more likely to be incorrect hits. Mass offset or multi-notch searches[45] rectify this by restricting possible mass deviations to a subset of the modification space that is most likely, such as the masses of commonly occurring modifications.

Although the strategies discussed thus far treat modifications as static molecules attached to fragmenting peptides, they can also undergo fragmentation themselves. One such example is phosphorylation. It can dissociate from the peptide as a neutral species during fragmentation--a neutral loss[46]. Phosphorylation is commonly lost during fragmentation and consequently fails to produce intense fragment ions shifted by the mass of the modification that can be identified during a standard variable modification search. Some search engines allow modification-specific neutral losses to be searched as an additional ion series alongside standard peptide fragments.

In other cases, PTMs undergo much more complex fragmentation than can be captured with this method. N-glycosylation exemplifies this, spawning an entire ecosystem of tools designed around finding glycopeptide-specific fragmentation patterns[47-49]. It can produce three distinct fragmentation types in addition to fragmentation of the peptide itself, that frustrate traditional search algorithms: intact peptide ions with partial glycan attached, peptide fragment ions with partial glycans attached, diagnostic ions from the glycan chain with no sequence specificity[50]. Glycopeptide search engines utilize these to score glycopeptides, evaluate glycan composition and structure, identify spectra containing glycopeptides, and more. Other modifications, such as

ADP-Ribosylation (ADPR), can exhibit the same types of complex modification fragmentation and are amenable to similar search techniques[51]. MSFragger's labile search mode is general enough to accommodate other labile PTMs such as these. This application involves a modified mass offset search, wherein additional ion series encode partial modifications attached to intact and fragmented peptide ions.

While all these techniques produce roughly the same output about the peptide backbone fragments, they do not all provide the same information about the modifications on them. *De novo* and database searches rely on prior knowledge of PTMs before the search, and as a result produce output that includes the modification identity and localization (although these can be inaccurate). Contrarily, open searches do not. A peptide's open search excess mass contains no information about modification identity or localization. Derivatives of the open search such as the mass offset or labile search provide slightly more information about modification identity due to the constrained search space, but still no localization information. Open searches are a powerful tool for PTM discovery, but interpreting their results can still be challenging. Shedding light on PTMs derived from open searches will be the focus of this dissertation.

1.4 Validation of modified peptides

Peptide sequencing is an error prone process, so post processing is done to control the False Discovery Rate (FDR) among peptide candidates[38]. Empirical null distributions of false peptide hits can be estimated by searching decoy peptides alongside a reference database[52]. The decoy peptides consist of reversed peptide sequences to maintain theoretical fragmentation characteristics of true peptides. By constructing a ranked list of all peptide-spectrum-matches (PSMs), the local FDR can be estimated and controlled via examination of the local proportion of decoys. Search engine scores alone do a poor job of separating true from false PSMs, so tools have been developed to better model these distributions and increase the sensitivity of analyses.

The first such tool is PeptideProphet[53]. PeptideProphet fits a mixture model of two distributions to the search engine scores. At a given score, the probability that a PSM corresponds to a true hit

is the portion of the mixture model contributed by the higher-scoring (true positive) component. By accounting for other peptide properties such as length and number of enzymatic termini, PeptideProphet better captures the complexity of PSM assignment than a single search engine score. It can also explicitly take open search mass shifts into account by independently modeling mass errors within 1 Da bins spanning the search range. It also supports mass offset and labile searches by modeling mass errors as deviations from the closest mass shift in the mass shift list rather than from the unmodified peptide[15].

Percolator[54] can also perform flexible PSM post-processing. Percolator implements a support vector machine to find characteristics that best separate target from decoy PSMs. Unlike PeptideProphet, Percolator takes a table of scores as input and is flexible enough to include any numeric score. This leaves the door open for supporting any search method, including mass offset and open searches. Despite its success, Percolator is restricted to linear models, so nonlinear variants have been introduced to capture additional complexity[55].

1.5 Characteristics of modified peptides

The wide array of effects PTMs can have on mass spectra means that the same algorithm may not be appropriate for all cases, and spectrum interpretation can remain difficult regardless of which approach is used. In these cases, secondary characteristics of PTMs can augment search engines and increase PTM recovery.

Unequivocally, the most sought-after information about a PTM is its localization within a peptide sequence and ultimately the corresponding protein sequence. Search engines with variable modifications enabled provide some information about where the PTMs present on a peptide but rarely a measure of confidence about the recovered residues for it. This is problematic in cases where gaps in peptide ion series overlap multiple possible modification sites and a PTM's localization is ambiguous. Many scores and tools have been developed to perform post-hoc localization of PTMs after peptide identification[56-59]. Some rely on decoy amino acids to calculate local False Localization Rates on arbitrary scores (FLRs)[59] while others attempt to

calculate the probability of true hits directly[56]. Localization algorithms tend to be geared toward relocalizing variable modifications after peptide identification based on a handful of possible amino acids that can be modified. But open searches are residue-agnostic and considering every residue on a peptide as a potential localization site can violate the assumptions of localization algorithms and scores.

Peptide retention time is another such characteristic. Liquid chromatography separates peptides based on their polarity prior to entering the mass spectrometer, and the amount of time they are retained in the column before exiting is referred to as their retention time. Modifications affect peptide retention times by changing a peptide's polarity, for example by adding highly polar or nonpolar moieties or changing the peptide's conformation, and their effects on retention time can be gathered by comparing modified and unmodified versions of the same peptide[14]. Some analyses, such as DIA peptide identification, rely on peptide retention time[60,61] to restrict the search space because spectra containing dozens of cofragmented peptides would otherwise be too noisy to extract reliable identifications. This principle can also be extended to DDA analyses. Percolator can rescore peptide identifications by incorporating additional features--such as a deviation from the predicted retention time--to better separate true from false peptide hits[62]. Advances in RT prediction with deep learning have extended peptide RT prediction to modification-bearing peptides. Indeed, some clever formulations have even been developed to predict peptide RTs for open search-derived peptides bearing unknown modifications by first learning the effect of atomic composition on RT from the peptide backbone then transferring that to predict the effect of atomic composition on RT from the modification[63].

We previously discussed the effect that modifications can have on peptide fragment ions by shifting their location in the spectrum, i.e., by adding mass to the ions. However, modifications can also change the characteristics of the ions themselves, even ions that do not bear the modification[64]. Isobaric tandem mass tags (TMT) are commonly used to reduce batch effects in proteomics experiments. Proton mobility, one of the factors affecting peptide backbone ion formation, is reduced in TMT-labeled peptides, which has the effect of increasing the number of

neutral losses from peptides[65]. Spectrum prediction tools such as Prosit [66] and DIA-NN[67] have developed PTM-specific models to overcome this problem rather than shifting the ions of unmodified peptides by the mass of the modification.

When these features are incorporated into mainstream proteomics platforms, they tend to be specific to a handful of common PTMs. Their lack of generalizability across PTM types means most modifications are suboptimally identified, so incorporating general models into modified peptide identification is still a fertile field of research with many open questions.

1.6 Outline

Comprehensive PTM searches have many underappreciated benefits. For one, samples can contain unexpected sample specific PTMs that dramatically lower sensitivity when unaccounted for. Additionally, multiple classes of PTMs can interact with each other in ways that are obfuscated from traditional searches. Open searches, a strategy for the unbiased assessment of PTMs, have shown great promise in these regards and have the potential to both facilitate a deeper understanding of the sample and unlock new modes of analysis. Hence, generalizable solutions to characterizing modifications derived from open searches have tremendous potential to augment search engines and increase PTM recovery to better understand the proteome at large.

Open search analysis still presents challenges to proteomics experiments. It provides no information on the identity or localization of modifications. Mass shifts can even correspond to concurrent modification events on the same peptide or be within the range of multiple ambiguous modifications, further frustrating analysis. Furthermore, its inability to identify shifted peptide fragment ions reduces its sensitivity by suppressing the scores of most modification-bearing peptides. Even looking for ions shifted by the supposed modification mass may miss the mark; labile modifications produce unexpected fragment mass shifts and concurrent modifications will produce partially shifted ion series. All these challenges hamstring sensitivity and interpretability.

The goal of this dissertation is to begin addressing some of these shortfalls in PTM analysis using shotgun proteomics with a focus on open search characterization. In Chapter Two, I develop a method for comprehensively characterizing open search results, enabling new applications for unbiased PTM discovery. In Chapter Three, I extend the algorithm developed in Chapter Two, introducing additional metrics that allow researchers to peer into the spectra of PTMs and extract PTM-specific fragmentation patterns and diagnostic ions as well as examining for the first time the general characteristics of PTM-specific diagnostic features. In Chapter Four, I expound my methods' utility by applying it in multiple settings to characterize synthetic and chemical PTMs.

CHAPTER II

PTM-Shepherd: Analysis and Summarization of Post-translational and Chemical Modifications from Open Search Results

This chapter was published in its entirety as *PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications from Open Search Results* in Molecular & Cellular Proteomics

2.1 Introduction

Database searching of shotgun proteomics data is a commonly used strategy for identification of peptides and proteins from complex protein mixtures [30,68]. Peptide identification in this strategy most commonly relies on matching tandem mass spectrometry (MS/MS)-derived peptide spectra to their theoretical counterparts using MS/MS database search tools, which requires prior knowledge of the potential modifications that might be present in a sample. This is problematic, as proteins can exist in myriad forms outside of their canonical sequences. For example, protein function is commonly modulated by post-translational modifications (PTMs), and additional chemical modifications from sample processing can hinder identification. Because the search space of all potential peptides including their modifications is so large, when using conventional database search strategies, researchers are forced to limit the modifications considered by their searches, leading to large number of unexplained spectra [44,69-71].

Open searching, or mass-tolerant searching, is one strategy that allows researchers to expand their search space and reduce the number of unexplained MS/MS spectra. It has proven to be an effective strategy for identifying both known and unknown modifications in shotgun proteomics experiments [44,45,69,72,73]. Rather than being limited to user-specified modifications, open searches identify peptides with mass shifts corresponding to potential modifications or sequence variants. These mass shifts do not, however, contain the same information present in closed searches, most importantly the identity of the modification and what amino acids within the peptide sequence

may contain it. Deciphering open search results thus requires subsequent computational characterization to recover this information [42,69,74,75].

Existing tools for open search postprocessing perform a limited set of analyses on a spectrum-level basis. PTM-Prophet [76], for example, is limited to localizing mass differences for each PSM but does not provide data summaries that can inform subsequent searches nor does it provide identities for mass differences. Philosopher [77] only provides mappings of mass differences to UniMod and generates a basic mass shift histogram. Here we present PTM-Shepherd, an automated tool that calls modifications from open search peptide-spectrum match (PSM) lists and characterizes them based on attributes such as amino acid localization, fragmentation spectra similarity, effect on retention time, and relative modification rates. PTM-Shepherd can also perform multi-experiment comparisons for studying changes in modification profiles under differing conditions. We utilize these profiles in a wide array of situations to show how additional metrics, interexperiment comparisons, and bulk analytical profiles can be helpful in PTM analysis. Overall, we expect that PTM profiles produced by PTM-Shepherd will greatly enhance understanding of the data at both the macro level for quality control and the micro level for specific PTM identification.

2.2 Materials and methods

2.2.1 The PTM-Shepherd algorithm

Mass shift histogram construction

A histogram of identified mass shifts is constructed using all PSMs from the PSM.tsv file (or multiple PSM.tsv files in the case of multi-experiment analysis) generated by the MSFragger/Philosopher pipeline (**Figure 1**). These PSM.tsv files are typically (by default) filtered to 1% PSM-level and 1% protein level FDR using target-decoy counts, as determined by the Philosopher filter command. The widths of each bin the histogram is 0.0002 Da (by default). This histogram is extended by 5 Da on either side of the most extreme values in order to prevent peaks at the maximum and minimum of the histogram from being truncated after smoothing. Random noise between -0.005 Da and 0.005 Da is added to break ties occurring between bin

boundaries and mass shifts. After bin assignment, the histogram is smoothed to make peaks more monotonic. Bin weight is distributed across 5 bins (by default), with the weights assigned to each bin being determined by a Gaussian distribution centered at the bin to be smoothed such that 95% of the bin's weight is distributed between them. Peaks, representing mass shifts of observed modifications, are called from this histogram.

Peak picking

PTM-Shepherd picks peaks based on a mixture of peak prominence and signal-to-noise remainder (SNR) as measures of quality and quantification, respectively. A peak's prominence is calculated as the ratio of its apex to the more intense of either its left or right shoulder, found by following a peak downward monotonically (Figure 1). To improve monotonicity for this procedure, adjacent histogram bins are temporarily grouped into small sets and flattened to the minimum bin height within the set, with set size internally calculated based on the total number of histogram bins. Peaks are called when their prominence exceeds 0.3 (by default). A peak's SNR is calculated with a 0.004 Da sliding window (by default) against a background of 0.005 Da on either side (scaling linearly with peak picking width). The average height per histogram bin is computed for the signal and noise regions, then the signal remainder is calculated by subtracting off the noise. From this list of peaks, the top 500 by SNR (by default) are sent to downstream processing. Peak boundaries are considered to be either the observed peak boundary or the defined precursor tolerance, whichever is closer to the apex. PSMs are assigned to the peak if their mass shift falls within the reported peak boundary.

Mass shift annotation

Detected peaks are iteratively annotated using entries from the Unimod (retrieved: 2 Oct 2019)[2] modification database (including single residue insertions and deletions and isotopic error peaks), supplemented with a user-specified list of mass shifts. Each peak is allowed to be decomposed into at most two modifications. Some exceptionally rare or protocol-based modifications (e.g., O18 labeling, N15 labeling) that regularly confounded annotation were removed. Mass differences within 0.01 Da (by default) of a known mass shift are annotated

immediately. If a mass shift does not meet this condition, it is then tested against combinations of user-defined mass shifts and known annotations before being checked against combinations of two modifications identified at the previous step above. Failing both of these assignments, mass differences are marked as "Unannotated" and appended to the list of potential modification combinations.

Mass shift localization

PTM-Shepherd constructs localization profiles for each mass shift peak. Localization profiles are constructed for each experiment, reporting an N-terminal localization rate and a normalized amino acid propensity for each peak. The localization step is performed for every PSM by placing the mass shift at each amino acid in turn and re-scoring the PSM (with the original spectrum) using the same scoring function as in MSFragger. PSMs are considered localizable if there is a position(s) within the peptide sequence that, when the mass shift is placed there, results in more matched fragment ions than using unshifted fragment ions only (i.e. without adding the mass shift anywhere). Localizable PSMs corresponding to the same peak in the mass shift histogram are aggregated, and their characteristics are analyzed. The localization rate for a peak is calculated by counting the number of instances a mass shift was localized to a particular amino acid. If the localization is ambiguous (i.e. several sites scored equally high), the weight of the localization is distributed among all localized residues. Counts are normalized to the rate of localization for a given residue, then divided by each residue's background content. Background residue content is computed by counting the number of occurrences of each residue in every localizable PSM in the entire dataset (by default). Options for experiment-level normalization at the unique peptide level, and bin-wise normalization at the PSM and unique peptide level are also available.

Modified-unmodified comparisons

Cosine spectral similarity between modified and unmodified peptides is used to determine how mass shifts affect MS/MS spectra. Unmodified PSMs, i.e. PSMs with a mass shift less than 0.001 Da (by default), are aggregated based on their identified peptide sequence and charge state. If

there are more than 50 unmodified spectra for a peptide, 50 are randomly selected for downstream comparisons. Then, for every mass shifted PSM at a given charge state, the average cosine similarity score between this PSM and its corresponding unmodified PSMs at the same charge state is recorded. These similarity scores are aggregated for all PSMs for each mass shift peak, then averaged and reported as that peak's spectral similarity profile. Retention time effects are also examined. Peptide retention times are extracted from Philosopher's PSM.tsv output. For every PSM with a mass shift, the average difference in retention time between that PSM and all its corresponding unmodified PSM is calculated. These average retention time differences are aggregated for each mass shift peak, then averaged across all peptides in that peak and reported as that peak's retention time difference profile.

2.2.2 Experimental datasets

Four formalin-fixed, paraffin-embedded (FFPE) datasets were used for identifying modifications associated with the fixing process and storage as selected by Tabb et al. for their study [78]. Two of these datasets, titled "Nielsen" (PXD000743) and "Buthelezi" (PXD013107), were acquired on SCIEX TripleTOFs [79]. The Nielsen dataset was acquired on a TripleTOF 5600+ and consists of 218449 scans across 20 SCIEX .wiff files, and the Buthelezi dataset was acquired on a TripleTOF 6600 and consists of 474726 scans across 12 SCIEX .wiff files. Two other datasets, titled "Zimmerman" (PXD001651) and "Nair" (PXD013528), were acquired on Thermo Q-Exactive instruments [80]. The Zimmerman dataset consists of 79803 scans across 5 .raw files, and the Nair dataset consists of 245589 scans across 10 .raw files. Files were acquired as .raw or .wiff files and converted to mzML using ProteoWizard's MSConvert version 3.0.18208.

The synthetic peptide dataset was obtained from ProteomeXchange (PXD004732) in mzML format [81]. Only MS runs with the 3xHCD label were included in our analysis. Peptide pools labeled as "SRM" were also excluded. This synthetic peptide dataset consists of unmodified proteotypic human peptides fragmented on a Thermo Fisher Orbitrap Fusion Lumos instrument. Cysteines were incorporated as alkylated cysteines during synthesis.

Additional datasets used in this work were obtained from the Clinical Proteomics Tumor Analysis Consortium (CPTAC) Data Portal in mzML format [82]. These were limited to MS runs generated from the CompRef samples, a CPTAC reference material created using breast cancer xenograft pools for quality control and data harmonization purposes. The samples were analyzed using TMT-10 labeling based technology. The first cohort of six experiments (TMT 10-plexes) consists of samples processed at three sites (2 experiments from each site) - the Broad Institute (BI), Johns Hopkins University (JHU), and Pacific Northwest National Laboratory (PNNL) - acquired on an Orbitrap Fusion Lumos as part of the CPTAC harmonization study [83]. The second cohort consists of the same CompRef samples processed as longitudinal QC samples as part of three CPTAC datasets: the Clear Cell Renal Cell Carcinoma (CCRCC) dataset generated at JHU (three experiments), the Lung Adenocarcinoma (LUAD) dataset generated at BI (four experiments), and the Uterine Corpus Endometrial Carcinoma (UCEC) dataset generated at PNNL (four experiments). All these data, at all sites, were generated using the CPTAC harmonized data generation protocol [83]. These data were processed together using PTM-Shepherd's multi-experiment setting to generate a single report.

2.2.3 Database search and statistical validation

All analysis was performed using a database constructed from all human entries in the UniProtKB protein database (retrieved 29 July 2016). Reversed protein sequences were added as decoys and common contaminants from were appended (total targets and decoys: 141,585). Unless specified otherwise, all datasets were processed with the following parameters. Data was searched using MSFragger v2.1[69] with a precursor mass tolerance of +/- 500 Da. Isotope error correction was disabled, and one missed tryptic cleavage was allowed for peptides of 7 to 50 residues in length. Oxidation of methionine was included as a variable modification and cysteine carbamidomethylation was included as a fixed modification. MSFragger mass calibration and parameter optimization was performed for all datasets [84], including fragment ion tolerance. Shifted ions were not used in scoring.

FFPE data was processed using a -200 to 500 Da mass range to match that used in the original

publication [78]. CPTAC data was processed with protein N-terminal acetylation and peptide N-terminal TMT mass of 229.1629 Da as variable modifications (TMT was also specified as fixed modification on Lys). In addition, CPTAC data was searched against combined UniProtKB mouse plus human protein database (retrieved 10 February 2020), with its respective reversed decoys appended to the database, resulting in 252,401 total target and decoy proteins.

PSMs identified using MSFragger were processed using PeptideProphet [85] via the Philosopher v2.0.0 toolkit [77]. All processing and filtering was performed on a per-experiment basis. The four FFPE datasets were processed as four experiments. The chloroacetamide-labeled HeLa cells dataset was processed as one experiment containing all 39 fractions. CPTAC samples were grouped experiment-wise, with each experiment containing all 24 fractions. Due to large size, the ProteomeTools synthetic peptide dataset was processed as 11 subsets split based on the five-digit identifier at the beginning of each filename. PeptideProphet parameters for all analyses were default open search parameters: semi-parametric modeling, clevel value set to -2, high accuracy mass mode disabled, masswidth of 1000, and using expectation value for modeling. Resulting PSM matches were filtered to 1% FDR using target-decoy strategy with the help of Philosopher filter command.

2.3 Results

2.3.1 Overview of PTM-Shepherd

The overview of PTM-Shepherd computational workflow is shown in Figure 2-1. The process starts with PTM-Shepherd reading FDR-filtered PSM lists (produced by MSFragger and Philosopher, optionally with open search artifacts removed using Crystal-C [86]), and the mass shift for each PSM, to construct a mass shift histogram [74]. After smoothing the histogram, PTM-Shepherd picks peaks based on a mixture of peak prominence and signal-to-noise ratio. From this list of detected peaks, the top 500 (by default) are selected for downstream processing. Basic abundance statistics are then calculated for this list of detected peaks. PSMs are assigned to a particular peak if their mass shift falls within the reported peak boundary, and abundance of the peak is calculated based on spectral counts. PTM-Shepherd can also operate in a multi-

experiment mode. In this mode, peak detection is performed on an aggregate mass shift

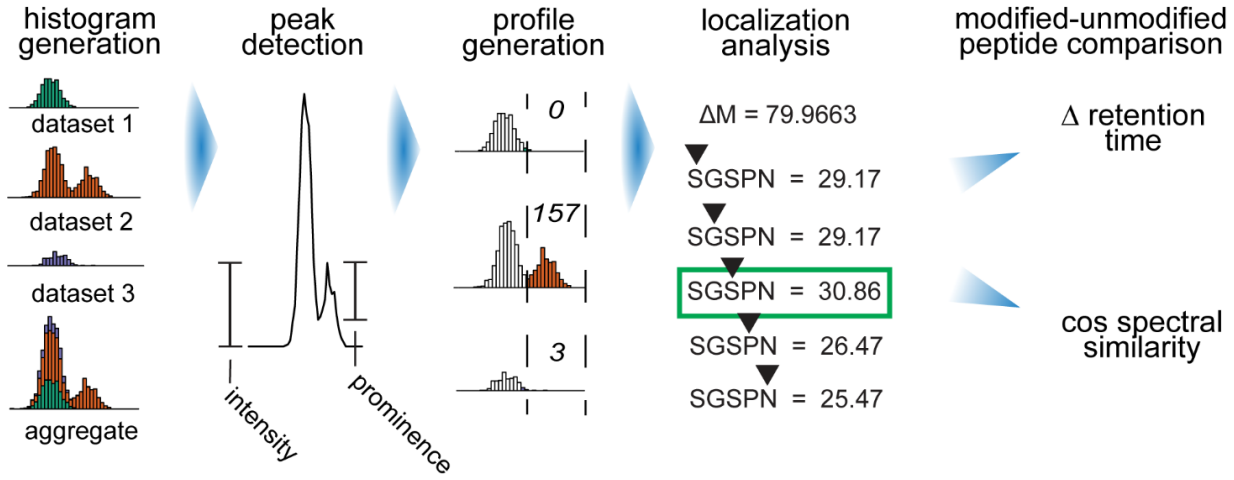

Figure 2-1: PTM-Shepherd workflow. Data processing begins by aggregating the mass shifts across all datasets into a common histogram. Peaks are determined based on their prominence. The 500 most intense peaks in aggregate are then quantified for each dataset and normalized to size. Peptides with each mass shift are iteratively rescored with the modification at each position, producing localization scores for each peptide and an aggregate localization enrichment for each mass shift. Finally, modified peptides and their unmodified counterparts are analyzed to have their pairwise cosine spectral similarity and change in retention time calculated.

histogram from all experiments, generated from the mass shifts of each experiment weighted according to the proportion of the total PSMs they comprise. The use of a combined histogram for peak detection can greatly simplify comparisons between modifications detected in different conditions and experiments. In this multi-experiment mode, the summary attributes for each detected peak are generated separately for each experiment, and for all data combined.

Once peaks in the mass shift histogram have been called, PTM-Shepherd attempts to determine their identities. Mass shifts are iteratively annotated using entries from the Unimod [2] modification database, isotopic error peaks, and user-specified mass shifts, allowing the mass difference to be decomposed into at most two modifications. PTM-Shepherd also constructs localization profiles for each peak. Localization profiles are constructed for each experiment, reporting an N-terminal localization rate and a normalized amino acid propensity for each modification. This analysis is performed for every PSM by placing the mass shift at each amino acid in turn and re-scoring the PSM (with the original spectrum) using the scoring function presented in MSFragger (see Methods).

