The Confounding Effects of Concurrent Metaphosphate Neutral Loss and Peptide Backbone Fragmentation on Phosphorylation Site Localization

The Confounding Effects of Concurrent Metaphosphate Neutral Loss and Peptide Backbone Fragmentation on Phosphorylation Site Localization

Tyler Morris Levy

A Thesis in the Field of Bioinformatics

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2022

Abstract

Phosphorylation is a critical post-translational modification (PTM) involved in many cellular signaling processes. Tandem liquid-chromatography mass spectrometry (LC-MS) based experiments are commonly used to analyze levels of protein expression and their PTMs. To accurately identify phosphorylation sites through LC-MS experiments, the amino acid sequence of a given peptide identification must be correctly assigned, and the phosphate localized to the appropriate residue. A variety of localization algorithms are utilized to control the false localization rate (FLR) in LC-MS experiments. Some localization algorithms rely on the identification of site determining ions which discriminate between potential phosphorylated residues. These algorithms often interpret fragment ions which lack a phosphate as positive evidence to suggest the phosphate must reside on the complementary fragment. However concurrent HPO_3 neutral loss and peptide backbone fragmentation is a possible alternative explanation. Using a phosphopeptide library containing over 2,000 phosphopeptides, we show that concurrent HPO_3 neutral loss and peptide backbone fragmentation correlates with incorrect localization assignments. These confounding fragments can be more problematic in some higher energy collision dissociation (HCD) based datasets as opposed to collision-induced dissociation (CID) based datasets. When restricting a test localization algorithm to only consider phosphate harboring site determining ions, we see a consistent decrease in FLR's across a variety of fragmentation types. Our results implicate HPO_3 neutral loss as a problematic neutral loss for common phosphorylation localization algorithms.

Dedication

This work is dedicated to my parents, Neal and Stephanie Levy, and wife Monica Adler for their continued love and support.

Acknowledgments

I would like to thank many members of the Cell Signaling Technology community for their support over many years. Particularly, Mike Aguiar and Anthony Possemato for their essential roles in designing, executing, and interpreting these experiments. Jeff Knot and others in the peptide lab for generating reference peptide library materials. Sean Beausoleil, my thesis director, for his patience, guidance, and support of this and other projects.

Table of Contents

I. Introduction

Mass Spectrometry Based Proteomics

Proteins are the most common drug targets, with G-Protein Coupled Receptors (GPCRs) alone