PTM-Shepherd also computes several metrics that are useful for gaining a better understanding of the nature of those detected mass shifts. For each peak, PSMs containing that mass shift are compared to their unmodified counterparts if present within the same run. First, cosine spectral similarity between modified and unmodified peptides is computed, which is useful for determining how the modifications affect spectra. Then retention time effects are examined, and the average difference in retention time between the peptide with and without modification are reported.

2.3.2 PTM-palette discovery: analysis of FFPE samples

Understanding how sample processing and storage affect proteins is critical to maximizing their identification. Analysis of tissue samples preserved using formalin-fixing paraffin-embedding (FFPE) technique warrant the inclusion of additional modifications to reflect changes in proteins following formalin fixation. FFPE samples are also typically analyzed after long-term storage, during which they could be exposed to high temperatures and sunlight [87]. Although previous studies have examined which modifications should be included when analyzing proteins from FFPE samples [87], this was revisited recently by Tabb et al. [78] using a two-pass search. First, an open search was used to identify prevalent mass shifts. Second, they performed a traditional search and informed the localization of their mass shifts with chemical knowledge. We sought to investigate how PTM-Shepherd could be used to validate their findings and streamline this analysis for other datasets and sample preparation protocols.

After their first pass open search, Tabb et al. [78] found five modifications that were consistently present across the four datasets analyzed: methylation, di-methylation, single oxidation, double oxidation, and variable carbamidomethylation. Automated processing with PTM-Shepherd replicates most of these findings. Based on PSM counts and using the same criteria, we find mass shifts of methylation, mono-oxidation, and di-oxidation within the top 10 mass shifts (excluding isotopic error peaks) for every dataset (Table A-1A). Interestingly, PTM-Shepherd also finds a notable discrepancy with respect to di-methylation levels. PTM-Shepherd identifies two peaks in

close proximity: 27.9954 Da (corresponding to formylation) and 28.0320 (corresponding to di-methylation). Di-methylation is only higher than formylation in one dataset (Nielsen, Figure 2-2a), while the others have formylation between three- and nine-fold higher than di-methylation. To confirm that this was not an artefact of PTM-Shepherd's signal-to-noise peak picking, we reanalyzed these results with the DeltaMass software that implements an alternative (Gaussian mixture modeling) strategy for peak picking [74]. For all these four datasets, DeltaMass found that the region of mass shifts from 27.90 to 28.10 contained two peaks (Figure I-1). For Nielsen, Nair, and Zimmerman, these are easily visible. Even the Buthelezi dataset, while not exhibiting as clear of a separation as the others, places the more abundant peak apex closer to the mass shift value corresponding to formylation. The presence of formylation within a list of most abundant PTMs also makes logical sense given the nature of preservation method.

Tabb and colleagues relied on chemical knowledge and other tools [42,72] to arrive at the final search configuration that included oxidation of Met to methionine sulfone. We chose to investigate this further using PTM-Shepherd. Because a single oxidation of Met was already included as a variable modification in our open search, a Met oxidation to methionine sulfone may appear as either a variable modification and a +15.9949 Da mass shift localized to Met or a +31.9898 Da mass shift localized to Met. However, we do not observe enrichment of Met (Table A-1A) localization in either of these instances. In contrast, the enrichment scores for Pro were 9.3 and 5.6 for mono- and di-oxidation, respectively. Tabb and colleagues' gain in the number of PSMs when adding Met sulfone and dihydroxy Pro in the search may be explained by the diffuse nature of oxidation. When using a closed search strategy with a dynamic +31.9898 Da modification on Met, any occurrence of two +15.9949 Da events - for example on two alternative oxidation sites - might be interpreted as a Met sulfone because peptide ions will converge downstream of the theoretical and experimental oxidation sites. The same phenomenon can occur with multiple instances of hydroxyproline. Collagen is known to contain massive amounts of hydroxyproline, and as such is likely to produce peptides with multiple hydroxyprolines [88]. To determine whether the +31.9898 Da mass shift was attributable to multiple hydroxyprolines co-occurring on the same peptide, we checked whether PSMs containing it were more likely to map to collagen proteins than non-

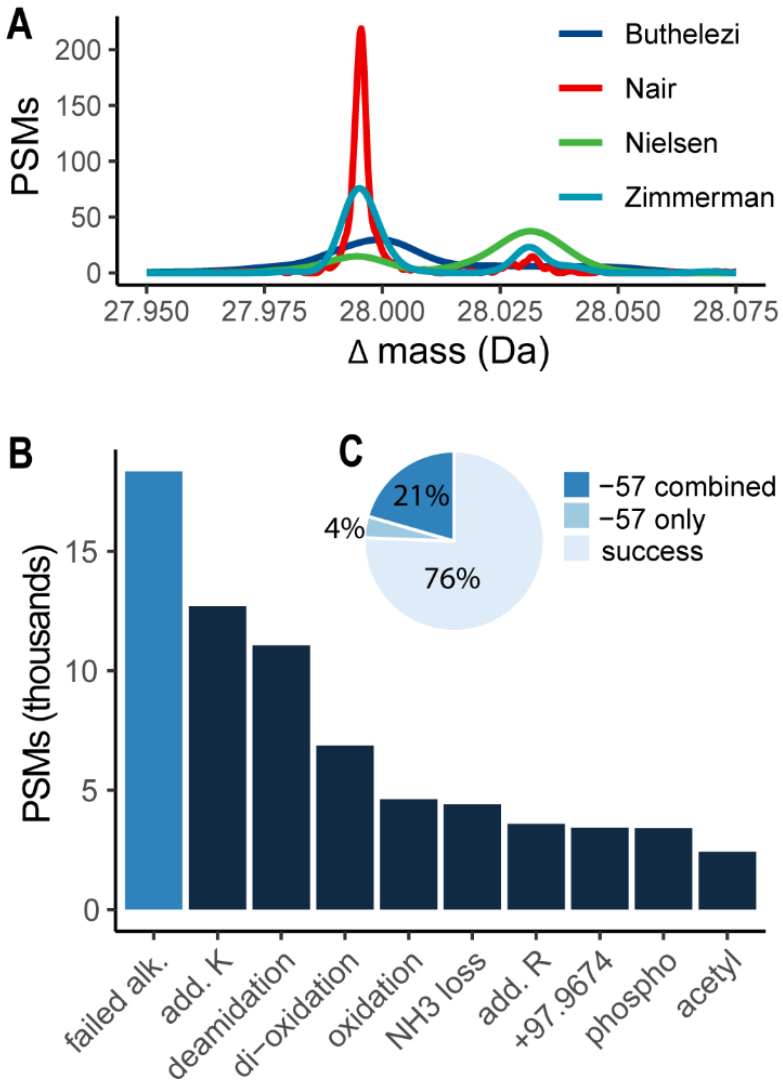

Figure 2-2: Basic PTM-Shepherd applications. A: PTM-Shepherd identifies two peaks in close proximity for Tabb et al.'s four datasets. All four datasets (Zimmerman, Nair, Nielsen, and Buthelezi) show a mixture of two Gaussian peaks about 28 Da. The consistently more intense peak is at 27.9949, formylation. Only in the Nielsen dataset does di-methylation (28.0313) approach formylation's intensity. B: PTM-Shepherd identifies more failed alkylation than other common modifications such as deamidation and not-Met oxidation. C: PTM-Shepherd modification decomposition identifies six times as much failed alkylation as

collagen proteins. This mass shift was overwhelmingly more likely to occur on collagen proteins (OR = 43.5, $p < 10^{-5}$ by Fisher's exact test), confirming that it is a combination of multiple hydroxyprolines and can be captured via including hydroxyproline in a PTM palette. Overall, in

our experience in this and other datasets, PTM-Shepherd provides a very reasonable estimate of the most likely modification sites for a particular mass shift.

Formaldehyde adduct (+30.0106 Da, annotated as methylol in Unimod) is a known modification observed on peptides from FFPE samples, and it was detected at high levels in the Nair and Zimmerman datasets (in the top 10). According to PTM-Shepherd analysis, this mass shift lacks any significant localization characteristics (localized less than 10% of the time), indicative of a non-covalent adduct. In general, identification of labile modifications is one of the advantages of open database searching with MSFragger compared to other PTM-focused tools or closed searches with variable modifications, all of which are less effective at finding labile modifications that cannot be localized. Summarizing our observations, PTM-Shepherd suggests a slightly modified version of the PTM palette from FFPE samples proposed by Tabb et al.: oxidation of Met (+15.9949 Da), hydroxyproline (+15.9949 Da on Pro; ideally, specified for collagens only), formylation on Lys and N-termini (+27.9949 Da), and methylation of Lys and N-termini (+14.0157 Da). It may also be beneficial to include the methylol (+30.0106) adduct, but using the mass offset search option of MSFragger that allows both shifted and non-shifted fragment ions in scoring rather than as variable modification [45,84,89].

2.3.3 Detection of Cysteine artefacts following underalkylation

Cysteine is an extremely reactive amino acid, frequently picking up an array of chemical modifications when exposed [90]. Unspecified mass shifts, such as those resulting from chemical modifications of Cys, confound peptide identifications and lead to lower recovery rates. Cys alkylation restricts the number of chemical derivatives it can form and prevents interference from disulfide bonds, and as such has been a mainstay of proteomics processing for decades [91]. Chloroacetamide (CAA) and iodoacetamide (IAA) are the two most common alkylating reagents used in proteomics workflows. Previous comparisons of these reagents found that IAA generally has a higher rate of cysteine alkylation than CAA when applied at the same concentration, but with the caveat of higher rates of off-target effects [92]. Here, we have tested the ability of PTM-Shepherd to uncover cysteine artefacts in proteomic datasets. Bekker-Jensen et al. [93] rigorously

tested shotgun proteomics protocols to determine an optimal strategy for rapidly generating a comprehensive profile of human proteomes, ultimately producing a valuable repository of high-quality, deep proteomics data. Their protocol also included a 10 mM treatment with the alkylating agent CAA, which, per Schnatbaum et al. [92], only achieved two-thirds the alkylating efficiency of 10 mM IAA in complex mixtures. Unlike the other samples we analyzed for this manuscript, this protocol also did not denature protein samples before adding the alkylating agent. This likely contributes heavily to underalkylation as well. As such, it presents an exceptional opportunity to examine these cysteine artefacts.

Open search analysis followed by PTM-Shepherd shows a number of prevalent mass shifts enriched on Cys, consistent with what we expect from underalkylated samples. Note that because Cys alkylation was searched as a fixed modification, in order to elucidate the identity of modifications occurring on unalkylated Cys residues the mass shift must be decomposed into two components: a failed alkylation event (Δm = -57.0215 Da from the theoretical mass of the identified peptide) and the modification itself. Consider the mass shift -9.03680 Da detected in this dataset. PTM-Shepherd decomposes this mass shift into a failed alkylation event (-57.0215 Da) and a triple oxidation of Cys to cysteic acid (Δm = +47.9847 Da). This becomes particularly important when trying to directly assess the number of failed alkylations in a sample. Strictly counting the number of -57.0215 Da mass shifts will severely under count their total occurrences because it ignores cases where it is found in conjunction with another modification, which are very likely given Cys's reactivity. We implemented an additional parameter in PTM-Shepherd to account for this that prioritizes user-defined modifications and allows them to identify mass shifts that do not directly correspond to entries in Unimod [2]. On its own, failed alkylation was the sixth most abundant mass shift and was more prevalent than other common events that are often accounted for in closed searches, such as pyroglutamate formation, and accounted for 3.9% (3450 PSMs) of the 89281 total Cys-containing PSMs (Table A-2A). However, because it frequently occurs with other mass shifts as demonstrated above, we also pooled the instances in which it was annotated as one of two mass shifts on a peptide. Remarkably, the total number of failed alkylation events jumps to 20.5% (18343 PSMs) when considering all instances of failed

alkylation annotations (Figure 2-2c). When considering decomposed mass shifts, failed alkylation is nearly twice as common as deamidation and four times as common as non-Met oxidation events (Figure 2-2b, Table A-2B).

After applying an abundance cutoff of 0.01% of total spectra, we detected 10 mass shifts exhibiting strong Cys localization (>10-fold enrichment) that were also annotated with a failed alkylation of Cys. These were a large portion of the broadly occurring Cys-enriched PTMs in the samples (Table A-2A). The most abundant of these modifications correlate with what would be expected in a poorly alkylated sample. The +1 and +2 isotopic error peaks in conjunction with unalkylated Cys were particularly abundant, accounting for 1601 combined spectra. The aforementioned -9.0368 Da - triple oxidation on Cys without alkylation - was most prevalent aside from these. Its heavily enriched localization to Cys (48.5-fold) lends credence to this compound identification that would be missed by other annotation tools and, consequently, a count of total failed alkylation events. Surprisingly, failed alkylation combined with a "formaldehyde adduct" (+12.0000 Da) was also common. The combined mass shift of -45.0216 Da was heavily localized to Cys and had a 96% N-terminal localization rate, pointing to potential thiazoladine formation via N-terminal Cys cyclization. These are known to occur in formaldehyde treated data, however the authors did not report the use of formaldehyde [94]. A lone "formaldehyde adduct" mass shift accounting for 305 PSMs and heavily localized to Trp (42.5-fold enrichment) was also detected in the dataset, however. Taken together, these indicate that the thiazoladine was probably an artifact of formaldehyde exposure rather than underalkylation, though the latter may be required for there to be available Cys to react with formaldehyde. Failed alkylation of Cys and the subsequent triple and double oxidations conform to our chemical knowledge of Cys artefacts and, along with glutathione disulfide as a biological modification, comprise 8.7% of all Cys-containing PSMs. Including these modifications should increase Cys-peptide recovery in underalkylated samples.

2.3.4 PTM-Shepherd computed metrics facilitate granular PTM identification

Open searches are inherently limited in the information they provide, providing only peptide lists and their associated mass shifts [69]. Data interpretation efforts are further complicated by the ambiguity of mass shifts. Two methylation events and an ethylation event, for instance, would be indistinguishable from each other based on mass. However, more granular identities can be discerned by incorporating additional metrics: changes in retention time (RT), spectral similarity (SS), and localization. To demonstrate that these additional metrics improve open search result comprehension, we analyzed the synthetic unmodified tryptic peptide dataset generated as part of the ProteomeTools project [81]. This dataset allows us to examine and characterize instrumental artefacts apart from confounding biological factors.

In-source losses from peptides result from low-energy fragmentation pathways that can occur during tandem MS as well as during ionization and transmission, resulting in artefactual changes to the observed precursor mass [95]. Because in-source losses occur after column elution and consequent retention time recording, they have no effect on peptide retention time. This property can be used to distinguish them from sample modifications [96]. We used PTM-Shepherd to elucidate the origins of two of this dataset's most common mass shifts attributable to both in-source losses and real modifications: loss of H_2O and loss of NH_3 (Table A-3). Peptides with multiple spectra corresponding to each loss had their RT shifts pooled and collapsed to their median. Interestingly, both losses of H2O and NH3 exhibited bimodal changes in retention time (Figure 2-3).

For peptides presenting losses of H2O (Δm = -18.0104 Da; 2961 peptides), both composite RT distributions were approximately Gaussian with approximate means of 0 and 450 s. As anticipated, many peptides (16.7%) fall within the mean 0 distribution, indicating that they do not experience increases in column RT despite the loss of a highly polar group. This is characteristic of in-source losses and indicates that peptides within this distribution are exhibiting in-source loss of H2O in the mass spectrometer prior to precursor selection. Peptides presenting losses of NH3 (Δm = -17.0270 Da, n = 1094) showed a similar pattern. Note that H2O and NH3 are only two examples of in-source loss and, in some cases, entire residues can be lost via this

mechanism. As a significant source of instrumental bias, it is important to be able to classify in-source losses properly and remove them from experimental sample pools. In fact, for researchers studying the isobaric biological forms of these mass shifts, it is critical to exclude these.

The second population of peptides with H_2O and NH_3 loss exhibited a RT-shift consistent with a

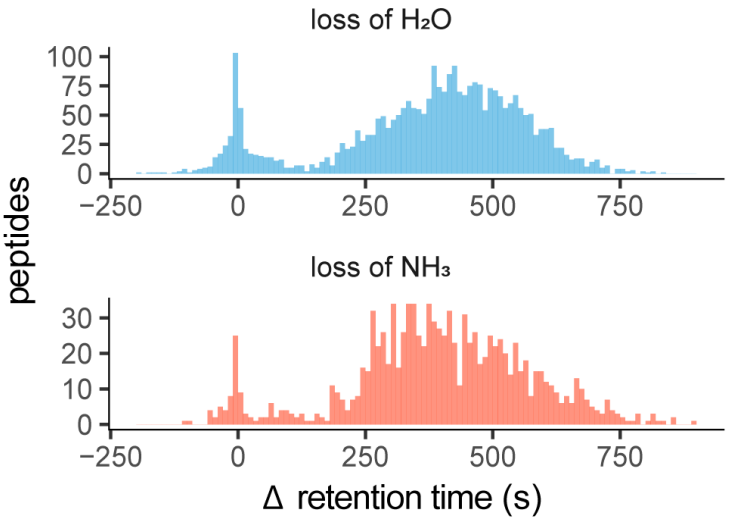

Figure 2-3: Retention time profiles for peptides with losses of H_2O and NH_3. Modified peptides are compared to their homologous unmodified peptides, with multiple retention time changes being collapsed to their median. The effect on retention time for losses of H_2O (top) and NH_3 (bottom) are distributed bimodally. These mass shifts are known to correspond to both in-source losses and spontaneous conversions. In-source losses should not have an effect on retention time, and as such are suspected to fall within a Gaussian distribution centered at zero.

pre-elution modification - the loss of a polar group increased retention time by an average of 450 s. H_2O losses are known to manifest as a conversion to pyroglutamic acid from Glu [97] as well as on Asn, Gln, Ser, Thr, Tyr, Asp, and Cys as sample-derived modifications [2]. NH_3 losses are known to manifest as a conversion to pyroglutamic acid, but from Gln rather than Glu [98]. Other losses of NH_3 are known to occur on some N-termini occupied by Thr, Ser, and Cys and on any Asn [2]. While RT shifts alone do not contain enough information to fully identify these modifications, additional metrics calculated by PTM-Shepherd - localization propensity and modified-to-unmodified peptide similarity - allowed us to delineate the primary sources of this mass shifts.

In the case of a loss of NH_3, two primary sources (in addition to the in-source losses) were identified: a spontaneous conversion of Gln to pyroglutamate, and a cyclization of Cys. Cys cyclization is expected to be present in peptides being synthesized with carbamidomethylated

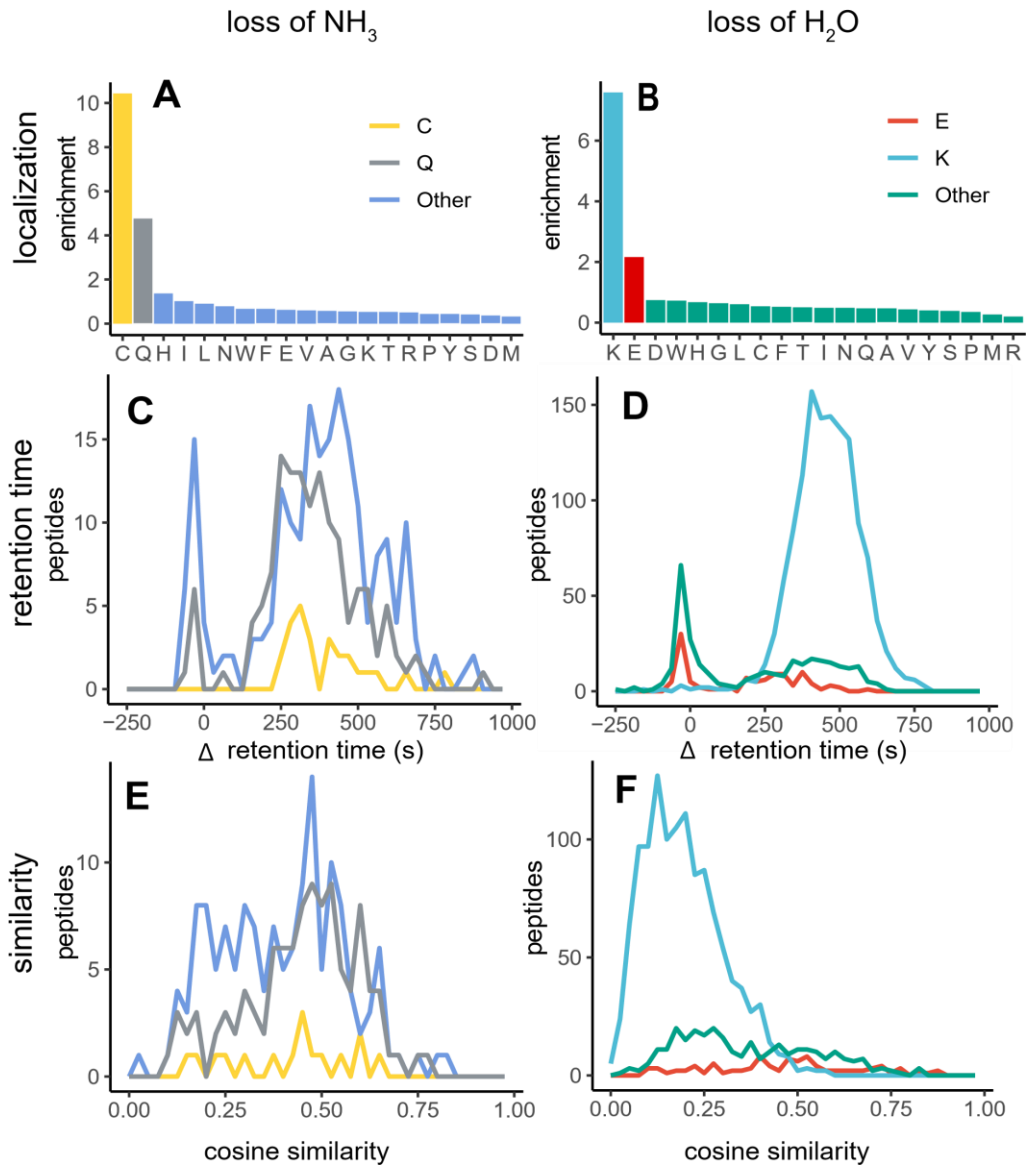

Figure 2-4: Analytical profiles for losses of H2O and NH3. A,B: Localization profiles reveal a non homogeneous landscape with specific residues showing enrichment. C,D: Select modifications are distinguishable from background in-source decays in their effect on retention time. E,F: Similarity scores show lower profiles for C-terminal modifications on Lys, whereas N-terminal modifications on Glu, Cys, and Gln have higher similarity.

Cys, as the reaction is known to occur after alkylation [99]. Localization analysis showed that Cys (enrichment score = 10.4) and Gln (enrichment score = 4.7) were the two most enriched residues for this mass shift (Figure 2-4a). In-source losses localized to Cys are rare and in-source losses localized to Gln are relatively common [100], which is reflected in the number of peptides each residues produces with ΔRT = 0. To illustrate, Cys has an RT profile very different from other residues in aggregate, while Gln has a similar distribution (Figure 2-4c); none of those containing Cys localized NH_3 losses were in-source losses, as compared to 19.6% of other residues in aggregate.

Unlike NH_3, we expected losses of H_2O to only be heavily enriched in glutamate. Glutamate is known to have two sources for loss of H_2O; it is known to be a source of water in-source loss but can also spontaneously undergo N-terminal cyclization *in vitro* to produce the same mass shift [101]. The localization enrichment profile for loss of H_2O (Figure 2-4b), however, revealed that two residues were exceptional contributors to this PTM's prevalence: Glu (enrichment score = 2.2) and Lys (enrichment score = 7.6). We identified populations of peptides with losses of H_2O corresponding to both of these populations based on ΔRT as described above (Figure 2-4b, red). An intense peak near ΔRT = 0s indicates that many unique peptides are capable of producing in-source losses of H_2O (22.8%) on Glu, consistent with the conclusions of Sun et al. [100] Another peak near ΔRT = 300s indicates that the remainder of the peptides had a loss of H_2O that was present before column elution, and as such is likely to be an N-terminal cyclization of glutamate occurring in vitro.

Even more so than glutamate, lysine was the largest contributor to losses of H_2O (example spectrum at Figure I-2). Puzzlingly, lysine's side chain does not have a hydroxyl group that it can readily lose, and as such any H_2O losses attributable to lysine must be derived from the C-terminal hydroxyl group of tryptic peptides. It is worth noting that this phenomenon is unique to lysine, as Arg had the lowest localization enrichment score (0.2) of all 20 residues (Figure 2-4b). Based on retention times, there were also no appreciable H_2O in-source losses attributable to lysine - consistent with previous findings [100] - indicating that these lysines were being dehydrated

prior to elution (Figure 2-4d). We believe this is most likely due to a C-terminal lysine cyclization event. Though undescribed in proteomics, lysine derivative cyclization has been induced in other settings[102]. This theory is supported by spectral similarity calculations between peptides with and without this lysine-localized mass shift (Figure 2-4f). There is exceptionally low spectral similarity between the spectra of modified and unmodified peptides. Peptides containing non-labile modifications and modifications near the C-terminus both result in low spectral similarity; non-labile modifications are likely to be retained in the MS/MS spectra rather than being removed during MS1 analysis, and C-terminal modifications shift the intense y-ion series. A covalently bound lysine ring structure on peptide C-termini fits both of these criteria and may be the underlying cause of low spectral similarity**.**

The similarity profiles for losses on Glu, Cys, and Gln were distinct from Lys in that they were not enriched for peptides with MS/MS showing low similarity to their unmodified counterparts. This may be accounted for by the fact that Glu cyclization, like Cys and Gln, occurs at the N-terminus; shifting the b-ion series has less of an effect on the MS/MS spectra than shifting the y-ion series. Unsurprisingly, all three of these have similarity profiles roughly corresponding to the proportion of their spectra experiencing *in vitro* modifications, which are the only modifications we can be sure are occurring at the N-terminus.

Overall, by including metrics beyond mass shifts in PTM identification, we show that much more information beyond the chemical composition of a mass shift can be deduced. Retention times can be used to discriminate between in-source losses and sample modifications, localization profiles can be used to deduce biological or chemical origins, and spectral similarity provides additional localization and lability metrics. Incorporating all of these, we found, to our knowledge, a previously unknown or at least underappreciated modification (C-terminal lysine cyclization) in a deeply consequential synthetic peptide library.

2.3.5 PTM-Shepherd in multi-experiment settings

PTM-Shepherd can be run in a multi-experiment mode to analyze modification profile across a

large number of experiments. Such an analysis could be performed for visualization of interesting biological trends and in search of experiment-specific biological modifications. It could also be useful for quality control and detection of batch effects – a common source of variation in high throughput data [103]. Previous efforts have been made to identify MS performance metrics [104], and some groups have shown how these can be leveraged to identify quality control issues [105] and better understand intra- and inter-laboratory variability [106,107]. We posited that open-search derived modification profiles could be used to determine interexperiment variation while simultaneously providing insight into its origins.

To evaluate PTM profiling in multi-experiment settings we used CPTAC CompRef reference material data (pooled tumor xenografts comprising ten samples each from two different breast cancer subtypes, cryopulverized and shipped to different processing locations) obtained from the CPTAC Data Portal (see **Methods**). The samples were processed at three different locations (PNNL, JHU, BI), and analyzed using TMT 10-plex labeling technology as part of the CPTAC3 Harmonization study [83]. The same CompRef samples were also analyzed as longitudinal QC samples as part of the three large cancer profiling studies, CCRCC [108] (MS data collected at JHU; UCEC [109] (MS data collected at PNNL), and LUAD [110] (MS data collected at BI).

We first investigated the most abundant mass shifts identified by MSFragger and PTM-Shepherd in these data (Table 2-1, see Table A-4A for the full list), which revealed several interesting observations. In general, PTM-Shepherd accurately reconstructed expected trends including the localization profiles of the most abundant modifications. For example, carbamylation and formylation were most highly enriched on N-terminus, phosphorylation on Ser, and oxidation on Trp. Not considering isotope errors, the mass shift of 229.1629 Da (TMT overlabeling) was the most common modification, localized predominantly to Ser (enrichment factor of 5.6). Of note, only 74.5% percent of peptides found with TMT on Ser were also found in “unmodified” form (i.e., with unlabeled Ser). In contrast, many other abundant modifications, such as formylation

Table 2-1: Top mass shifts from CPTAC quality control samples. Assigned modifications correspond to automated Unimod matches, with * indicating a partially manually reannotated mass shift. The "% in Unmodified" column corresponds to the percent of PSMs with a matching unmodified PSM in the unmodified bin. Top two enriched amino acid localizations are shown in columns denoted "AA."

Mass Shift	PSMs	Assigned Modification	% in Unmodified	Similarity	Delta RT	N-term rate (%)	AA_1	AA_2
0.000	4372080	None	100.00	0.98	0	0		
1.002	913405	+1 isotopic error	79.43	0.81	5	3		
229.163	312295	TMT	74.52	0.38	-70	6	S (5.6)	T (2.4)
2.005	177121	+2 isotopic error	75.95	0.74	4	1		
-0.984 (15.011*)	156315	NH addition to M *	90.25	0.71	-173	1	M (15.8)	
230.166	141498	+1 isotopic error + TMT	72.51	0.42	-153	6	S (3.8)	
0.017	117922	+1 isotopic error + NH addition to M *	69.97	0.78	1	1	M (5.8)	
0.984	110967	Deamidation	68.84	0.68	53	6	N (12.7)	R (5.3)
17.026	92862	Deuterated methyl ester	94.96	0.76	-50	1		
15.011	87013	Conversion of carboxylic acid to hydroxamic acid	95.22	0.54	-434	4	E (5.5)	D (3.5)
-1.002	72465	-1 isotopic error/ +1 isotopic error + Half of a disulfide bridge *	81.40	0.78	16	1	C (4.8)	W (4.5)
15.995	59846	Oxidation	91.19	0.49	-342	21	W (24.0)	M (11.7)
100.016	59070	Succinic anhydride labeling reagent	95.76	0.55	596	77	P (3.6)	S (2.1)
27.995	52323	Formylation	93.22	0.61	265	61	S (2.7)	
79.967	48379	Phosphorylation	55.30	0.64	569	4	S (6.2)	
21.981	47471	Sodium adduct	98.72	0.25	-148	2		
115.027	47022	Cleavage product of EGS protein crosslinks by hydroxylamine	97.81	0.58	610	86	H (2.5)	
43.010	45450	Carbamylation	97.10	0.64	640	70		
-18.010	45445	Dehydration/Pyro-glu from E	97.48	0.64	-175	12	D (3.2)	T (2.5)
1.987	42988	+1 isotopic error +	69.08	0.66	45	4	N (11.9)	R (3.0)

and carbamylation, were found in both modified and unmodified forms in almost all cases. Interestingly, the second most abundant modification was a mass shift of 15.0107, predominantly

localized to Met (that was indistinguishable in MSFragger output from a combination of oxidation and -0.9842 Da loss on Met). This mass shift may represent the addition of an NH group to Met due to exposure to hydroxylamine, a reagent used in TMT labeling. At present, UniMod database annotates a 15.0107 Da mass shift only as conversion of carboxylic acid to hydroxamic acid, with Asp and Glu as only possible sites (which were observed in these data, but at a lower frequency than on Met, see Table 2-1).

The PTM profiles resulting from PTM-Shepherd analysis of these data are presented in Figure 2-5. Sample-wise K-means clustering revealed distinct sample clusters, and mass shift-wise clustering on correlation between columns (transposed PTM-Shepherd output) revealed some highly similar modifications. Sample clustering precisely reconstitutes sample processing location . For example, Cluster 4 in Figure 2-5 shows a series of mass shifts related to TMT overlabeling, or TMT labeling that was not captured by fixed sequence expansion on Lys and dynamic sequence expansion on peptide N-termini. PNNL data consistently shows lower TMT overlabeling than BI and JHU for every mass shift in this cluster, and PSMs corresponding to a single additional TMT are 5-8 times lower than at the other two locations. BI and JHU also show enrichments of TMT labeling on Ser and, to lesser degree, Thr.

Though we expect to see differences in TMT labeling fidelity, PTM-Shepherd also allows us to explore unexpected batch effects. Lenčo *et al*. recently raised concerns about the use of formic acid in sample preparation, specifically stating that an excess of Ser, Thr, and N-terminal formylation events are present in samples reconstituted with it [111]. Within the CPTAC harmonization study, the localization profile did exactly match that described by Lenčo and colleagues: Ser enrichment of 2.7, Thr enrichment of 1.9, and a 61% potential N-terminal rate (Table A-4A). Interestingly, this formylation peak appears to disproportionately high in the first of the two BI replicates from the harmonization studies (Figure 2-5, Cluster 5). BI01 replicate

had formylation 10- to 20-fold higher than JHU01 or PNNL01 and was even 4-fold higher than BI02. Ser and Thr also exhibit inflated formylation localization for this sample, consistent with Lenčo *et al.*'s results (Table A-4B,C**)**. Overall, a deeper analysis may be warranted in future

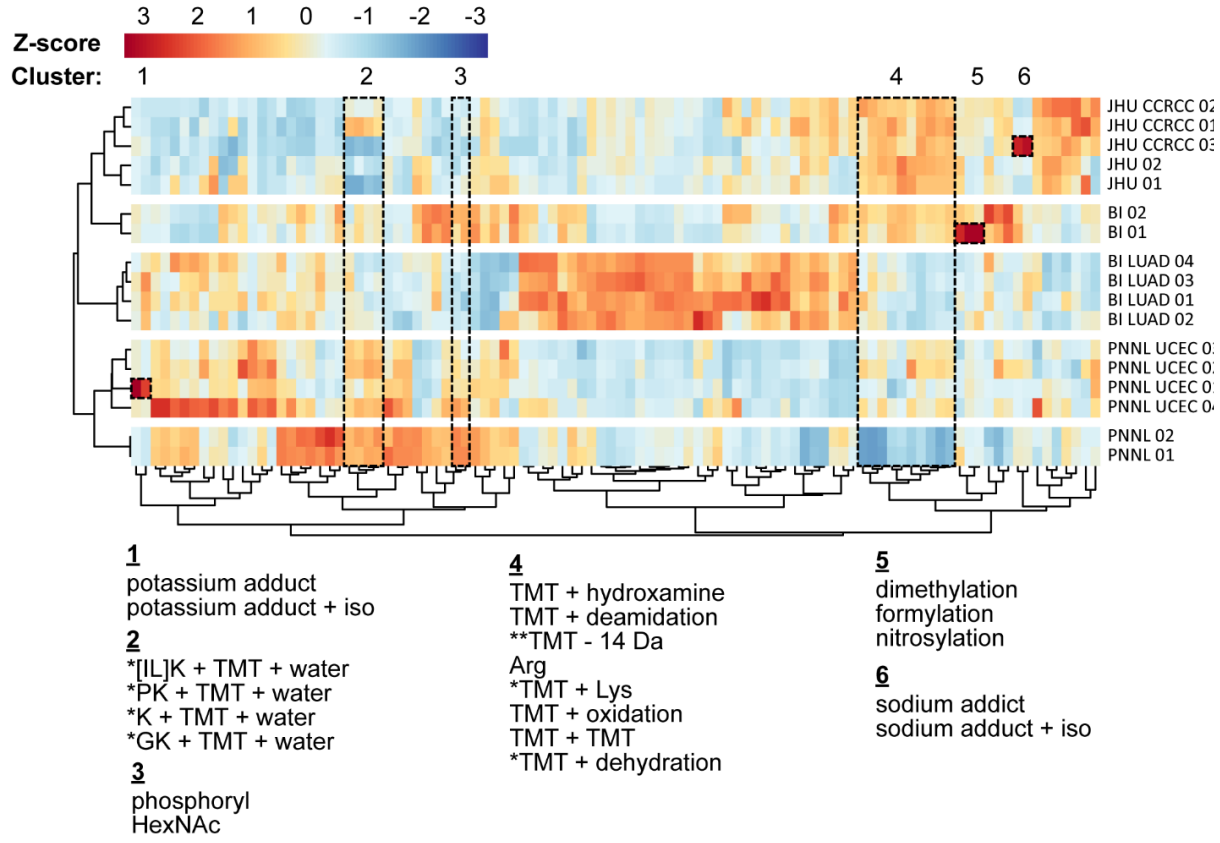

Figure 2-5: Clustered heatmap representation of CPTAC3 quality control samples transposed from PTM-Shepherd output. Values shown are column-wise z-scores of spectral counts. Column clustering shows highly related modifications, and row clustering shows experiments clustering by processing location. Mass shift clusters discussed in the text are numbered, and their corresponding mass shifts are shown left-to-right in the bottom of the figure. Samples processed longitudinally throughout their respective studies are indicated using tumor type label (LUAD, UCEC, or CCRCC). Samples with no tumor type label were processed as part of the CPTAC harmonization study. Mass shifts in Cluster 2 corresponds to negative mass shifts. *This annotation was constructed manually. **-14 Da can correspond to a large number of modifications and single residue mutations.

studies to reduce batch effects caused by formic acid use. These analyses could be extended to other sample handling artifacts, e.g. potassium adducts (Cluster 1) and sodium adducts (Cluster

6), which also exhibit marked longitudinal variability.

PTM-Shepherd also reveals how instrument parameters might be playing a role in lab-specific batch effects. We noted four large, negative mass shifts that were differentially identified across labs (Figure 2-5, cluster 2). Retention time profiling showed that these mass shifts were likely to be in-source losses (Figure 2-3). Their profiles were similar to loss of ammonia, a known in-source loss, and dissimilar to formylation, a known pre-elution modification (Figure I-3). Sequence analysis facilitated manual decomposition, revealing that three of the mass shifts were composed of a hydrophobic residue (Iso, Leu, Pro, or Gly), a C-terminal Lys, a TMT tag, and a water. The final mass shift was the loss of a C-terminal Lys, a TMT tag, and a water. It is possible that slight differences in fragmentation or ionization energies between labs may be manifesting as disparities in precursor charge states and proton mobility, however determining the mechanism through which these are occurring is outside the scope of this work.

Aside from analyzing how samples cluster together, there is also useful information to be gleaned from analyzing how mass shifts cluster together. We expect that related modifications should have highly correlated abundances between experiments, with co-clustering of isotopic error peaks as the most obvious example (e.g., potassium and sodium adducts and their related +1 isotopic error peaks, Figure 2-5, Clusters 1 and 6). Using a *Z*-score normalization to coerce PTMs derived from related sources across experiments to cluster together and clustering on the correlation between columns, we were able to identify some unknown mass shifts based on their co-clustering with other (known) modifications. In the TMT-related cluster noted above (Figure 2-5, Cluster 4), we observed three mass shifts that were missed by automatic annotation. Of the six that were annotated, five of the mass shifts are directly attributable to TMT overlabeling, and one to a missed tryptic cleavage or additional Arg at one end of the peptide sequence. This knowledge allowed us to explain one of the unannotated mass shifts as combinations of missed cleavages and TMT labeling. One modification (+357.2584 Da) is precisely the mass of a TMT-labeled Lys residue. This was missed by automatic annotation because both "Addition of Lys" and "TMT10 Plex" would have to be more abundant than the combination of the two during

PTM-Shepherd analysis. The other unexplained mass shift (+213.1680 Da) can be explained by a combination of TMT overlabeling and a dehydration event or a misattributed Met oxidation included as a variable modification. Overall, this analysis demonstrates the utility of multi-experimental analysis with PTM-Shepherd to better identify the mass shifts are not amenable to automatic annotation.

Finally, while our analysis above focused mostly on modifications introduced due to labeling and other sample handling steps, clustering of mass shifts may also be useful for uncovering correlated biological modifications. Of note, clustering of PTM-Shepherd results shows that phosphorylation (+79.9663 Da) is most correlated with HexNAc (+203.0794 Da) (Figure 2-5, Cluster 3). Interestingly, co-enrichment of glycopeptides was recently observed in datasets experimentally enriched for phosphopeptides [112], however, no phosphopeptide enrichment steps were applied to generate the data used in this work.

2.4 Discussion

Despite advancements in computational proteomics, many MS/MS spectra remain unexplained. Open searching with tools like MSFragger has proven to be an effective way to overcome the limitations of traditional database searches by removing the requirement of having prior knowledge of the peptide modifications present in the sample. The modifications elucidated by open searches, however, lack many of the metrics necessary to make proper determinations about their identities and origins. We addressed these challenges in PTM-Shepherd, which produces comprehensive PTM profiles for open search derived mass shifts, including multiple Unimod annotations, retention time changes, spectral similarity, and localization profiles.

We demonstrated the utility of PTM-Shepherd in four examples, providing a broadly applicable guide for others interested in utilizing open searches for PTM analysis in their own research. First, in the development of an FFPE-treatment PTM palette, we showed how PTM-Shepherd disambiguated two overlapping peaks: formylation and demethylation. We also demonstrated how PTM palettes can be easily constructed for other sample preparation methods without

extensive postprocessing. Second, we showed how PTM-Shepherd's unique ability to decompose mass shifts into multiple Unimod modifications allows us to identify and quantify the degree of failed alkylation, though this is easily extensible to other scenarios, e.g., identifying mass shifts corresponding to an absent variable modification and another co-occurring modification. We also demonstrated how incorporating additional metrics into PTM identification provides researchers with more granular, high confidence PTM identities, including the ability to distinguish between sample-derived and instrument-derived artefacts. Finally, when applied to data from a large, multi-center proteomics study, PTM-Shepherd helped us to visualize batch effects and the effect of sample processing location, as well as elucidate the identities of unannotated mass shifts. We believe PTM-Shepherd will become a widely used component in our MSFragger-based pipeline for comprehensive analysis of post-translational and chemical modifications, including searches for rare and even novel modifications, across a wide range of biological applications.

2.5 Data and software availability

All raw mass spectrometry data used in the manuscript can be found from the ProteomeXchange Consortium via the PRIDE partner repository, or (CPTAC data) from the CPTAC Data Portal, using specific dataset identifiers cited in the text. PSM lists can be accessed at 10.5281/zenodo.4042962. PTM-Shepherd is available as a standalone JAR executable (https://github.com/Nesvilab/PTM-Shepherd) and also fully integrated into the FragPipe Graphical User Interface (http://fragpipe.nesvilab.org/).

2.6 Acknowledgements

The authors would like to thank the users of our tools for their feedback. This work was funded in part by NIH grants R01-GM-094231 and U24-CA210967. D.J.G. was supported in part by the Proteogenomics of Cancer Training Program (T32-CA140044). A.I.N. and A.T.K. conceived the project, A.T.K. developed the first version of the software, later extended and improved by D.J.G., D.J.G. performed all analyses, D.M.A., F.Y., and F.L. contributed to the software development, D.J.G and A.I.N. wrote the manuscript with input from all authors, and A.I.N. supervised the entire project.

CHAPTER III

Mining for Ions:
Diagnostic Feature Detection in MS/M Spectra of
Post-translationally Modified Peptides

3.1 Introduction

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 tremendously 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 herculean 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 [13], but newer tools have incorporated additional features. Unfortunately, they are designed to work with a narrow subset of chemoproteomics probes, requiring isotopic labeling signatures to be present at the MS1 and

MS2 levels [17]. This limits their applications to chemical probes that are labeled non-isobarically, thus they cannot be used for some PTM probes [18], biological PTMs, or the development of isobaric mass tags [19]. Other approaches to score PSMs from modified peptides are trained for specific PTMs [20] or perform model refinement that focuses on distances between experimental peaks, ignoring diagnostic ions and discarding information about matched ions from the peptide backbone [21].

We present a novel 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 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.

3.2 Methods

3.2.1 Diagnostic feature detection algorithm

Spectral feature calculation

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 $\boldsymbol{m}$ ions, where each $\boldsymbol{ion_i}$ has a corresponding $\boldsymbol{mz_i}$ and $\boldsymbol{int_i}$ 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 $\boldsymbol{U}$ of length $\boldsymbol{m}$. This can be represented as $\mathrm{U} = [(\mathrm{mz}_1, \mathrm{int}_1), \cdots, (\mathrm{mz}_\mathrm{m}, int_m)]$.

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 ***V*** of length ***m***, where ***m*** is the number of ions remaining in the spectrum after filtering. Each component $\boldsymbol{V_i}$ contains the pairwise distance between ***P*** and $\boldsymbol{mz_i}$ as well as the intensity $\boldsymbol{int_i}$**.** This can be represented as $\mathrm{V} = [\mathrm{V}_1, \cdots, \mathrm{V}_k]$ with each component $\mathrm{V}_i = (\mathrm{mz}_i - \mathrm{P}, \mathrm{int}_i)$. Intuitively, each component can be interpreted as what the precursor remainder mass and intensity would be if the ***i***th 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. 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 ***F*** holds each of ***n*** theoretical fragment ions, where ***n*** is the length of the peptide minus one and $\boldsymbol{F_j}$ corresponds to the ***j***th fragment ion. Then, the pairwise distance between each remaining ion in the MS/MS spectrum and each theoretical fragment ion $\boldsymbol{F_j}$ is calculated and stored in a matrix ***W*** of size ***m*** by ***n***, where ***m*** is the number of ions remaining in the spectrum. This can be represented as

$$W = \begin{bmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{m1} & \dots & W_{mn} \end{bmatrix}$$

with each component $W_{ij} = (mz_i - F_j, int_i)$. Intuitively, each matrix component can be interpreted as what the ***j***th fragment's remainder mass and intensity would be if the ***i***th ion in the

spectrum were the ***j***th 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. 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. To account for uncertainty of ion position and smooth the histogram, the intensity of each ion is placed uniformly over an area equal to the MS/MS spectrum tolerance for the average ion in the histogram, i.e., the average inserted peptide or fragment mass. For diagnostic ions, a mass of 150 Da is used as the mass for smoothing. Furthermore, insertions into the histogram are normalized by the number of PSMs matching a particular peptide ion—that is, a grouping based on sequence, modification state, and precursor charge state—to prevent the inflation of features from abundant peptide ions.

Peaks in the histogram are defined by descending each side of a local maximum bin until a bin with zero intensity or a higher value is reached. 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.

Rather than using every PSM for what is inherently a noisy process, we select only those that are most likely to have the cleanest spectra. To do this, PSMs 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 number of representative PSMs for each mass shifts is then capped at 1000 as we found that these gave reliable values across iterations. The 1000 representative PSMs are selected randomly, but internally a seed is provided for reproducibility.

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 $\boldsymbol{W}$ for any ion type.

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 required 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 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 F_j from the experimental ion i, the same experimental ion responsible for that offset will also produce an "echo" offset from fragment 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 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.

3.2.2 Calculation of classification metrics from PTM-Shepherd

PTM-Shepherd internally calculates a series of metrics to characterize diagnostic features. One of these is the Area Under the Curve of the Receiver Operating Characteristic Curve (AUC-ROC, or just AUC). This metric can be computed directly from the ***U***-statistic of the test group computed as part of the Mann-Whitney *U* test. The formula for this statistic is:

$$AUC_t = \frac{U_t}{n_t n_c}$$

where ***t*** and ***c*** stand for test and control groups and ***n*** is the respective groups sample sizes. This metric has a useful interpretation as a rank probability statistic, i.e., it can be directly interpreted as the probability that a randomly chosen value from group ***t*** will be higher a randomly chosen value from group ***c*** [113]. Importantly, when calculated from the ***U***-statistic, AUC is not sensitive to class imbalances. This allows comparisons between diagnostic ions for mass shifts that have different numbers of PSMs, as in Figure 3-5b. The second metric calculated in the manuscript is precision, also known as positive predictive power. In classification problems, this metric is the inverse of FDR. Whereas FDR can be interpreted as the probability that a hit is a false positive given that it is positive, precision can be interpreted as the probability that a hit is a true positive given that it is positive. When using the presence of a particular feature to classify whether the spectrum contains the PTM of interest, this can be calculated easily from PTM-Shepherd output by the equation:

$$precision = \frac{proportion_m}{proportion_u + proportion_m}$$

where ***u*** and ***m*** correspond to unmodified and modified PSMs and ***proportion*** is the proportion of spectra containing the ion. Again, this metric is not sensitive to class imbalances when calculated this way, enabling direct comparisons between mass shifts with different numbers of PSMs.

3.2.3 Efficient data access and storage

This process was refined after encountering computational bottlenecks for large datasets in several places. In one approach, the delta mass bins are processed one-by-one (Fig B-3). Because spectral files can be very large and it is not reasonable to keep all of them open at once, spectral files would need to be parsed for every delta mass bin. Parsing spectral files is computationally expensive due to IO limitations and may also be CPU intensive based on the file format. Redundantly parsing files for every delta mass bin, potentially hundreds of times, is too wasteful. Another approach would be to pre-initialize histograms for every mass bin, then loop through spectral files one-by-one and parse and process them (Fig B-4). However, pre-initializing hundreds to thousands of histograms at the resolution required for MS fragment analysis would be extremely memory intensive. The volume of insertions going into these histograms may preclude the use of sparse arrays to save memory due to speed, and for large consortium-level datasets adequate sparsity cannot be assumed. A third option is to do this in two passes (Fig B-5). In the first pass, spectral features for all PSMs are calculated and cached in intermediate files that can be accessed more quickly, allowing them to be parse and opened hundreds of times throughout sample processing.

This also solves a second issue. Storing spectral features from the prior step in memory for use in the statistical testing step is also not feasible. This is because each PSM's features requires a space in memory equal to roughly

$$n * [(k - 1) * 2] + 2n$$

peaks, where ***n*** equals the number of peaks in a spectrum and ***k*** equals the length of the peptide, primarily because each fragment ion series produces ***k – 1*** arrays for every ion in the experimental spectrum. By storing them in intermediate files, we can take advantage of storage rather than RAM.

To facilitate fast file access, we store precalculated spectral features as binary arrays, with each PSM's position in the file indexed in a manifest (Fig B-6). Scans can be located by loading the small manifest, then selectively accessed rather than loading the whole file. This strategy cuts

loading time from seconds to single digit milliseconds. Because it is so much faster (~6 orders of magnitude) than recalculating diagnostic features for the statistical testing (Fig B-7), and because calculating diagnostic features is one of the more computationally demanding portions of this algorithm, it also cuts processing time.

3.2.4 Data processing

Three datasets were used throughout this manuscript. The first dataset consists of the Clinical Proteomics Tumor Analysis Consortium (CPTAC) phosphorylation-enriched clear cell renal cell carcinoma (ccRCC) samples [25] from the CPTAC data portal [46]. 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 [47] 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" [37] was turned on for the PTM-Shepherd [22] 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. 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 Figure 3-3. This was performed using the mzMLs rather than the deneutrallossed and deisotoped [23] 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.”

The second dataset consists of a novel protocol for photoactivatable ribonucleoside-crosslinking from the ProteomeXchange repository PXD023401 [33]. 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 [48] was used to clean up open search results. Filtering to 1% PSM, peptide, and protein levels was performed by Philosopher [49].

The third dataset consists of two samples from the ProteomeXchange repository PXD004245 corresponding to ADPR -enriched samples of mouse and HeLa origin [40]. 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.

We attempted to do searches for ADP-Ribosylated (ADPR) peptides using parameters as close to FragPipe defaults as possible to make them reproducible for other users. The default ADPR workflow includes both a labile and variable modification search for ADPR, allowing competition between fragmented and intact forms of the modification. Because the zero bin is defined as peptides with no mass shift and PTMs identified as variable mods have no mass shift, some ADPR-containing peptides are also present in the unmodified peptide bin. However, these only account for roughly 1 in 7 PSMs in the unmodified bin, and as such cannot be driving the trends discussed here.

3.3 Results

3.3.1 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 [23], 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).

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 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. Peptide remainder masses are calculated by computing mass differences between the theoretical, unshifted peptide ion and all ions in the spectrum. Fragment remainder masses are calculated by iteratively computing mass differences between every theoretical ion from the peptide backbone 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 abundance 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 (Figure 3-1). 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

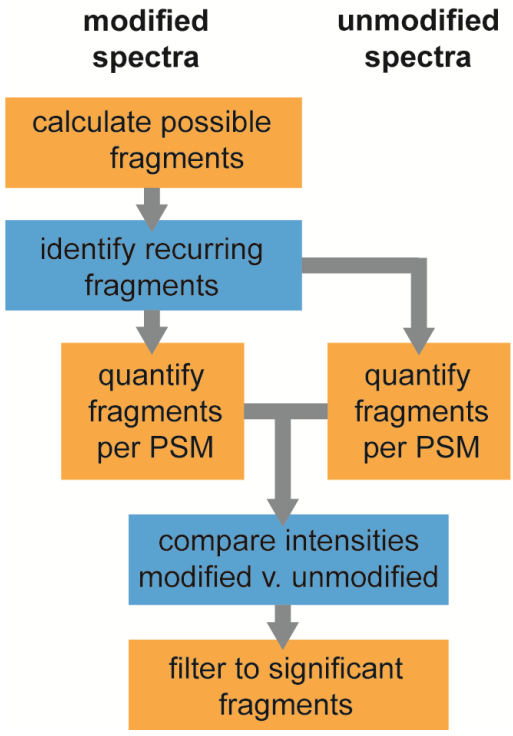

Figure 3-1: Workflow for diagnostic feature selection. First, all possible diagnostic features are calculated for every PSM. PSMs are then grouped by mass shift, and features that recur across PSMs are identified. For every recurring feature, the feature's intensity is extracted from every representative PSM within the mass shift bin and within the unmodified bin. These intensities are statistically compared between modified and unmodified spectra, then filtered based on statistical significance and different abundance metrics.

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 correct 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.

3.3.2 Data-driven discovery of diagnostic features

Glycopeptides contain labile modifications that produce rich sequences of diagnostic ions and peptide and fragment remainder masses [24]. We reasoned that detecting known glycopeptide fragmentation patterns would a good way to validate our algorithm's performance given the

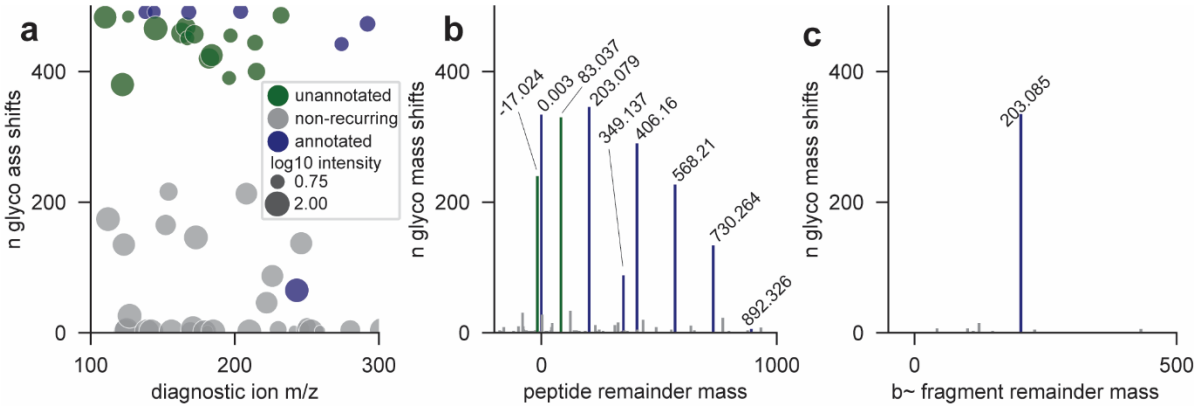

Figure 3-2: Diagnostic features of IMAC-enriched glycopeptides under high energy conditions. (a) Scatterplot of recovered diagnostic ions across glyco mass shifts. Ions occurring in >50% of mass shifts are considered recurring and are included in Fig 4-3. Color schemes for Fig 4-4c,d are consistent with Fig 2-2a. (c) Histogram of peptide remainder ions across glyco mass shifts. (d) Histogram of fragment remainder *b*-ions across mass shifts. Recovered features are generally explained by preexisting knowledge, and in cases where new features were found could be overwhelming determined empirically.

extensive literature characterizing glycopeptide fragmentation. To this end, we searched for glycopeptides in a large phosphorylation-enriched, TMT-labeled clear cell Renal Cell Carcinoma (CCRCC) dataset [25]. Phosphorylation enrichment by IMAC, the method employed in this publication, has been shown to simultaneously enrich glycopeptides, particularly those bearing sialic acids [26,27], 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.

We first wanted to verify that we could detect diagnostic ions associated with glycopeptides. After discarding any mass shifts less than 50 Da we were left with 493 likely glycan mass shifts,

each of which should be enriched for diagnostic ions 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 [7,28], including three known sialic-acid related oxonium ions at 274, 292, and 657 m/z (Figure 3-2a, Table B-1). 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 (Figure 3-3). Known ions clustered together meaningfully, with annotated GalNac, Hex, HexNac, and PhosphoHex 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 (Figure 3-3, cluster 5) that was highly correlated with both oxonium ions (Table B-2) 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. Recent manuscripts make no mention of these as diagnostic ions [12,29,30], 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 (Figure

3-2b). 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 surprising. This pattern—consisting of a cross-ring fragmentation event at the core GlcNAc and a loss of an ammonium, respectively—has previously been identified as a conserved fragmentation pattern for glycopeptides [31], but appears not to be used in current state-of-the-art tools [6,7,32]. This indicates that even for very well characterized 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 (Figure 3-2c, *y*-ion series at Figure II-1) 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 for 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 considered 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.

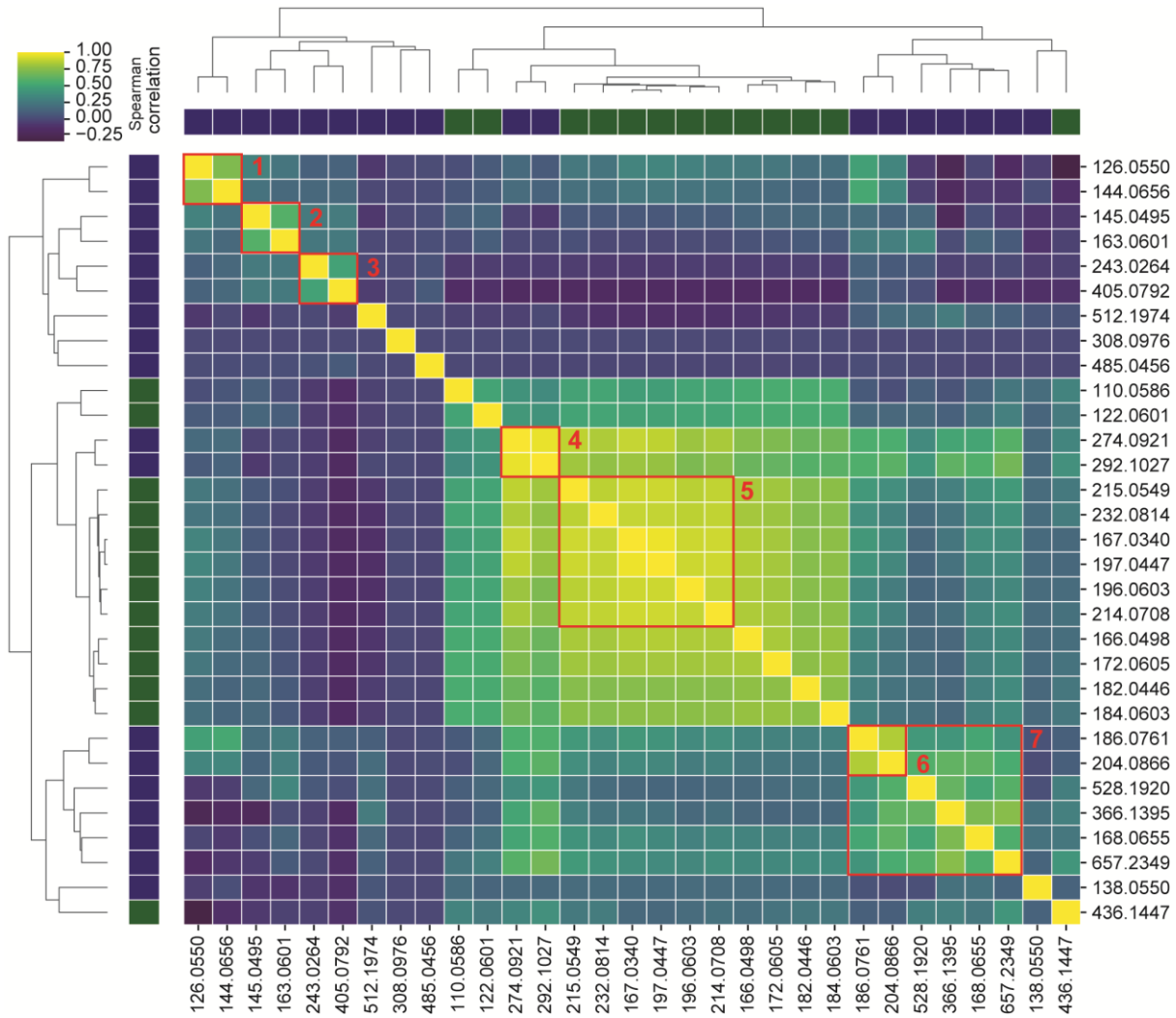

Figure 3-3: Clustering between known and unknown diagnostic ions for sialic acid enriched TMT-glycopeptides. Spearman correlation clustering between diagnostic ions across all glyco spectra. Previously annotated ions are labeled blue, unannotated ions are labeled green. Identifiable clusters are as follows: 1: GalNac, 2: Hex, 3: PhosphoHex, 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 *b*-ions across mass shifts.

To conclude, we showed that PTM-Shepherd was sensitive to known diagnostic features for glycopeptides. New features detected by PTM-Shepherd 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. Finally, we proved that the

information provided by PTM-Shepherd can be incorporated into subsequent searches to increase to fine-tune parameters for different experimental settings.

3.3.3 Fragmentation of complex post-translational modifications

ADP-ribosylation (ADPR) has seen a surge of interest in recent years, with many enrichment methods [38,39], and instrumental techniques [40] 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 painstaking analysis and examination of individual spectra [41]. We believed that PTM-Shepherd's diagnostic feature mining module could expedite fragmentation studies and reveal new, useful insights to their behavior. To demonstrate this, we reanalyzed ADPR-enriched data from Martello et al. [40] 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 [41]. As expected, we found previously annotated diagnostic ions (Figure 3-4a, Table B-3a,b) corresponding to almost every expected breakpoint on the ADPR side chain (Figure 3-4b). 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 found at high levels in unmodified spectra (73%), meaning its presence was not specific to PSMs with ADPR. 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

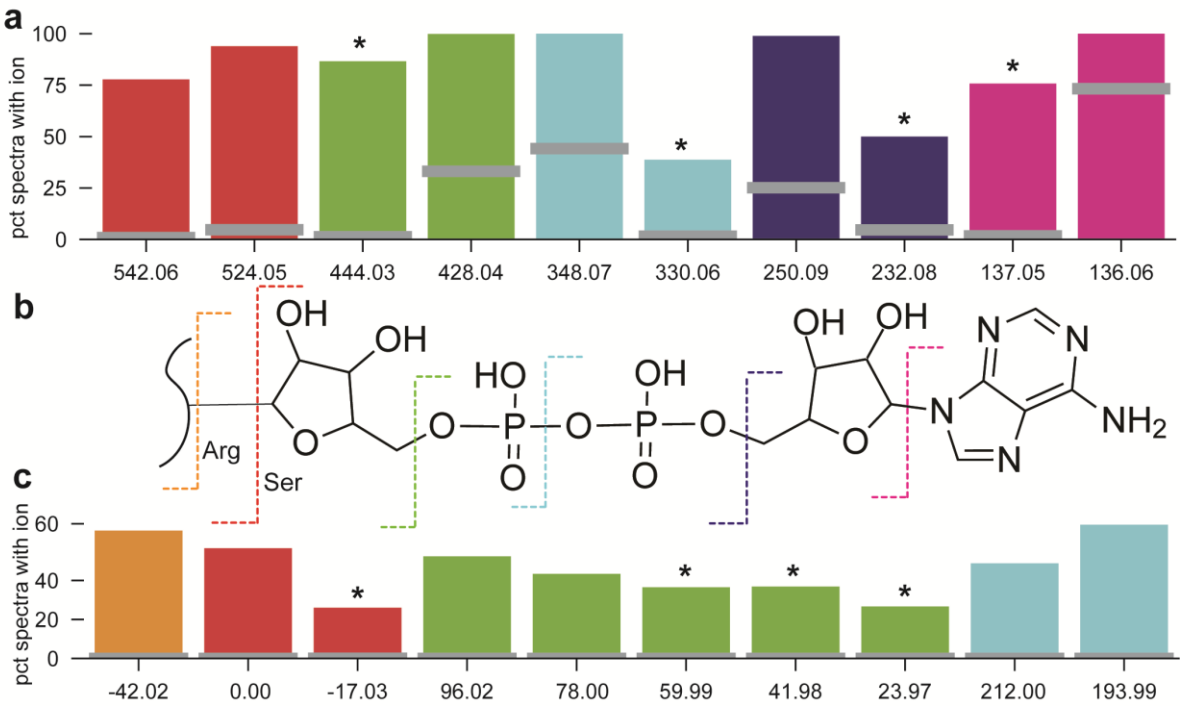

Figure 3-4: Analysis of ADPR fragmentation patterns. (a) ADPR diagnostic ions. Colored bars show the percentage of ADPR PSMs containing the diagnostic ion, while gray bars show the percentage of unmodified PSMs containing the diagnostic ion. Ions detected in both datasets were averaged across all values. A "*" denotes novel features discovered by PTM-Shepherd. (b) Structure of ADPR. Dashed lines correspond to breakpoints in the molecule, and color corresponds to the diagnostic features produced during fragmentation. (c) ADPR peptide remainder masses. PTM-Shepherd revealed many derivatives of known diagnostic ions that were more specific to ADPR as well as a series of unreported successive neutral losses from the peptide backbone.

their annotated counterparts and thus may be of interest to other 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 (Figure 3-5, Spearman's R^2: mouse = 0.857; HeLa = 0.884). We have previously commented on this phenomenon when

looking at biotin-derived Cysteine probes [16]. 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. It is worth noting that the issue of co-fragmented ions has been well-studied in the context of isobaric tandem mass tags [42]. But, to our knowledge, there has been little discussion of parallel issues when using diagnostic ions for ADPR analysis.

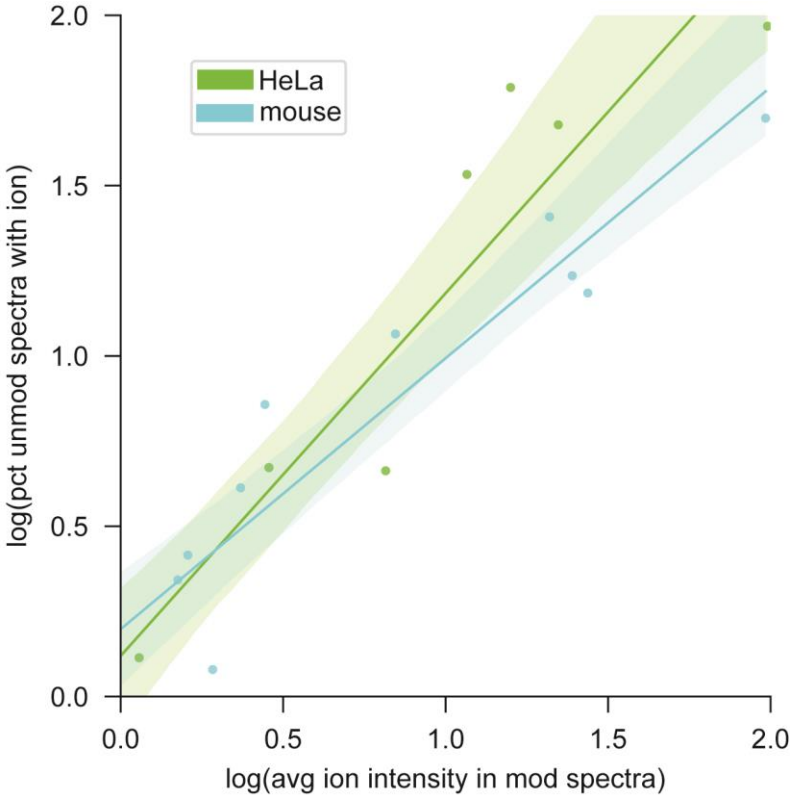

Figure 3-5: Correlation between diagnostic ions' average intensities and their presence in unmodified PSMs. Across both HeLa and mouse datasets, there is a strong correlation between an ion's intensity and its lack of specificity for the modification. This is caused by it being more likely to cross the detection threshold in unmodified PSMs when the ADPR is the minor product in a spectrum. Confidence intervals for Pearson's correlation are highlighted.

PTM-Shepherd also identified both types of remainder ions in this dataset, peptide (Figure 3-4c) and fragment. Of note was PTM-Shepherd's recovery of a -42 Da peptide remainder mass from the Arg-directed ADPR dataset (Table B-3b). 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 (Table B-3b). 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 (Figure 3-4b, 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 five 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 found in no 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 3-4b, red) will appear at the precursor *m/z* - 17 for unmodified peptides, but at precursor *m/z* – 558 from modified peptides.

3.3.4 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 them in large numbers from the CPTAC phosphorylation-enriched glycosylation 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 (Table B-1). Every diagnostic ion was evaluated individually for its ability to separate glyco and unmodified spectra based on its precision and AUC. 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*** (Figure 3-6a). 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). Because there is a detection limit for ions in mass spectrometers, less intense ions are also less likely to show up in spectra. For co-fragmented spectra, the presence of the ions from the minor product is inversely proportional to the spectral purity and proportional to the ion's intensity when its peptide is the major product. In other words, more intense diagnostic ions are more likely to appear in unmodified spectra because they can exceed the lower detection limit even for relatively pure unmodified PSMs. This has profound implications for researchers using diagnostic ions for PTM research. First, the most intense diagnostic ion for a PTM, a typical choice for diagnostic-ion triggered methods, might not be the optimal one. Second, less intense versions of the same ion—such as neutral losses or isotopic peaks—might have better statistical properties.

But this is a trend can be reversed by taking intensity information into account rather than only checking for the presence or absence of the ion (Figure 3-6b). 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

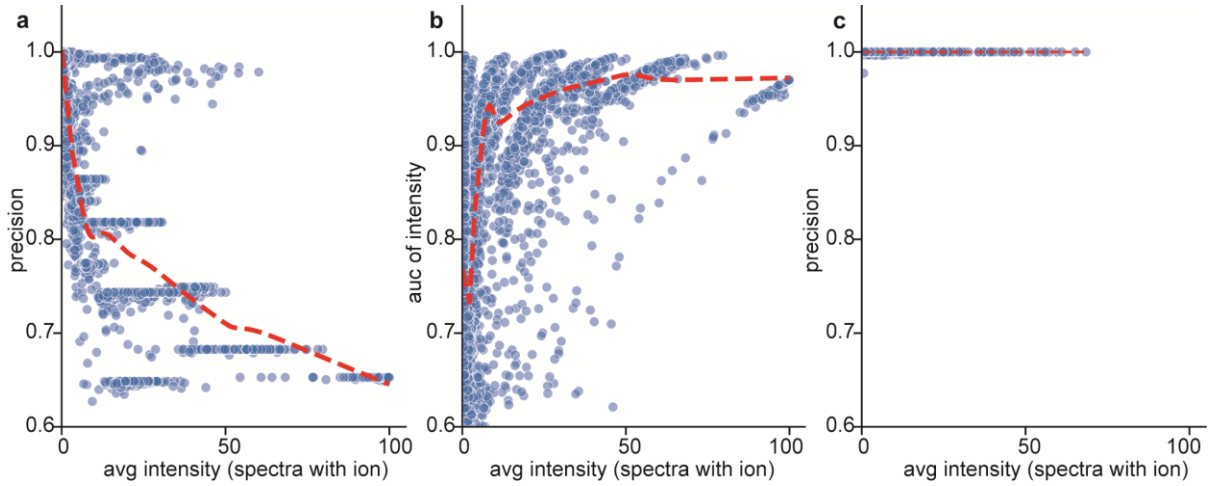

Figure 3-6: Trends in diagnostic and peptide remainder ions. Dotted red lines track a LOESS fit. (a) Relationship between diagnostic ions' observed intensity and the precision of their presence. (b) Relationship between diagnostic ions' observed intensity and the classification strength of their intensity as measured by AUC. (c) Relationship between peptide ions' observed intensity and the precision of their presence.

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 (Table B-3a). 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 (Figure 3-6c). 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 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 [43].

3.4 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 painstaking experimental characterization could be recreated with ease 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—could be automated and democratized to reach a broader audience. 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 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 [16]. It could be advantageous to develop custom modification scores for localization-by-proxy strategies [44] or as rescoring features in Percolator [45]. 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.

3.5 Data availability

Raw mass spectrometry files are available from public repositories. Our method has been implemented within PTM-Shepherd [22] and is freely available as part of the FragPipe suite of tools (https://fragpipe.nesvilab.org/).

3.6 Acknowledgements

DJG was supported in part by the Proteogenomics of Cancer Training Program (5T32CA140044). We would also like to thank our users for providing feedback on our tools. DJG developed the algorithm and wrote the software, with DAP and FY assisting in the development of the algorithm and DAP assisting in the development of the software; DJG and

AIN jointly analyzed the data and conceived the project while AIN supervised the project; DJG and DAP wrote the manuscript with input from all authors.

CHAPTER IV

Open Search Modification Characterization: Applications in Synthetic Modifications

The portion of this chapter dealing with the characterization of a cysteine-specific probe was adapted from *Enhancing Cysteine Chemoproteomic Coverage through Systematic Assessment of Click Chemistry Product Fragmentation* published in Analytical Chemistry

4.1 Introduction

One of the benefits of mass spectrometry is its versatility. By changing the sample input, existing proteomics workflows can sometimes be adapted to answer manifold experimental questions. Answering some biological questions can be as simple as enriching for a particular PTM[112,114-117]. Other biological questions, however, require specialized techniques. Researchers have shown extraordinary ingenuity in designing synthetic modifications to unlock new modes of analysis.

One such example is found in the field chemoproteomics. Chemoproteomics deals with building chemical probes that attack specific functional groups inside cells, covalently labeling them[29]. These can be used a few ways, such as by adding drug-like chemical probes to a sample to identify druggable residues[118]. Alternatively, probes that structurally mimic a hard-to-study biological PTM of interest can also be used to increase their signal[119,120]. Another example is found in studying the interactions between RNA and proteins[121-124]. Synthetic nucleotide analogs that crosslink proteins and RNA fix their interactions in place and can be used to map how the RNA-binding proteome changes across conditions. Finally, protein-protein crosslinking can be used to study protein-protein interactions at proteome scale or even to study protein structure dynamics[125-129].

The diversity of methods employed above underscores a major point: the number of PTMs researchers have access to is exploding and techniques to analyze them are in short supply. Here

we show that the PTM characterization techniques developed in Chapters Two and Three are generalizable, widely applicable, and fill important voids in the proteomics toolkit.

4.2 Methods

Cysteine click chemistry probe

Samples collected at high resolution using a Thermo Scientific Orbitrap mass spectrometer in DDA mode were processed using FragPipe v15, including MSFragger[15] v3.2 and Philosopher[49] v3.4.13. Mass spectrometry data from this experiment has been uploaded to ProteomeXchange under the identifiers PXD028853 and PXD030737[131]. Files were processed without conversion to mzML. Several modifications to the default "Mass-Offset-Common-PTMs" workflow were made to facilitate diagnostic feature extraction from the probe. The default list of mass offsets was replaced with 0 and 463.2364, the mass of the probe, and "Write calibrated MGF" was enabled to enable diagnostic feature extraction. Files were searched against the Uniprot[132] reviewed human sequence databased with decoys and common contaminants appended retrieved on 13 June 2021. Filtering to 1% PSM, peptide, and protein levels was performed by Philosopher[130]. Following this, an early version of the diagnostic feature extraction algorithm described in Chapter III was used to identify characteristic features.

RNA crosslinking

The second dataset consists of a novel protocol for photoactivatable ribonucleoside-crosslinking from the ProteomeXchange repository PXD023401[33]. Samples were acquired on a Thermo Fisher Orbitrap Fusion Lumos using HCD fragmentation in DDA mode. Only the two 4SU-specific raw files from the repository were used 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[132] database described above. Crystal-C[48] was used to clean up open search results. Filtering to 1% PSM, peptide, and protein levels was performed by Philosopher [49].

Protein crosslinking

The third dataset consists of protein crosslinking data retrieved from the ProteomeXchange repository PXD020704[133]. Data was collected on a Thermo Fisher Q-Exactive Plus in DDA mode. Three replicates each from the No_XL, FoAl, and DSS groups were run as three separate experiments in FragPipe v18.0. The "Diagnostic-Ion-Mining" workflow was used with default parameters to process the data. Files were searched against the Uniprot[132] reviewed human sequence databased with decoys and common contaminants appended retrieved on 25 June 2022 as background. Sequences for BSA, ovotransferrin, and alpha-amylase were appended. Filtering to 1% PSM, peptide, and protein levels was performed by Philosopher[130].

4.3 Results

4.3.1 Applications in chemoproteomics

Chemoproteomics probes are designed to covalently bind to specific functional groups such as those found on amino acid side chains[29]. This platform enables small molecule screening for drugs at proteome scale in multiple ways. In one use case, probes are mixed into a sample in two conditions: alongside a pharmaceutical agent or alone. When the drug is bound to a protein, it prevents the probe from binding to residues that are blocked. By quantifying the change in probe abundance for a particular site, binding efficiency and off-target hits can be determined. In another use case, the covalent probe acts as a proto-drug and provides a straightforward approach to perform structure-guided drug design even for proteins with structures that cannot be determined by crystallography[134]. While one end of the chemical probe contains a warhead

specific to a residue, the other end commonly contains a click chemistry tag, to which enrichable moieties such as biotin can be attached[127,135,136]. Chemical probes frequently form diagnostic ions and other diagnostic fragmentation patterns, which has led to an interest in tools that characterize them to increase peptide recovery[137]. Indeed, the tool that comes closest to identifying diagnostic features in the manner of PTM-Shepherd is pChem[138], a program designed specifically for chemoproteomics experiments, although it requires isotopic labeling signatures to be present at the MS1 and M2 level.

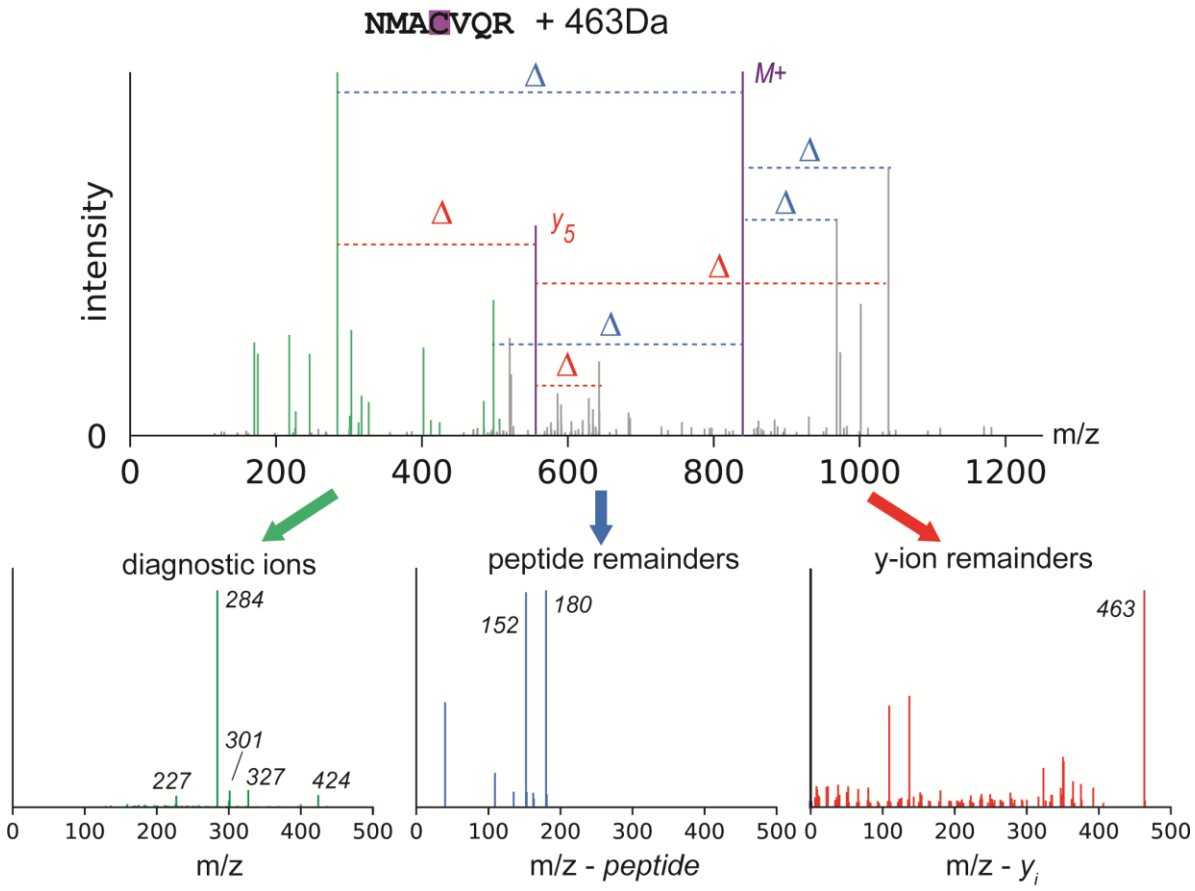

Figure 4-1: Identified diagnostic features for a Cys-biotin probe. A model spectrum shows how features are calculated, with potential diagnostic ions in green, theoretical fragment and peptide ions in purple, potential peptide remainder masses in blue, and potential fragment remainder masses in red.

Cysteine's unique chemistry has made it a target for chemoproteomics[139]. Prior studies have identified swaths of probed cysteines, but proteome-wide coverage has yet to be achieved[140].

One possible cause is that the probes employed are labile, leading to a mismatch between search parameters and experimental spectra. With the goal of increasing coverage of the ligandable cysteine proteome, we comprehensively characterized the fragmentation behavior of a biotinylated azide-alkyl cysteine probe (BAAC)[131].

Table 4-1: Diagnostic features of a biotintyl azide-alkyl cysteine probe

Ion type	Adjusted mass	P-value	AUC	Porportion of modified spectra	Proportion of modified spectra
diagnostic	284.1428	2.29E-184	0.894745	1	0.85
diagnostic	301.1694	3.35E-171	0.866078	0.81	0.1
diagnostic	327.185	5.38E-165	0.866029	0.86	0.16
diagnostic	424.2488	8.00E-139	0.827486	0.78	0.12
diagnostic	227.0846	5.42E-129	0.819819	0.79	0.17
Y	152.0994	2.80E-105	0.737092	0.47	0
Y	180.1038	4.96E-96	0.720232	0.44	0
b	463.2362	3.24E-106	0.749554	0.52	0.03
y	463.2364	2.06E-133	0.790352	0.59	0.02

The first feature I investigated was diagnostic ions. Some search engines use diagnostic ions during the search[47,141], but they can also be used during data acquisition[142]. PTM-Shepherd identified eight diagnostic ions. One was discarded due to its mass being larger than the mass of the modification, which we initially believed was improbable, and another two were discarded due to their comparatively low statistical significance. The other five (Fig 4-1) were subjected to additional scrutiny. The most abundant ion was found at an m/z of 284. It was deduced to be a biotin-oxonium ion (Fig III-1) unique to the probe's structure. Another ion at 227 is commonly observed as a product of biotinylation fragmentation and was expected due to the presence of biotin on the probe[137]. Three additional ions at 424, 327, and 301 were not expected, however, and had not been observed for this class of chemical probe before. It was determined that the most likely structures for two of these ions, 327 and 301, correspond to fragmentation within the triazole ring, which was previously assumed to be mostly stable under fragmentation conditions[128]. I also observed a striking lack of specificity for the most intense fragment, the biotin oxonium ion. Although it was present in 100% of PSMs containing a probe, it was also present in 85% of unmodified spectra. This is another example of the phenomenon from Chapter

III where intense ions can become major products in co-fragmented spectra, and as such are poor diagnostic ions in most circumstances (Table 4-1, full table can be found at Table C-1).

I then explored mass shifts that remained on the peptide or fragment ions. Two peptide remainder masses were particularly abundant: 152 and 180 (Table 4-1). These masses map once again to breakpoints within the triazole reagent within the linking arm, providing additional evidence for the lability of the group in the gas phase. These masses were uniquely found as peptide remainder masses and were not identified as fragment remainder masses. Useful fragment remainders were restricted to those that appeared for both *b*- and *y*-ions, ultimately limiting fragment remainder masses to the mass of the mod itself. In other words, the probe is found in the intact form when identifying fragment ions. Although the peptide remainder masses can be included in a labile search, modifications that remain intact during fragmentation see little benefit from this mode. Accordingly, the BAAC probe saw minimal increase in cysteine coverage when incorporating them into the search.

While it was expected that a labile modification yielding intense diagnostic ions would also leave intense peptide and fragment remainder masses that would benefit from labile mass offset searches, this was proven to not be the case. Failure to improve cysteine coverage notwithstanding, this study revealed new insights about common click chemistry tags that will find utility in other fora.

4.3.2 Applications in RNA crosslinking

RNA crosslinking studies also feature labile modifications that are hard to characterize. RNA crosslinking studies aim to characterize the RNA-bound proteome. This is done using a combination of synthetic nucleotide analogs and UV treatment[123,124]; a phosphate group normally present in the nucleotide is replaced by a sulfur and is incorporated into RNA polymers, and when exposed to UV light reacts with nearby molecules. When this happens *in vitro*, proteins that normally transiently interact with RNA are lashed to them. The RNA can then be detected as a PTM using LC-MS. Of course, this introduces complexity to proteomics analyses. Repeating

sugar molecules can fragment in myriad ways, frustrating attempts to localize or even identify RNA moieties. Furthermore, the crosslinking can occur on several amino acids. Bae et al. recently developed pRBS-ID, an RNA crosslinking workflow utilizing photoactivatable nucleotides and chemical RNA cleavage to overcome these challenges[122]. 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[34].

Table 4-2: Open search of pRBS-ID data

Peak Apex	PSMs	Mapped mass	Localized PSMs	N-term localization	AA	AA enrichment	AA psm count
0.0002	7751		291				
226.0594	4241	226.0594 mass shift	1343		R	2	160
1.0032	2378	First isotopic peak	658		F	1.6	65
57.0216	945	Iodoacetamide	806	38.94	H	11.9	184
94.0168	830	94.0168 mass shift	799	7.47	H	9.2	140
-48.0032	568	Homoserine lactone	559	26.23	M	53	373
227.0634	1204	First isotopic peak + 226.0594 mass-shift	252		H	2.5	12
41.0266	450	Amidination	124		M	16.4	26
283.0806	486	226.0594 mass-shift + Iodoacetamide	128	5.56	H	2	5

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 two unannotated mass shifts of 226 Da and 94 Da at high amounts likely corresponding to the modification of interest (Table 4-2). These mass shifts localized only 32% of the time. This can be explained by the lability of the 226 modification. Notably, the fragment remainder masses PTM-Shepherd identified for both mass shifts were nearly identical, 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 (Fig 3ab, Table 4-3). Like the loss of ammonia described from glycopeptide's *Y*-ion series described in Chapter III, this mass shift appeared to be diagnostic for RNA-crosslinked peptides (226 mass shift AUC = 0.57, 94 mass The localization rates for these two mass shifts differ markedly, with the 94 Da mass shift being localized 96% of the time but the 226 mass shift being

Table 4-3: Diagnostic fragmentation patterns for pARS-ID products

Peak apex	Ion type	Mass	Remainder propensity	Delta Mod Mass	Percent Mod PSMs	Percent Unmod PSMs	Avg Intensity Mod	Avg Intensity Unmod	Intensity Fold Change
226.059	diag	133.050			39	12.9	7.81	4.76	5.0
226.059	diag	115.040			31.1	10.4	6.04	3.84	4.7
226.059	b	94.029	31.7	-132.031	36.9	1	47.65	24.91	70.6
226.059	b	77.005	18.5	-149.054	17.9	1	29.43	13.13	40.1
226.059	y	94.030	22.2	-132.029	35.7	1	46.95	24.91	67.3
226.059	y	77.007	15.1	-149.053	17.4	0.9	28.81	13.05	42.7
94.017	diag	215.058			30.62	16.5	14.28	7.01	3.8
94.017	b	94.017	36.7	0.000	39.53	0.9	38.37	26.18	64.4
94.017	b	77.002	22.5	-17.015	21.71	1	34.51	12.29	61.0
94.017	b	-19.064	20.8	-113.081	15.89	0.6	36.8	21.02	46.4
94.017	b	66.027	20.1	-27.990	15.89	1.3	19.82	23.58	10.3
94.017	y	94.017	23.8	0.000	39.53	0.9	38	26.18	63.8
94.017	y	77.006	14.4	-17.011	19.38	0.9	30.34	13.05	50.1

ammonia) was identified from both mass shifts' *b*-ions (Fig 3ab, Table 4-3). Like the loss of ammonia described from glycopeptide's *Y*-ion series described in Chapter III, this mass shift appeared to be diagnostic for RNA-crosslinked peptides (226 mass shift AUC = 0.57, 94 mass shift AUC = 0.58). Utilizing the 77 Da and 94 Da fragment remainder masses together results in a 3.8% increase in crosslinked PSMs, and a remarkable 47.4% increase over a standard mass offset search. With this fragmentation information in hand, we reprocessed the data using

FragPipe's labile search mode, which is fully compatible with built-in tools for quantitation [35] and localization [36,37] and fills the requirements of their workflow.

After a more targeted search, we also wondered whether any additional diagnostic features might appear for the RNA-crosslinked peptides and performed a second pass at diagnostic feature

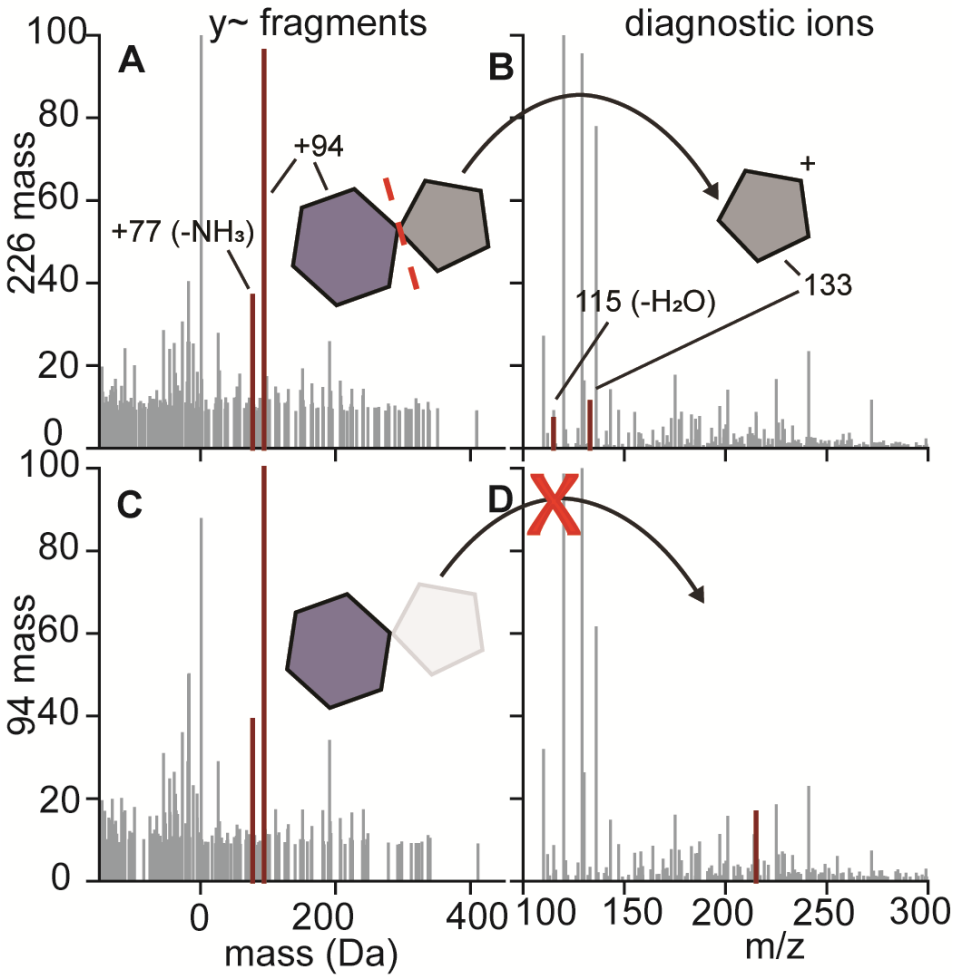

Figure 4-2: Characterization of 4SU fragmentation from a pRBS-ID experiment. A) All possible fragment remainder masses for *y*-ions from the 226 Da mass shift. Remainder masses that passed PTM-Shepherd's filtering are highlighted in dark red, corresponding to the retention of the 94 Da fragment on the peptide. Figure 4-3b,c,d use the same color scheme as Fig 3a. (B) Diagnostic ions derived from the fragmentation of the nucleoside analog from the 226 Da mass shift. (C) All possible remainder masses for *y*-ions from the 94 Da mass shift. (D) All possible diagnostic ions from the 94 Da mass shift.

mining (Table C-2). Diagnostic ions can be of particular interest for future analyses, such as the ion-triggered instrument routines mentioned above, even if they are left unused at the present. We found two easily explicable diagnostic ions for the intact nucleoside (Fig 4-2b): 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. Accordingly, these ions were not diagnostic for the MS1

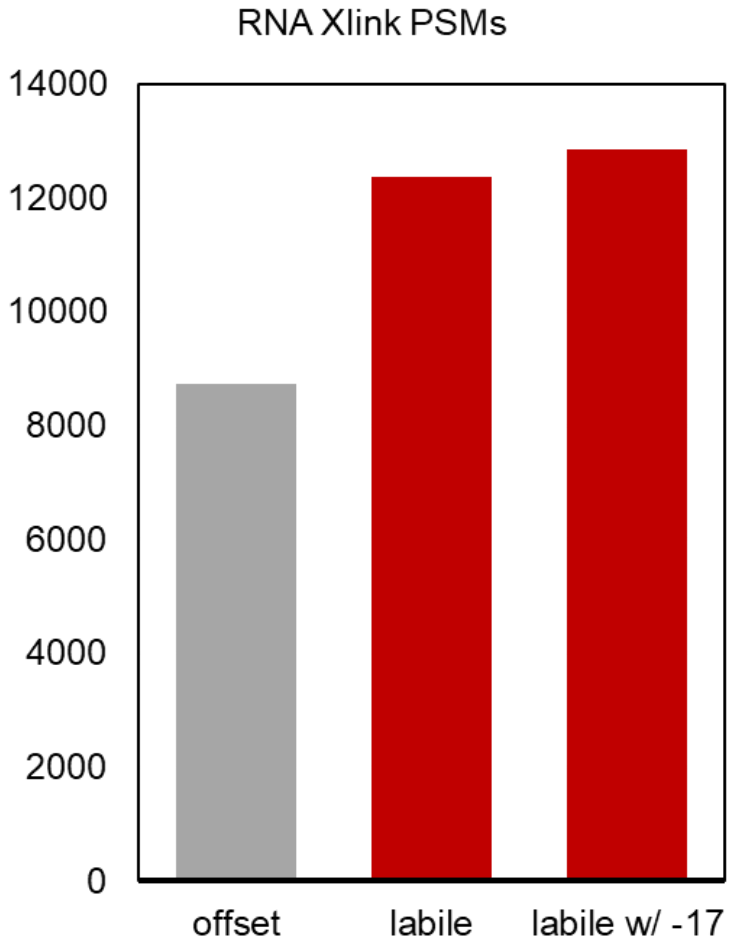

Figure 4-3: Improvements in RNA crosslinked PSMs with different parameter settings. Comparisons between traditional offset searches, labile searches with the partial nucleotide mass shift, and the partial nucleotide mass shift with a loss of 17 Da discovered by PTM-Shepherd showed an increase in cross-linked PSMs when including additional information.

mass shift corresponding the nucleoside without the ribose (Fig 4-2d), as with the ribose already dissociated there is nothing left to form the diagnostic ion.

4.3.3 Applications in protein-protein crosslinking

Proteins can interact in various ways, but direct protein-protein interactions (PPIs) require them to be in close proximity. PPI experiments have traditionally been done using one protein as bait, fixed in place, while other proteins from the sample flow over it. The idea behind this strategy, called affinity purification (AP), is that proteins interacting with bait proteins will be enriched in the sample and identifiable by MS[143,144]. Protein crosslinking can also be used for interactome studies by capturing which proteins are found in close proximity when the crosslinkers are added[125]. This has the added benefit of being able to characterize the entire protein interactome at once with simple sample preparation.

This technique introduces quite a bit of computational complexity into the database search. For one, peptide precursors now correspond to the mass of two peptides plus the crosslinker mass. Searching the MS/MS spectra for combinations of peptides increases the search space exponentially. Furthermore, individual MS/MS spectra contain fragments not just from two peptides, but peptide fragments ions that are shifted by unknown masses when the crosslinked peptide fragments. When using non-cleavable crosslinkers, a peptide p_1 of length 5 crosslinked to a peptide p_2 of length 7 could produce 4 *y*-ions, each of which could be shifted by the masses of 6 $_p2$ *y*-ions or 6 p_2 *b*-ions. Cross linking search engines assume that only the intact peptide *p2* exists as a PTM on peptide p_1's fragments[145], but the likelihood of one peptide remaining intact while the other fragments is small. Cross linking search engines thus rely primarily on unshifted fragments to identify two peptides, each of which is present in a noisy spectrum with a large search space. Better computational strategies are needed to deal with this challenge.

Recent work by Slavin et al. (2020)[133] identified open searches as a way to find the mass of a crosslinker. The approach involves a series of expensive computations, including open searching, identifying multiple peptides per spectrum, then checking for recurring offsets in the spectra as

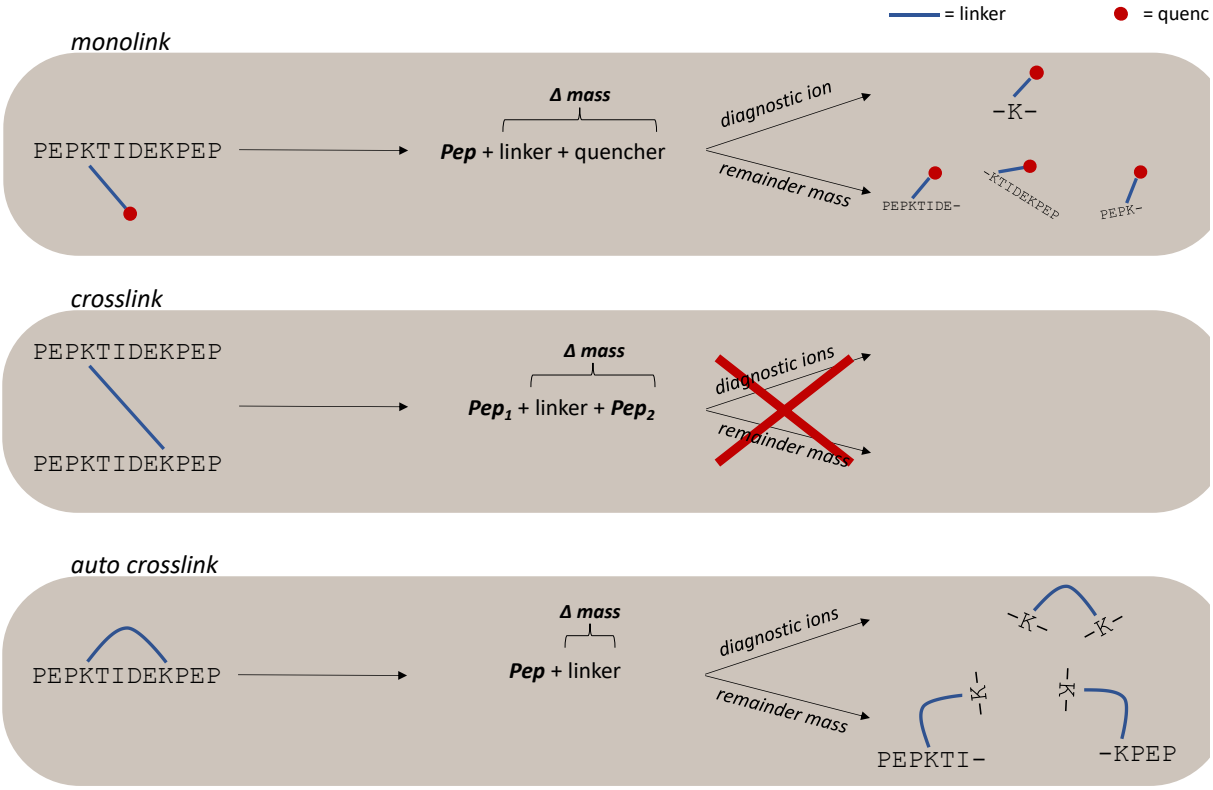

Figure 4-4: Crosslinking options and the ions they lead to. A) Monolinked peptide. The dead end crosslink produces a consistent and analyzable delta mass. Diagnostic ions and fragment remainder masses show a linker that does not fragment well. B) Crosslinked peptides. The delta mass produced by crosslinked peptides is dependent on the linker mass and the mass of the distal peptide. Because the distal peptide is not a constant mass, PTM-Shepherd cannot bin these observations in order to extract diagnostic information. C) Auto crosslinked peptide. Peptides crosslinked to themselves present a consistent mass of the linker alone. Their fragmentation patterns mimic those of true crosslinked peptides and can act as their proxy, but auto crosslinked peptides can be identified more easily.

combinations of the two peptides. The result is the mass of the crosslinking portion of the reagent. I reasoned that our pipeline could reproduce their results while providing additional information on the fragmentation characteristics of crosslinked peptides.

Rather than looking at peptides crosslinked to other peptides, I reasoned that the same information about crosslinker masses and fragmentation patterns could be garnered from monolinks and auto crosslinks. To extract diagnostic fragmentation patterns from a PTM, PTM-

Table 4-4 Open search mass shifts from a DSS crosslinking experiment

Peak apex	PSMs	Mapped mass	Localized PSMs	N-term localization rate	AA1	AA1 enrichment score	AA1 PSM count
-0.0002	1498		20				
26.016	129	Acetaldehyde	129	92.25	R	2.2	6
57.0216	180	Iodoacetamide	147	27.22	K	5	57
40.031	69	Propionaldehyde	69	95.65	F	2.2	7
-14.0154	96	-14.0154 mass-shift	93		I	22	67
155.095	156	Monolink (ammonia) of DSS/BS3 crosslinker	153	10.26	K	10.9	130
-17.0274	69	Loss of ammonia	68	49.28	C	7.5	26
-91.0092	62	-91.0092 mass-shift	61		C	15.4	48
0.9844	80	Deamidation	70	9.88	N	19.5	54
43.0054	44	Carbamylation	24	47.73	K	2.9	6
27.9938	31	Formylation	31	12.9	K	11.5	28
53.917	50	Replacement of 2 protons by iron	43		D	2	7
11.999	35	formaldehyde adduct	35	68.57	H	8.4	8
1.0032	67	First isotopic peak	17				
156.079	46	Monolink (water) of DSS/BS3 crosslinker	46	13.04	K	11.1	40
28.0246	66	di-Methylation	65		A	10.3	58
-31.0442	24	-14.0154 mass-shift + Loss of ammonia	21	20.83			
162.0504	17	Hexose	3	5.88			
138.0674	35	Intact DSS/BS3 crosslinker	33	20	K	7.3	19

Shepherd requires the PTM to produce consistent MS1 delta masses. Monolinks are crosslinkers that only reacted with a single residue, so the mass shift detected on a monolinked peptide corresponds to the mass of the linker and the mass of the reagent used to quench the crosslinking reaction (Fig 4-4A). Autolinks, being crosslinks between two residues of the same peptide, also produce a consistent mass of just the linking arm (Fig 4-4 4C). Finally, the case of two crosslinked peptides produces a mass shift of the combined linker arm and secondary peptide. Since the secondary peptides are not consistent masses, PTM-Shepherd cannot detect them as a single modification (Fig 4-4B). However, we can use the fragmentation patterns of monolinks and auto crosslinks to deduce the fragmentation patterns of peptide-peptide crosslinks.

One of the most abundant mass shifts detected by PTM-Shepherd was a mass of 155.0950 Da, the expected combined mass of the monolinker and the quenching reagent. Interestingly, I found several diagnostic ions associated with this mass shift. The top diagnostic ion (239.1752 m/z) corresponded to a Lys immonium ion (84.0814 m/z, Fig 4-4A) with the monolink intact and was highly specific to the monolink mass shift (96.60% of monolink spectra compared to 12.06% of unmodified spectra, AUC = 0.98). I also detected the mass of the monolink on *b*- and *y*-ions. Taken together, these indicate that the linker arm is not prone to breakage.

Table 4-5 Diagnostic features from a DSS crosslinking experiment

Peak apex	Ion type	Mass	Remainder propensity	Delta mod mass	Percent mod	Percent unmod	Avg intensity mod	Avg intensity unmod	Intensity fold change
155.095	diag	239.175			96.9	12.06	46.35	3.35	100.0
155.095	diag	256.202			48.8	0.23	2.39	0.27	100.0
155.095	diag	312.192			55.8	5.80	5.5	2.6	20.3
155.095	diag	174.112			38.0	0.23	1.41	2.82	82.2
155.095	diag	267.170			35.7	0.70	3.88	2.96	67.2
155.095	diag	156.102			44.2	3.71	2.96	3.13	11.3
155.095	b	155.095	37.7	0	55.8	1.16	67.9	4.87	100.0
155.095	y	155.095	35.2	0	55.0	0.70	57.21	4.26	100.0
138.067	diag	305.222			71.0	0.70	8.97	3.38	100.0
138.067	diag	222.149			51.6	0.93	5.2	2.69	100.0
138.067	diag	239.175			51.6	12.06	23.92	3.35	30.6
138.067	b	138.067	49.2	0	61.3	2.09	43.64	7.83	100.0
138.067	y	138.067	32.2	0	67.7	2.09	44.7	4.87	100.0
138.067	y	221.142	17.3	83.075	45.2	1.39	23.86	7.65	100.0

Because the mass of the quenching agent is known beforehand, the mass of crosslinker can be deduced from the monolink mass. It was also detected automatically at the correct mass of 138.0674 Da (Table 4-4). Like the monolink mass shift, a series of diagnostic ions (Table 4-5) was detected. The top hit corresponds to a double Lys immonium ion structure (Fig 4-4C), where a Lys immonium ion is attached to each side of the linking arm. This was highly specific to the crosslinked spectra (70.97% of crosslink spectra compared to 0.70% of unmodified spectra) and might be useable for the detection of crosslinked peptides. More interesting were the masses detected as fragment ion remainders. The mass of the linking arm was detected as a remainder

mass from fragments, although this would not be the case in true peptide-peptide crosslinks. More interestingly was another fragment remainder mass corresponding to the mass of the linker arm retaining a single Lys immonium ion as a mass on the fragment ions. To produce this, the distal peptide would need to have undergone two fragmentation events on either side of the linked Lys. The proximal peptide then fragments with the distal peptide's remnants grimly attached (Fig 4-4C). This is powerful, because it should allow us to identify additional ions for crosslinked peptides and boost their identification rate as well as localize the crosslink to the appropriate residue.

Steigenberger et al. (2019)[146] have previously reported the detection of these diagnostic ions from synthetic crosslinked peptides and commented on their abundance and utility, confirming that auto crosslinks can be used to determine fragmentation characteristics of peptide-peptide crosslinks. Although the authors above explored the conditions leading to the fragment remainder masses as well, there are holes in their applications that fit into this thesis. This crosslinker, despite being non-cleavable, produces consistent remainder masses that can be searched for during peptide identification and localization. Their addition into a crosslink-compatible search engine should dramatically boost the recovery of crosslinked peptides over attempting to identify the mass of the entre secondary peptide as a PTM. Similar to the aforementioned study, Slavin et al.'s (2020)[133] work in finding the crosslinker mass can also be easily performed using PTM-Shepherd.

In conclusion, I have demonstrated a novel technique for elucidating fragmentation pathways for protein-protein crosslinkers without the need for bespoke tools or synthetic peptides. Rather, monolinks and auto crosslinks facilitate the analysis of protein-protein crosslinked molecules.

4.4 Discussion

I showed that open searches coupled with diagnostic feature extraction have broad applicability across a range of PTMs that extends well beyond previously observed biological ones. With a mix of clever strategies and diagnostic ion mining, I characterized a Cys-specific

chemoproteomic probe and provided a roadmap for others to better understand future probes, finding five new diagnostic ions and two new peptide remainder masses. In the RNA-binding proteome experiment, I was able to recapitulate the results of their hand-crafted pipeline and found additional features that improved the number of PSMs identified by 47.4% over the prior state-of-the-art and by 3.8% over the strategy from the published analysis. Finally, I developed a novel approach to identifying characteristic fragmentation patterns for crosslinkers, patterns which can be deduced directly from the data in a straightforward manner rather than relying on heuristics or synthetic peptides.

4.5 Data availability

Mass spectrometry data generated for the chemical probe experiments has been uploaded to ProteomeXchange under the identifiers PXD028853 and PXD030737[131]. RNA crosslinking data was retrieved from the ProteomeXchange repository PXD023401[33]. Protein crosslinking data was retrieved from the ProteomeXchange repository PXD020704[133].

4.6 Acknowledgements

This study was supported by CA140044 Proteogenomics of Cancer Training Program. I would like to thank Jong Woo Bae et al. for providing us early access to ProteomeXchange repository PXD023401.

CHAPTER V
The Future of Open Searches

5.1 Conclusions

Many aspects of the proteome are dynamic, governed by post-translational modifications (PTMs) rather than transcription or translation. Understanding the proteome and realizing the promises of the post-genomics era requires a comprehensive understanding of the interplay between a protein's PTMs and the rest of the cell. Mass spectrometry-based proteomics coupled to open searches is currently the only method available for the comprehensive analysis of post-translational modifications (PTMs) at proteome scale. Unfortunately, open searches can have lower sensitivity than traditional searches for known modifications. Plus, the interpretation of open searches is a fraught and opaque exercise. In this dissertation, I presented computational methods to characterize open search-derived PTMs, aiding researchers in interpreting and utilizing open search results.

In Chapter Two of the dissertation, I developed a software platform called PTM-Shepherd for the analysis and summarization of open search results. PTM-Shepherd provides researchers with intuitive information about their open search-derived PTMs, including their identities, prevalence, localization profiles, and changes in retention time and spectral similarity. We then applied PTM-Shepherd to several different scenarios to demonstrate its utility. First, we examined chemical artifacts produced by FFPE, a common tissue preservation method, wherein PTM-Shepherd was able to distinguish between two isobaric PTMs to correct previous recommendations about which of these two modifications should be included when searching FFPE data. Second, we examined a frequently reprocessed, high-quality dataset and found their sample processing to contain an experimental error that introduced a huge variety of Cys-specific PTMs. Third, we examined a "gold standard" synthetic peptide dataset and found a novel PTM by combining PTM secondary metrics. Finally, we examined a large, multi-university

proteomics consortium's data and found massive, site-dependent batch effects in PTM profiles across experiments.

In our analyses we found that this information, while useful, was not sufficient to characterize all PTMs. Labile PTMs in particular could not be localized based on their delta mass and their recovery lagged other PTMs when using shifted ions in database searches. Some cases of isobaric PTMs also remained indistinguishable due to having similar retention time and spectral similarities. In Chapter Three of the dissertation, I extended PTM-Shepherd to characterize the fragmentation spectra of PTMs by identifying diagnostic features specific to individual PTMs. To do this, we calculate three features for each PTM-Shepherd mass shift: diagnostic ions, fragment remainder masses, and peptide remainder masses. The novelty of this technique was grounded in the realization that unmodified peptides can act as an empirically derived null distribution of spectral features and are used as a synthetic control when identifying PTM-specific diagnostic features. We applied our technique to two well-studied sets of PTMs to validate it and showcase how it could be used to discover new ions. Studying glycopeptides labeled with tandem mass tags (TMT) is particularly challenging due to incompatibility in fragmentation energies between glycans and the TMT moiety. We identified many expected ions, but we were able to identify a series of sialic acid-related ions in a data-driven manner that had not been previously described. We then turned our attention to another labile modification, ADP-ribosylation (ADPR). Once again, we recovered many of the known fragmentation patterns as well as new ones that had not previously been described. We hoped these analyses provide a roadmap for other researchers to streamline the traditionally painstaking work of fragmentation analyses in different contexts and for other PTMs. Finally, with our ability to identify diagnostic features in mass, we sought to understand their general characteristics. Importantly, we found that the most abundant diagnostic ions were often the least diagnostic due to co-fragmentation. Diagnostic ions are commonly used to validate the presence of PTMs, so this has major implications for studies utilizing them.

In Chapter Four, I applied PTM-Shepherd to several different synthetic PTMs and show that it is

broadly applicable across proteomics subdisciplines. Covalent chemoproteomics probes have gained popularity in recent years as a method to interrogate druggable protein active sites without crystallization. Rapid development of probes runs counter to thoroughly understanding their fragmentation patterns, an issue we address by completely characterizing the fragmentation of a Cys-specific chemoproteomic probe automatically with PTM-Shepherd. We also studied synthetic RNA analogs used in protein-RNA crosslinking, finding a fragment remainder that increased PSM recovery by a staggering 25% over the existing state of the art. Finally, I looked at protein-protein crosslinkers and showed how to identify a series of crosslinker-specific ions that have the potential to greatly increase recovery in protein crosslinking studies.

5.2 Future directions

5.2.1 Multiple localization

The principle of open searches is that a peptide can be identified based on ions that do not bear the modification. One of the benefits of this logic is that open searches have the capacity to identify peptides containing multiple modifications at once[44]. As others have shown, localization programs and algorithms designed for closed searches produce less than ideal results when confronted with the dramatically expanded localization possibilities required to annotate open search results[75]. Recent advancements in this area have led to tremendous improvements in localization fidelity, but they still fail to take advantage of open search's ability to identify multiple PTMs in tandem[75]. This has gained prominence alongside the study of PTMs--particularly the study of PTM crosstalk[147]. Localizing PTMs derived from open searches is challenging because the search space can include multiple possible modifications, every residue on the peptide, and comparisons between different numbers of modifications. The last presents a currently unanswered problem: when deciding to attribute a mass shift to a single PTM versus multiple PTMs, the configuration with multiple PTMs is overwhelmingly favored. The volume of possible configurations is so much larger when considering multiple PTMs that the chance of a random hit with no evidence is correspondingly inflated. In theory, this could be fixed by controlling for the number of peptide isoforms tested when calculating localization probabilities or potentially by treating each isoform as a hypothesis test and correcting resultant *p*-values

accordingly.

Discriminating between isobaric combinations of modifications is critical in at least three areas: pyrophosphorylation[148], histone proteins[149], and glycoproteomics[142]. Pyrophosphorylation is a PTM involved in a wide range of cellular processes such as bacterial toxicity and viral infection. It is isobaric to two separate phosphorylation (or sulfation) events and will consequently be outcompeted by two separate phosphorylation events when compared head-to-head. This means that many pyrophosphorylation events may be misattributed to multiple phosphorylations when analyzed via open search. Histone proteins suffer from a different problem: a wide variety of possible PTMs and high incidence rate motivates the need for a tool that can assign modifications and their potential combinations in an unbiased manner. Lacking this, the histone code and DNA regulation may prove to be puzzles impossible to crack. It is also common for O-glycans to appear in clusters, or to appear alongside N-glycans, but tools only exist for the analysis of O-glycans alone and even then do not correct for multiple occurrences on the same peptide[142,150]. Ultimately, putting the ability to study these issues into the hands of researchers has the potential to make seismic changes in the biological and medical landscape.

5.2.2 Open search modification rescoring with semi-supervised learning

Like other omics technologies, proteomics has benefitted from the surge in novel deep learning algorithms[151,152]. Spectrum prediction from peptide sequences is one of the most widely used applications for deep learning models in proteomics[153-155]. Predicted spectra are used to better distinguish correct and incorrect peptide matches to spectra, and have been shown to provide substantial benefits, especially when the size of the search space increases[156], as is often the case when searching for modified peptides. Predicting spectra for modified peptides requires building training sets for each modification of interest separately[66,67], however, and as such runs afoul of the "curse of generalizability." Even if this were possible, there is a dearth of spectrum prediction tools that do full spectrum prediction[153] rather than limiting the predictions to just a subset of the most common backbone ions. For modifications that produce intense secondary

series of ions (like neutral losses or remainder masses) or have diagnostic ions, skipping over these misses PTM-specific signals.

Others have taken it upon themselves to develop algorithms that will predict whether a spectrum contains a peptide of a particular modification class *a priori* without even identifying the peptide in the first place. PhoStar[157], for example, predicts whether a peptide contains a phosphorylation event before any search is performed. An ad hoc learning approach[158] (Fig 5-1) to the same problem was recently developed that was shown to increase the number of phosphorylated peptides recovered when included as a feature in Percolator. However, both algorithms can be improved.

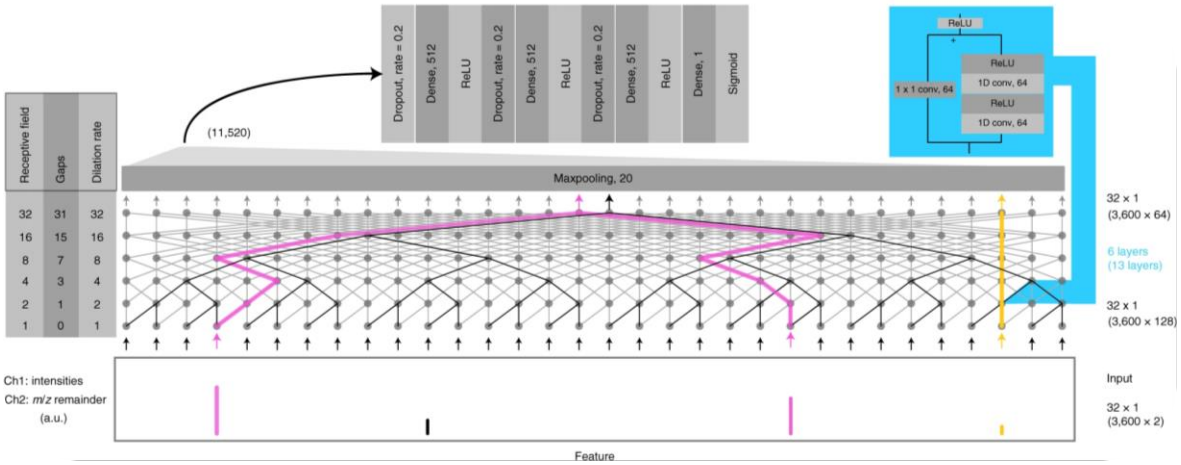

Figure 5-1: Ad hoc learning of peptide fragmentation. Schematic of deep learning architecture for ad hoc learning of peptide fragmentation. Long range associations between peaks are captured as the size of dilation increases with successive layers. Reprinted from *Ad hoc learning of peptide fragmentation from mass spectra enables an interpretable detection of phosphorylated and cross-linked peptides*, Altenburg et al. (2022)[158] under the Creeative Commons license.

PhoStar specifically and solely can detect phosphorylation. It encodes chemical features specific to phosphorylated residues, so it is ungeneralizable[157]. For researchers interested in examining the entire PTM space, this is insufficient. The ad hoc learning approach[158], while generalizable in theory, also has issues. When relying on distances between experimental peaks as it does, the input space becomes polluted with noise. Where ***false*** signifies an ion not attributable to a PTM

and ***true*** signifies an ion that is, pairwise computations between peaks in the spectrum can produce three classes of features: ***false-false***, ***false-true***, and ***true-true***. Only ***true-true*** features can encode useful information. Because only a portion of the peaks in a spectrum are ***true***, the signal-to-noise ratio drops off exponentially. Mathematically, this can be represented as $(true / total)^2$. By reframing pairwise distances as being between the theoretical peptide peaks and the experimental spectrum peaks, as we do in Chapter Three, the signal-to-noise ratio becomes $(true / total)$. An algorithm implementing pairwise calculations as described should thus be able to identify meaningful features with many fewer data points.

A deep learning architecture resembling ad hoc learning with several tweaks should be able to rescore most modifications from open searches without prior knowledge. First, rather than learning which features are associated with PTMs from a generic training dataset, general peptide fragmentation patterns can be learned from a generic dataset, then PTM-specific features can be identified in a semi-supervised way [54,159] via transfer learning by seeing which features separate a spectrum's top hits from its lower scoring hits. Second, using pairwise distances between theoretical and experimental peaks rather than pairwise experimental peaks should allow learning with much less data and reduce the number of PSMs required for model convergence.

5.2.3 Open search modification rescoring with PTM spectral libraries

Another area where PTMs are underutilized is in spectral library generation[160]. Spectral libraries are used to create a definitive fragmentation pattern for peptides, which allows them to be more confidently identified than by standard experimental-theoretical spectrum correlation scores. One common approach is to do DDA and DIA experiments in tandem[161]. Since DDA experiments have higher identification rates while DIA experiments have better quantitative accuracy, DDA experiments are used to build spectral libraries that increase the number of identifiable peptides in DIA experiments[160]. Another approach is to use pregenerated spectral libraries to rescore DDA search results to improve confidence in PSMs[13]. One key point is that DDA spectral libraries are designed to reidentify the same peptide in another context (a paired DIA experiment,

for example), whereas pregenerated libraries can also be used to rescore PSMs from DDA experiments[13]. These approaches are both peptide-centric in that model spectra are produced for every peptide. However, a PTM-centric approach incorporating aspects from each is likely to provide additional, complementary value.

I propose a PTM-centric spectral library generated from DDA data that can be used to rescore search results in the fashion of pregenerated spectral libraries. As I have shown throughout this dissertation, PTM fragmentation patterns are partially dataset dependent. Peptides follow the same logic, and DDA-generated spectral libraries for peptides are considered the gold standard for the same reason. By creating a list of spectral features—a spectral library—for each PTM, then checking for a correlation between the spectral features of a PSM with a matched mass, we would be creating a modification-agnostic score. The score could then be used for PSM rescoring in the same manner as pregenerated libraries. This would essentially bifurcate a PSM into two components: a measure of peptide confidence and a measure of PTM confidence. PSM hits that might have been filtered out due to low peptide scores would instead be boosted by knowing that their mass shift is likely to be correct. Ultimately, this enables identification of new peptides bearing old modifications.

Importantly, we have already shown that this works in practice. The prime example can be seen in Chapter Four. Diagnostic patterns were mined from the RNA crosslinking dataset before being reapplied in a second pass search. The result was a staggering nearly 50% increase in the number of cross-linked peptides identified. Rather than mining fragmentation patterns from filtered data and incorporating them into second pass search scores, they can be mined from high-confidence PSMs and fed into Percolator immediately following the primary search.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER II

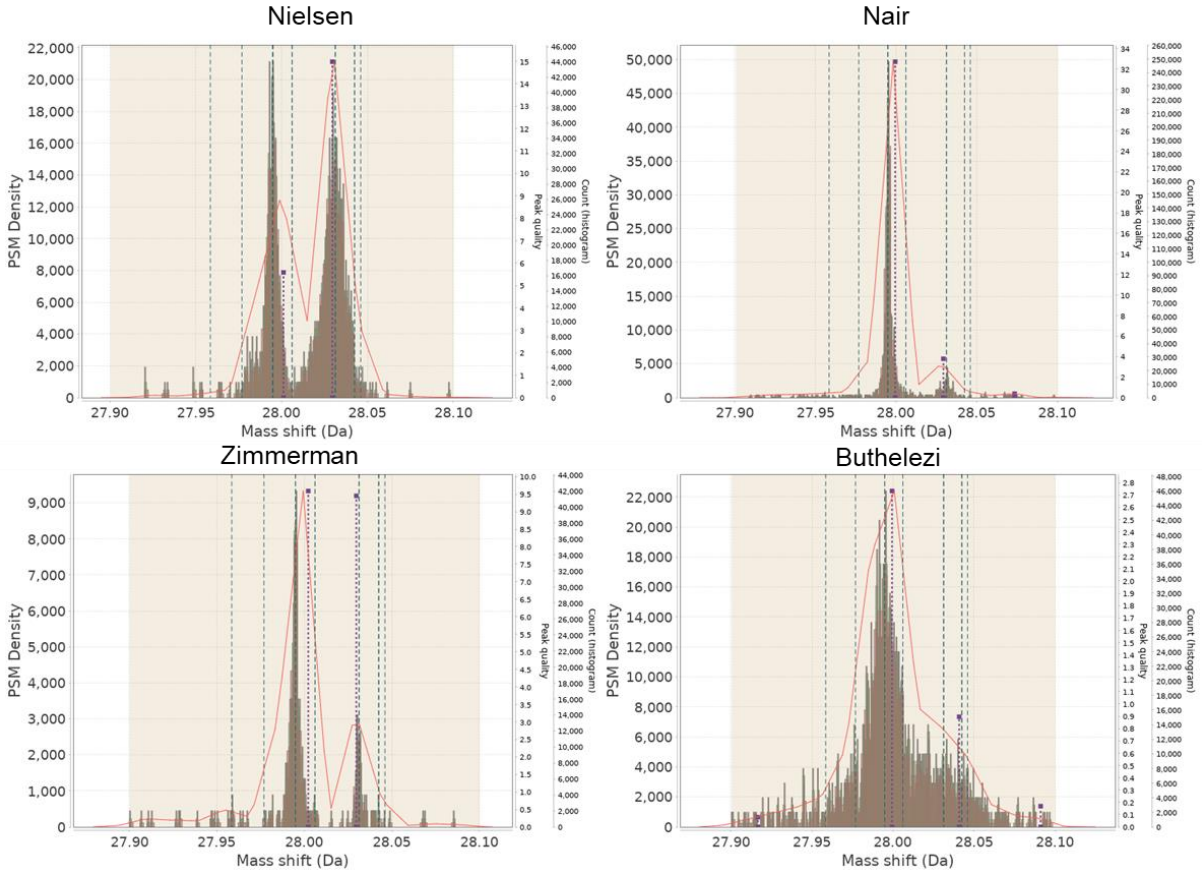

Figure A-1: DeltaMass mass shift profiles around 28 Da for FFPE treated data. Two peaks representing formylation (+27.9949 Da) and di-methylation (+28.0314 Da) are clearly visible for the Nielsen, Nair, and Zimmerman datasets. For the lower resolution Buthelezi dataset, the composite gaussian mixture (red outline) is still composed of two distributions centered at the expected values (purple squares).

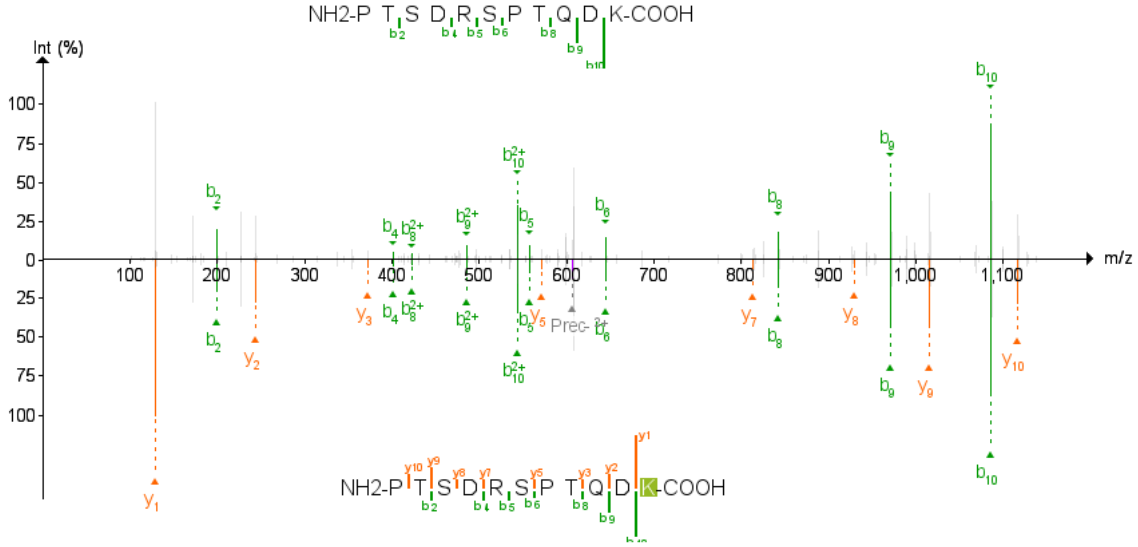

Figure A-2: PDV view of spectrum with Lys water loss. 01717a_BE3-TUM_second_pool33_01_01-3xHCD-1h-R1.6388.6388.2. Placing the loss of water (-18.0106 Da) on C-terminal K aligns the *y*-ion series. All discriminating ions to confirm localization are visible.

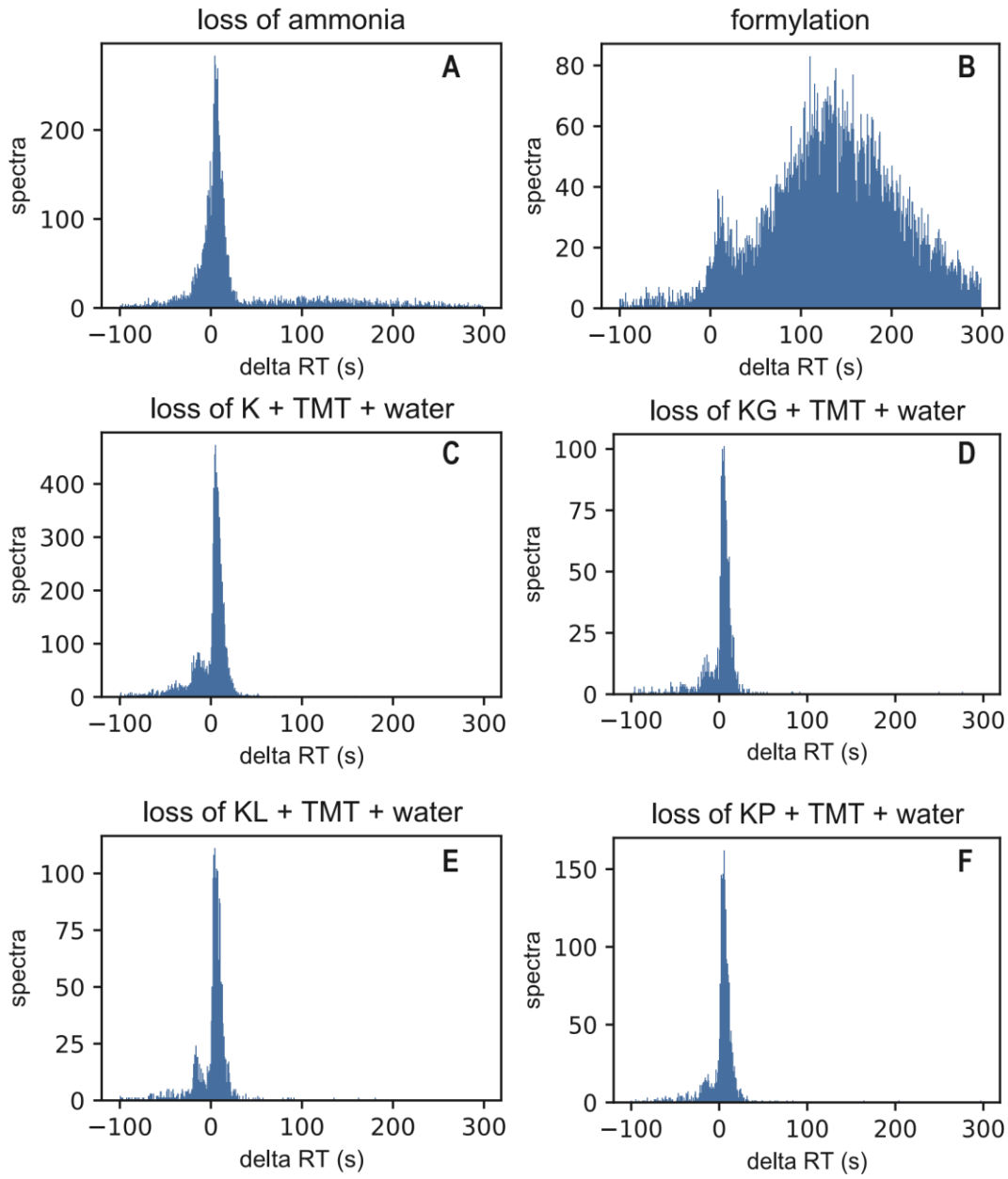

Figure A-3: Retention time profiles of presumed whole-residue in-source losses from all combined CPTAC reference samples. Retention time profiles for losses of ammonia (A) and formylation (B) are shown for comparison. Losses of ammonia can be both in-source or pre-elution; the distinct peak at delta RT = 0 is characteristic of an in-source loss. We observe that pre-elution modifications such as formylation (B) present with broader profiles and have an effect on retention time. All four presumed whole-residue in-source losses exhibit narrow peaks near delta RT = 0, indicating that they are occurring in-source.

Tables A-1, A-2, A-3, and A-4 can be found at 10.5281/zenodo.6975737.

Table A-1: PTM-Shepherd results for reanalysis of four Tabb et al. (2020) datasets. See "Chapter_2/Table_A-1.xlsx".

- global.profile.tsv table from PTM-Shepherd containing summary statistics across all four datasets

Table A-2: PTM-Shepherd results for reanalysis of Bekker-Jensen et al. (2017) dataset. See "Chapter_2/Table_A-2.xlsx".

- **A:** global.profile.tsv table from PTM-Shepherd containing summary statistics across the dataset
- **B:** global.modsummary.tsv table from PTM-Shepherd containing abundance information for annotated modifications rather than mass shifts

Table A-3: PTM-Shepherd results for reanalysis of Zolg et al. (2017) dataset. See "Chapter_2/Table_A-3.xlsx".

- global.profile.tsv table from PTM-Shepherd containing summary statistics across the dataset

Table A-4: PTM-Shepherd results for reanalysis of CPTAC3 quality control samples. See "Chapter_2/Table_A-4.xlsx".

- **A:** global.profile.tsv table from PTM-Shepherd containing summary statistics across the datasets
- **B:** 01BI.profile.tsv report from PTM-Shepherd containing summary statistics for this dataset
- **C:** 02BI.profile.tsv report from PTM-Shepherd containing summary statistics for this dataset

Appendix B

Supplementary Materials for Chapter III

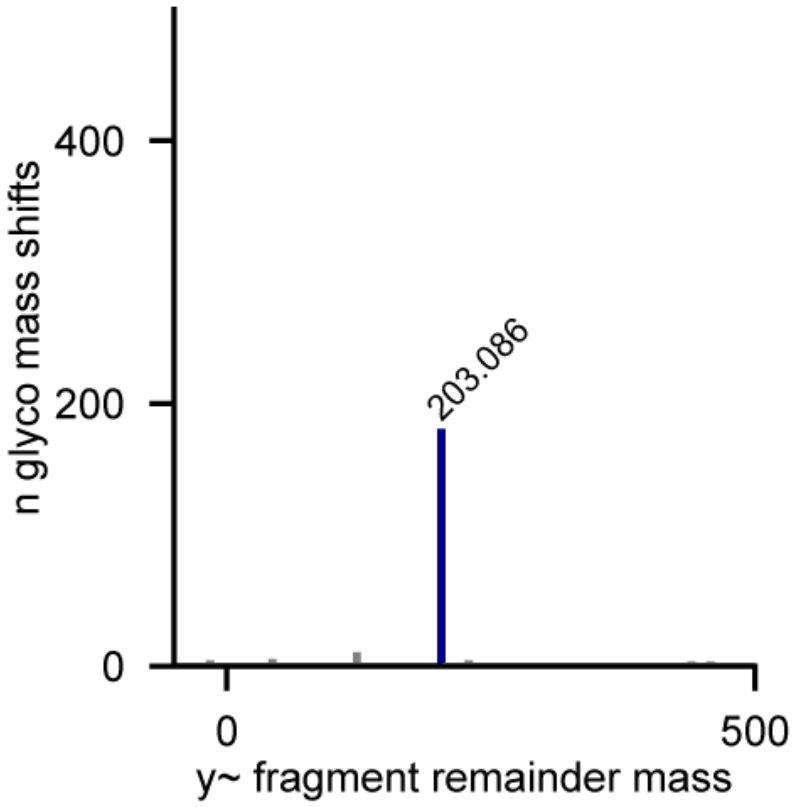

Figure B-1: Histogram of fragment remainder *y*-ions across mass shifts. Like *b*-ions, *y*-ions also show the expected fragment remainder mass pattern.

```
for delta mass bin do
    initialize feature histograms;
    for spectral file do
        parse spectral file;
        add spectral features to histograms;
        close spectral file;
    end
    process feature histograms;
end
```

Algorithm 1: Requires loading spectral files for every delta mass bin and doesn't allow dynamic histogram sizing.

Figure B-2: Algorithm for processing delta mass bins individually with reloading/reprocessing every spectral file for every mass bin.

```
for delta mass bin do
    initialize feature histograms;
end
for spectral file do
    parse spectral file;
    add spectral features to histograms;
    close spectral file;
end
for delta mass bin do
    process feature histograms;
end
```

Algorithm 2: Requires keeping every feature histogram in memory and doesn't allow dynamic histogram sizing.

Figure B-3: Single-pass algorithm for processing delta mass bins in parallel that requires storing every mass bin in memory at once.

initialize intermediate file;
for *spectral file* **do**
    parse spectral file;
    add spectral features to intermediate file;
    close spectral file;
end
index intermediate file;
for *delta mass bin* **do**
    find appropriate histogram boundaries;;
    parse cached features for histogram sizing;
    initialize feature histograms;
    add spectral features to histograms;
    process feature histograms;
end
close intermediate file;

Algorithm 3: Minimizes RAM usage and spectral file parsing but hits an IO bottleneck on the intermediate file.

Figure B-4: Two-pass algorithm where spectral features are precomputed and selectively re-accessed for each mass shift bin.

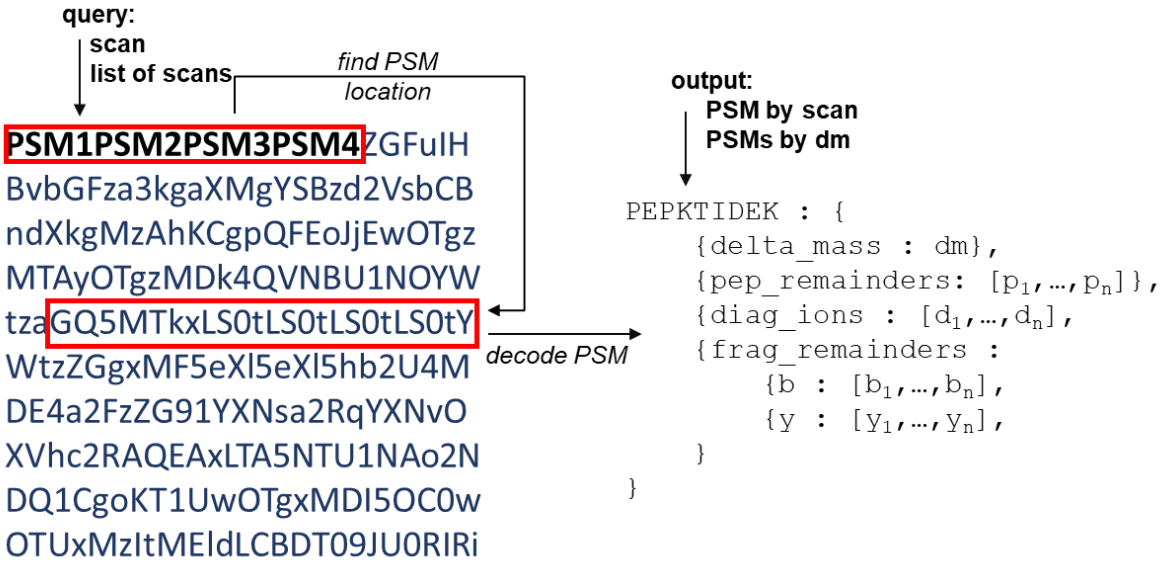

Figure B-5: Implementation of fast-access indexed binary spectral feature intermediate file for two-pass feature extraction.

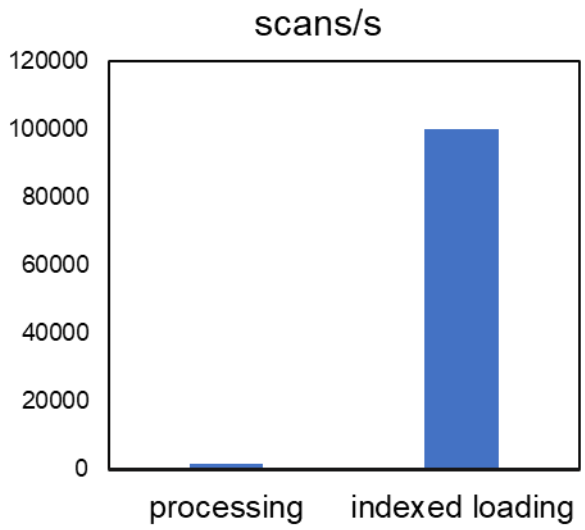

Figure B-6: Speed comparison for indexed loading vs. reprocessing spectral features at each stage where they are needed.

Tables B-1, B-2, and B-3 can be found at 10.5281/zenodo.6975737.

Table B-1: PTM-Shepherd diagnostic features for analysis of CCRCC phospho-glyco data. See "Chapter_3/Table_B-1.xlsx".

- global.diagmine.tsv table from PTM-Shepherd containing diagnostic feature information

Table B-2: Spearman correlation matrix between ion intensities across spectra for the analysis of CPTAC CCRCC phospho-glyco data. See "Chapter_3/Table_B-2.xlsx".

- correlation matrix of ion intensities extracted using PTM-Shepherd glyco mode

Table B-3: PTM-Shepherd diagnostic features for the analysis of ADPR data. See "Chapter_3/Table_B-3.xlsx".

- **A:** global.diagmine.tsv table from PTM-Shepherd containing diagnostic feature information for the HeLa dataset
- **B:** global.diagmine.tsv table from PTM-Shepherd containing diagnostic feature information for the mouse dataset

Appendix C

Supplementary Materials for Chapter IV

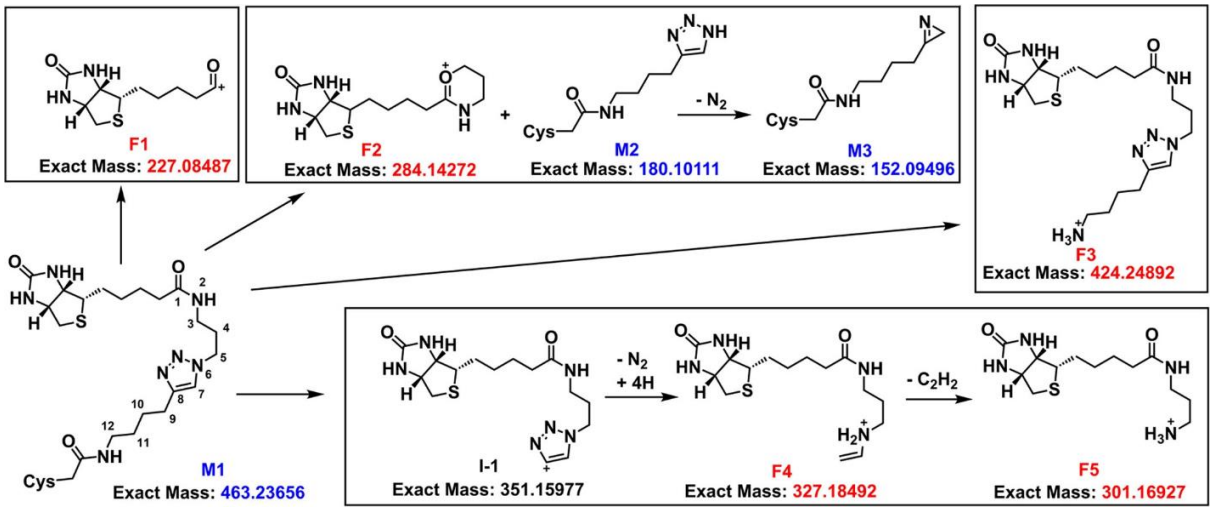

Figure C-1: Schematic for proposed fragmentation patterns of cysteine chemical probe. Reprinted with permission from Yan et al. (2022)[131]. Copyright © 2022 American Chemical Society.

Tables C-1 and C-2 can be found at 10.5281/zenodo.6975737.

Table C-1: PTM-Shepherd diagnostic features for the analysis of a cysteine probe. See "Chapter_4/C1_cysprobe_diagnostic.xlsx".

- global.diagmine.tsv table from PTM-Shepherd containing diagnostic feature information

Table C-2: PTM-Shepherd diagnostic features for the analysis of RNA crosslinking data. See "Chapter_4/C2_rnax_diagnostic.xlsx".

- correlation matrix of ion intensities extracted using PTM-Shepherd glyco mode

Appendix D
PTM-Shepherd Output Files

global.profile.tsv

global.profile.tsv reports the most prominent features from PTM-Shepherd analysis of mass shifts observed from FDR-filtered open search results. Each row corresponds to a different detected mass shift, thus not all PSMs will be represented in this table. Please note that mass shifts are annotated based on UniMod mapping, thus they are not definitive chemical identities and should be used as a starting point along with localization and amino acid enrichment information. Unless otherwise indicated, values are summed from all datasets in the analysis. Column contents are listed below.

- **peak_apex** apex of the detected delta mass peak (in Da)
- **peak_lower** lower bound of the detected peak (Da), determined by precursor tolerance or the detection of an adjacent peak
- **peak_upper** upper bound of the detected peak (Da), determined by precursor tolerance or the detection of an adjacent peak
- **PSMs** the number of PSMs contained within the peak boundary (bin), reported for each dataset if multiple datasets are used as input
- **peak_signal** relative measure of peak prominence/quality. In noisy regions of the delta mass histogram, values are penalized
- **percent_also_in_unmodified** the percentage of PSMs in this mass bin with a corresponding PSM in the unmodified bin
- **mapped_mass_1** primary modification annotation derived from Unimod, all isobaric modifications listed and separated by "/"
- **mapped_mass_2** if the delta mass peak is a combination of two masses, a second modification annotation is listed here. As with mapped_mass_1, all isobaric modifications are listed and separated by "/"

- **similarity** MS/MS spectral similarity of modified peptides compared to their unmodified counterparts. When multiple modified-unmodified comparisons are done for a single peptide, these cosine similarity scores are averaged for the peptide. The peptide scores are then averaged across all peptides in the mass shift bin. These comparisons are only done for peptides of the same charge state.
- **rt_shift** retention time shift comparing modified peptides to their unmodified counterparts. When multiple modified-unmodified comparisons are done for a single peptide, the retention time shifts are averaged for the peptide. The peptide shifts are then averaged across all peptides in the mass shift bin. Individual comparisons are only done for peptides in the same LC-MS run. Units are usually seconds but can vary by instrument type
- **int_log2fc** log2 fold-change of average intensity for matched shifted/unshifted peptides, computed as described above. Peptides affect by sample preparation artifacts tend to be lower abundance than their unshifted counterparts, thus this value will be low in these cases
- **localized_PSMs** number of PSMs for this delta mass that showed at least one additional matched ion when the mass shift is placed on a residue
- **n-term_localization_rate** percentage of PSMs with an uninterrupted string of localized residues from the N-terminus. This is calculated differently from other enrichment scores due to the difference in assumptions underlying N-terminal and residue-specific localization, so these values cannot be directly compared to the amino acid enrichment scores.
- **AA1** amino acid/residue most enriched (most likely to harbor the mass shift) compared to other residues
- **AA1_enrichment_score** equivalent to the odds the delta mass is localized to AA1 compared to other residues
- **AA1_psm_count** weighted number of PSMs where the mass shift localized to AA1. Shifts localizing to multiple residues are divided by the number of localized residues in the spectra, so this is an estimated number of PSMs localized to a particular residue

- (same enrichment_score and psm_count columns for AA2 and AA3 if multiple amino acids are likely to harbor the mass shift)
- **[experiment]_PSMs** number of PSMs with a mass shift in this bin
- **[experiment]_percent_PSMs** number of PSMs from the previous column as a percentage of total PSMs
- **[experiment]_peptides** number of unique peptide sequences with a mass shift in this bin
- **[experiment]_percent_also_in_unmodified** percentage of peptide sequences with a mass shift in this bin that are also found in the zero mass shift bin

global.modsummary.tsv

global.modsummary.tsv is a modification-centric table generated from PTM-Shepherd summarization of mass shifts observed in open search workflows. Please note that mass shifts are annotated based on UniMod mapping, thus they are not definitive chemical identities and should be used as a starting point along with localization and amino acid enrichment information. Contents of each column are listed below.

- **Modification** Name/annotation of the modification (as found in the global.profile.tsv file)
- **Theoretical Mass Shift** The theoretical mass (in Da) of the modification from Unimod if annotated, or the peak apex of an unannotated modification
- **[experiment]_PSMs** Number of PSMs with the modification, including any row from the global.profile.tsv file where the modification appears (e.g., a ‘Methylation’ entry in the will include PSMs corresponding to both ‘Methylation’ and ‘Methylation + First isotopic peak’)
- **[experiment]_percent_PSMs** The number of PSMs from the previous column as a percentage of the total PSMs

global.diagmine.tsv

global.diagmine.tsv is a mass shift-centric table that contains the diagnostic features identified for every mass shift. Please note that only mass shifts with diagnostic features detected are reported in the table. Contents of each column are listen below.

- **peak_apex** This field contains the apex of the detected MS1 peak (Da) present in the global.profile.tsv file from PTM-Shepherd.
- **mod_annotation** This field contains the mass shift annotations present in the global.profile.tsv file from PTM-Shepherd. When a mass shift is found to be the combination of two mass shifts, the "Potential Modification 1" and "Potential Modification 2" columns are merged with a semicolon.
- **type** This field can take one of several values. "diagnostic" refers to diagnostic ions, the ions that can be located directly in the spectrum. "peptide" refers to peptide remainder masses, mass shifts that indicate an ion's presence at a particular distance from an unshifted peptide. Six other values are possible based on parameter setting, each corresponding to one of the major ion series.
- **mass** This field contains the mass of the diagnostic feature. Peptide and fragment remainder masses will have the mass shift away from the theoretical ion. Diagnostic ions will have the m/z of the observed ion, so a non-neutral mass.
- **delta_mod_mass** This field contains the mass that was lost from the original mass shift to arrive at the remainder mass. (Note: only present for peptide and fragment remainder masses.)
- **remainder_propensity** This field contains the average percentage of ions from a particular series that are shifted. For example, a peptide capable of producing 10 *b*-ions with 2 ions identified ions shifted by the remainder mass and 2 identified ions unshifted would have a propensity of 50%. The propensity score for every representative PSM within a mass shift bin is averaged. (Note: only present for fragment remainder masses.)
- **percent_mod** This field contains the percentage of representative mass shifted PSMs that contain the ion at any intensity.
- **percent_unmod** This field contains the percentage of representative unshifted PSMs that contain the ion at any intensity.
- **avg_intensity_mod** This field contains the average intensity of the ion among representative mass shifted PSMs where the ion is present. To calculate the average across all representative mass shifted spectra, calculate (avg_intensity_mod *

percent_mod / 100). Because multiple ions can be matched for fragment remainder ions, this contains the average of the summed intensity of matched ions for each representative PSM.

- **avg_intensity_unmod** This field contains the average intensity of the ions among representative unshifted PSMs where the ion is present. To calculate the average across all representative mass shifted spectra, calculate (avg_intensity_mod * percent_mod / 100). Because multiple ions can be matched for fragment remainder ions, this contains the average of the summed intensity of matched ions for each representative PSM.
- **intensity_fold_change** This field contains the fold change in intensity when comparing the modified to unmodified peptides. This uses intensity across all spectra and can be calculated via (avg_intensity_mod * percent_mod) / (avg_intensity_unmod * percent_unmod).
- **auc** This column contains the AUC-ROC statistic for the intensity-based classification of this ion. It is calculated from the U statistic from the Mann-Whitney U Test. This statistic adjusts the two groups such that they are assumed to be of equal size.

REFERENCES

1 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

2 Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534-1536 (2004).

3 Smith, L. M. & Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nature methods* **10**, 186-187 (2013).

4 Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21**, 255-261 (2003).

5 Deribe, Y. L., Pawson, T. & Dikic, I. Post-translational modifications in signal integration. *Nature structural & molecular biology* **17**, 666-672 (2010).

6 Helenius, A. & Aebi, M. Roles of N-linked glycans in the endoplasmic reticulum. *Annual review of biochemistry* **73**, 1019-1049 (2004).

7 Bradshaw, R. A., Brickey, W. W. & Walker, K. W. N-terminal processing: the methionine aminopeptidase and Nα-acetyl transferase families. *Trends in biochemical sciences* **23**, 263-267 (1998).

8 Tasaki, T., Sriram, S. M., Park, K. S. & Kwon, Y. T. The N-end rule pathway. *Annual review of biochemistry* **81**, 261 (2012).

9 Nardozzi, J. D., Lott, K. & Cingolani, G. Phosphorylation meets nuclear import: a review. *Cell Communication and Signaling* **8**, 1-17 (2010).

10 Singh, V. *et al.* Phosphorylation: implications in cancer. *The protein journal* **36**, 1-6 (2017).

11 Clague, M. J. & Urbé, S. Ubiquitin: same molecule, different degradation pathways. *Cell* **143**, 682-685 (2010).

12 Freudenmann, L. K., Marcu, A. & Stevanović, S. Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. *Immunology* **154**, 331-345 (2018).

13 Kacen, A. *et al.* Uncovering the modified immunopeptidome reveals insights into principles of PTM-driven antigenicity. *bioRxiv* (2021).

14 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 Proteomics* **17**, 1850-1863 (2018).

15 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).

16 Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059-1061 (2018).

17 Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell. Proteomics* **11** (2012). https://doi.org:https://doi.org/10.1074/mcp.M111.010199

18 Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research* **40**, D261-D270 (2012).

19 UniProt: the universal protein knowledgebase. *Nucleic acids research* **45**, D158-D169 (2017).
20 Macek, B. *et al.* Protein post-translational modifications in bacteria. *Nature Reviews Microbiology* **17**, 651-664 (2019).
21 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).
22 Yates III, J. R. The revolution and evolution of shotgun proteomics for large-scale proteome analysis. *Journal of the American Chemical Society* **135**, 1629-1640 (2013).
23 Haag, A. M. Mass analyzers and mass spectrometers. *Modern Proteomics–Sample Preparation, Analysis and Practical Applications*, 157-169 (2016).
24 Hunt, D. F., Yates 3rd, J., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences* **83**, 6233-6237 (1986).
25 Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* **24**, 508-548 (2005).
26 Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989 (1994).
27 Tabb, D. L., Huang, Y., Wysocki, V. H. & Yates, J. R. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry* **76**, 1243-1248 (2004).
28 Dayon, L. & Sanchez, J.-C. in *Quantitative methods in proteomics* 115-127 (Springer, 2012).
29 Backus, K. M. *et al.* Proteome-wide covalent ligand discovery in native biological systems. *Nature* **534**, 570-574 (2016).
30 Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **73**, 2092-2123 (2010). https://doi.org:10.1016/j.jprot.2010.08.009
31 Annesley, T. M. Ion suppression in mass spectrometry. *Clinical chemistry* **49**, 1041-1044 (2003).
32 Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71 (1989).
33 Davis, M. T. *et al.* Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry II. Limitations of complex mixture analyses. *Proteomics* **1**, 108-117 (2001).
34 Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods* **1**, 39-45 (2004).
35 Dančík, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology* **6**, 327-342 (1999).
36 Ma, B. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry* **17**, 2337-2342 (2003).

37 Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22-24 (2013).
38 Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73**, 2092-2123 (2010).
39 Keich, U. & Noble, W. S. On the importance of well-calibrated scores for identifying shotgun proteomics spectra. *Journal of proteome research* **14**, 1147-1160 (2015).
40 Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114 (2014).
41 Han, X., He, L., Xin, L., Shan, B. & Ma, B. PeaksPTM: mass spectrometry-based identification of peptides with unspecified modifications. *Journal of proteome research* **10**, 2930-2936 (2011).
42 Dasari, S. *et al.* TagRecon: high-throughput mutation identification through sequence tagging. *J. Proteome Res.* **9**, 1716-1726 (2010).
43 Yu, F., Li, N. & Yu, W. PIPI: PTM-invariant peptide identification using coding method. *Journal of proteome research* **15**, 4423-4435 (2016).
44 Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **33**, 743-749 (2015).
45 Solntsev, S. K., Shortreed, M. R., Frey, B. L. & Smith, L. M. Enhanced global post-translational modification discovery with MetaMorpheus. *J. Proteome Res.* **17**, 1844-1851 (2018).
46 Mann, M. *et al.* Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends in biotechnology* **20**, 261-268 (2002).
47 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).
48 Liu, M.-Q. *et al.* pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nature communications* **8**, 1-14 (2017).
49 Roushan, A. *et al.* Peak filtering, peak annotation, and wildcard search for glycoproteomics. *Molecular & Cellular Proteomics* **20** (2021).
50 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).
51 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).
52 Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research* **7**, 29-34 (2008).
53 Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383-5392 (2002).

54 Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods* **4**, 923-925 (2007).
55 Fondrie, W. E. & Noble, W. S. mokapot: fast and flexible semisupervised learning for peptide detection. *Journal of proteome research* **20**, 1966-1971 (2021).
56 Shteynberg, D. D. *et al.* PTMProphet: fast and accurate mass modification localization for the trans-proteomic pipeline. *J. Proteome Res.* **18**, 4262-4272 (2019).
57 Savitski, M. M. *et al.* Confident phosphorylation site localization using the Mascot Delta Score. *Molecular & cellular proteomics* **10**, S1-S12 (2011).
58 Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology* **24**, 1285-1292 (2006).
59 Fermin, D., Walmsley, S. J., Gingras, A.-C., Choi, H. & Nesvizhskii, A. I. LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Molecular & Cellular Proteomics* **12**, 3409-3419 (2013).
60 Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature communications* **9**, 1-12 (2018).
61 Declercq, A., Bouwmeester, R., Degroeve, S., Martens, L. & Gabriels, R. MS2Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *bioRxiv* (2021).
62 Giese, S. H., Sinn, L. R., Wegner, F. & Rappsilber, J. Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. *Nature communications* **12**, 1-11 (2021).
63 Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods* **18**, 1363-1369 (2021).
64 Steckel, A. & Schlosser, G. Citrulline effect is a characteristic feature of deiminated peptides in tandem mass spectrometry. *Journal of The American Society for Mass Spectrometry* **30**, 1586-1591 (2019).
65 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).
66 Gabriel, W. *et al.* Prosit-TMT: Deep Learning Boosts Identification of TMT-Labeled Peptides. *Analytical Chemistry* (2022).
67 Steger, M. *et al.* Time-resolved in vivo ubiquitinome profiling by DIA-MS reveals USP7 targets on a proteome-wide scale. *Nature communications* **12**, 1-13 (2021).
68 Eng, J. K., Searle, B. C., Clauser, K. R. & Tabb, D. L. A face in the crowd: recognizing peptides through database search. *Mol. Cell. Proteomics* **10**, R111-009522 (2011).
69 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).
70 Nesvizhskii, A. I. *et al.* Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-

translational modifications, sequence polymorphisms, and novel peptides. *Molecular & cellular proteomics : MCP* **5**, 652-670 (2006). https://doi.org:10.1074/mcp.M500319-MCP200

71 Ning, K., Fermin, D. & Nesvizhskii, A. I. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* **10**, 2712-2718 (2010). https://doi.org:10.1002/pmic.200900473

72 Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature biotechnology* **36**, 1059-1061 (2018).

73 Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Molecular & Cellular Proteomics* **11**, M111.010199 (2012). https://doi.org:https://doi.org/10.1074/mcp.M111.010199

74 Avtonomov, D. M., Kong, A. & Nesvizhskii, A. I. DeltaMass: Automated Detection and Visualization of Mass Shifts in Proteomic Open-Search Results. *J. Proteome Res.* **18**, 715-720 (2018).

75 An, Z. *et al.* PTMiner: Localization and Quality Control of Protein Modifications Detected in an Open Search and Its Application to Comprehensive Post-translational Modification Characterization in Human Proteome. *Mol. Cell. Proteomics* **18**, 391-405 (2019).

76 Shteynberg, D. D. *et al.* PTMProphet: Fast and accurate mass modification localization for the trans-proteomic pipeline. *Journal of proteome research* **18**, 4262-4272 (2019).

77 da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods* **17**, 869-870 (2020). https://doi.org:10.1038/s41592-020-0912-y

78 Tabb, D. L. *et al.* Open search unveils modification patterns in formalin-fixed, paraffin-embedded thermo HCD and SCIEX TripleTOF shotgun proteomes. *Int. J. Mass Spectrom.*, 116266 (2019).

79 Nielsen, N. S., Poulsen, E. T., Klintworth, G. K. & Enghild, J. J. Insight into the protein composition of immunoglobulin light chain deposits of eyelid, orbital and conjunctival amyloidosis. *J. Proteomics Bioinform.*, Suppl 8: 002 (2014). https://doi.org:10.4172/0974-276X.S8-002

80 Nair, O. *Profiling medulloblastoma and juvenile pilocytic astrocytoma brain tumours in a South African paediatric cohort*, University of Cape Town, (2017).

81 Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259-262 (2017).

82 Edwards, N. J. *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **14**, 2707-2713 (2015).

83 Mertins, P. *et al.* Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. *Nat. Protoc.* **13**, 1632-1661 (2018).

84 Yu, F. *et al.* Identification of modified peptides using localization-aware open search. *Nat Commun* **11**, 4065 (2020). https://doi.org:10.1038/s41467-020-17921-y

85 Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392 (2002). https://doi.org:10.1021/ac025747h

86 Chang, H. Y. *et al.* Crystal-C: A Computational Tool for Refinement of Open Search Results. *J Proteome Res* **19**, 2511-2515 (2020). https://doi.org:10.1021/acs.jproteome.0c00119
87 Zhang, Y. *et al.* Unrestricted modification search reveals lysine methylation as major modification induced by tissue formalin fixation and paraffin embedding. *Proteomics* **15**, 2568-2579 (2015). https://doi.org:10.1002/pmic.201400454
88 Etherington, D. J. & Sims., T. J. Detection and estimation of collagen. *Journal of the Science of Food and Agriculture* **32**, 539-546 (1981).
89 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). https://doi.org:10.1038/s41592-020-0967-9
90 Chung, H. S., Wang, S.-B., Venkatraman, V., Murray, C. I. & Van Eyk, J. E. Cysteine oxidative posttranslational modifications: emerging regulation in the cardiovascular system. *Circ. Res.* **112**, 382-392 (2013).
91 Sechi, S. & Chait, B. T. Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Anal. Chem.* **70**, 5150-5158 (1998).
92 Schnatbaum, K., Zolg, D., Wenschuh, H. & Reimer, U. Fast and accurate determination of cysteine reduction and alkylation efficacy in proteomics workflows. *JPT Application Note* (2016).
93 Bekker-Jensen, D. B. *et al.* An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell systems* **4**, 587-599 (2017).
94 Metz, B. *et al.* Identification of formaldehyde-induced modifications in proteins reactions with model peptides. *J. Biol. Chem.* **279**, 6235-6243 (2004).
95 Cordero, M. M., Houser, J. J. & Wesdemiotis, C. The neutral products formed during backbone fragmentations of protonated peptides in tandem mass spectrometry. *Anal. Chem.* **65**, 1594-1601 (1993).
96 Savitski, M. M., Kjeldsen, F., Nielsen, M. L. & Zubarev, R. A. Relative specificities of water and ammonia losses from backbone fragments in collision-activated dissociation. *J. Proteome Res.* **6**, 2669-2673 (2007).
97 Kumar, A. & Bachhawat, A. K. Pyroglutamic acid: throwing light on a lightly studied metabolite. *Curr. Sci.* **102**, 288-297 (2012).
98 Dick, L. W., Jr., Kim, C., Qiu, D. & Cheng, K.-C. Determination of the origin of the N-terminal pyro-glutamate variation in monoclonal antibodies using model peptides. *Biotechnol. Bioeng.* **97**, 544-553 (2007).
99 Reimer, J. *et al.* Effect of cyclization of N-terminal glutamine and carbamidomethyl-cysteine (residues) on the chromatographic behavior of peptides in reversed-phase chromatography. *J. Chromatogr. A* **1218**, 5101-5107 (2011).
100 Sun, S. *et al.* Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. *J. Proteome Res.* **7**, 202-208 (2008).
101 Chelius, D. *et al.* Formation of pyroglutamic acid from N-terminal glutamic acid in immunoglobulin gamma antibodies. *Anal. Chem.* **78**, 2370-2376 (2006).
102 He, W., Tao, Y. & Wang, X. Functional Polyamides: A Sustainable Access via Lysine Cyclization and Organocatalytic Ring-Opening Polymerization. *Macromolecules* **51**, 8248-8257 (2018).

103 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733-739 (2010).
104 Rudnick, P. A. *et al.* Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics* **9**, 225-241 (2010).
105 Wang, X. *et al.* QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics. *Anal. Chem.* **86**, 2497-2509 (2014).
106 Paulovich, A. G. *et al.* Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* **9**, 242-254 (2010).
107 Tabb, D. L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography- tandem mass spectrometry. *J. Proteome Res.* **9**, 761-776 (2009).
108 Clark, D. J. *et al.* Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **179**, 964-983.e931 (2019). https://doi.org:10.1016/j.cell.2019.10.007
109 Dou, Y. *et al.* Proteogenomic Characterization of Endometrial Carcinoma. *Cell* **180**, 729-748.e726 (2020). https://doi.org:10.1016/j.cell.2020.01.026
110 Gillette, M. A. *et al.* Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* **182**, 200-225.e235 (2020). https://doi.org:10.1016/j.cell.2020.06.013
111 Lenco, J., Khalikova, M. A. & Švec, F. Dissolving peptides in 0.1% formic acid brings risk of artificial formylation. *J. Proteome Res.* **19**, 993-999 (2020).
112 Palmisano, G. *et al.* A novel method for the simultaneous enrichment, identification, and quantification of phosphopeptides and sialylated glycopeptides applied to a temporal profile of mouse brain development. *Molecular & cellular proteomics* **11**, 1191-1202 (2012).
113 Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* **27**, 861-874 (2006).
114 Post, H. *et al.* Robust, Sensitive, and Automated Phosphopeptide Enrichment Optimized for Low Sample Amounts Applied to Primary Hippocampal Neurons. *J. Proteome Res.* **16**, 728-737 (2017).
115 Yeom, J., Ju, S., Choi, Y., Paek, E. & Lee, C. Comprehensive analysis of human protein N-termini enables assessment of various protein forms. *Sci. Rep.* **7**, 6599 (2017).
116 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).
117 Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9**, 4632-4641 (2009).
118 Drewes, G. & Knapp, S. Chemoproteomics and chemical probes for target discovery. *Trends Biotechnol.* **36**, 1275-1286 (2018).
119 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).
120 Kielkowski, P. *et al.* A Pronucleotide Probe for Live-Cell Imaging of Protein AMPylation. *Chembiochem* **21**, 1285-1287 (2020).
121 Trendel, J. *et al.* The human RNA-binding proteome and its dynamics during translational arrest. *Cell* **176**, 391-403. e319 (2019).

122 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).
123 Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141 (2010).
124 Ule, J., Jensen, K., Mele, A. & Darnell, R. B. CLIP: a method for identifying protein–RNA interaction sites in living cells. *Methods* **37**, 376-386 (2005).
125 Wippel, H. H., Chavez, J. D., Tang, X. & Bruce, J. E. Quantitative interactome analysis with chemical cross-linking and mass spectrometry. *Current Opinion in Chemical Biology* (2021).
126 Iacobucci, C., Götze, M. & Sinz, A. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. *Current opinion in biotechnology* **63**, 48-53 (2020).
127 Sohn, C. H. *et al.* Designer reagents for mass spectrometry-based proteomics: clickable cross-linkers for elucidation of protein structures and interactions. *Analytical chemistry* **84**, 2662-2669 (2012).
128 Chowdhury, S. M. *et al.* Identification of cross-linked peptides after click-based enrichment using sequential collision-induced dissociation and electron transfer dissociation tandem mass spectrometry. *Analytical chemistry* **81**, 5524-5532 (2009).
129 Sinz, A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. *Mass spectrometry reviews* **25**, 663-682 (2006).
130 da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869-870 (2020).
131 Yan, T. *et al.* Enhancing Cysteine Chemoproteomic Coverage Through Systematic Assessment of Click Chemistry Product Fragmentation. *Anal. Chem.* **94**, 3800-3810 (2022).
132 Magrane, M. & UniProt, C. UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, bar009 (2011).
133 Slavin, M., Tayri-Wilk, T., Milhem, H. & Kalisman, N. Open Search Strategy for Inferring the Masses of Cross-Link Adducts on Proteins. *Analytical chemistry* **92**, 15899-15907 (2020).
134 Bantscheff, M. & Drewes, G. Chemoproteomic approaches to drug target identification and drug profiling. *Bioorganic & medicinal chemistry* **20**, 1973-1978 (2012).
135 Havelund, J. F. *et al.* A biotin enrichment strategy identifies novel carbonylated amino acids in proteins from human plasma. *Journal of Proteomics* **156**, 40-51 (2017).
136 Udeshi, N. D. *et al.* Antibodies to biotin enable large-scale detection of biotinylation sites on proteins. *Nature methods* **14**, 1167-1170 (2017).
137 Renuse, S. *et al.* Signature fragment ions of biotinylated peptides. *Journal of the American Society for Mass Spectrometry* **31**, 394-404 (2020).
138 He, J.-X. *et al.* pChem: a modification-centric assessment tool for the performance of chemoproteomic probes. *bioRxiv* (2021).
139 Backus, K. M. Applications of reactive cysteine profiling. *Activity-Based Protein Profiling*, 375-417 (2018).
140 Yan, T. *et al.* SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome. *Chembiochem* **22**, 1841-1851 (2021).

141 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**, 1794-1805 (2011).
142 Bagdonaite, I. *et al.* Glycoproteomics. *Nature Reviews Methods Primers* **2**, 1-29 (2022).
143 Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440 (2015).
144 Ewing, R. M. *et al.* Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular systems biology* **3**, 89 (2007).
145 Hoopmann, M. R. *et al.* Kojak: efficient analysis of chemically cross-linked protein complexes. *Journal of proteome research* **14**, 2190-2198 (2015).
146 Steigenberger, B., Schiller, H. B., Pieters, R. J. & Scheltema, R. A. Finding and using diagnostic ions in collision induced crosslinked peptide fragmentation spectra. *International Journal of Mass Spectrometry* **444**, 116184 (2019).
147 Venne, A. S., Kollipara, L. & Zahedi, R. P. The next level of complexity: crosstalk of posttranslational modifications. *Proteomics* **14**, 513-524 (2014).
148 Penkert, M. *et al.* Electron transfer/Higher energy collisional dissociation of doubly charged peptide ions: identification of labile protein phosphorylations. *Journal of The American Society for Mass Spectrometry* **30**, 1578-1585 (2019).
149 Rothbart, S. B. & Strahl, B. D. Interpreting the language of histone and DNA modifications. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1839**, 627-643 (2014).
150 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).
151 Meyer, J. G. Deep learning neural network tools for proteomics. *Cell Reports Methods* **1**, 100003 (2021).
152 Wen, B. *et al.* Deep learning in proteomics. *Proteomics* **20**, 1900335 (2020).
153 Liu, K., Li, S., Wang, L., Ye, Y. & Tang, H. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. *Analytical chemistry* **92**, 4275-4283 (2020).
154 Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods* **16**, 509-518 (2019).
155 Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature methods* **16**, 519-525 (2019).
156 Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nature communications* **12**, 1-12 (2021).
157 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).
158 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).
159 Halloran, J. T., Urban, G., Rocke, D. & Baldi, P. Deep Semi-Supervised Learning Improves Universal Peptide Identification of Shotgun Proteomics Data. *bioRxiv* (2020).
160 Craig, R., Cortens, J. P. & Beavis, R. C. The use of proteotypic peptide libraries for protein identification. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry* **19**, 1844-1850 (2005).

161 Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass spectrometry reviews* **39**, 229-244 (2020).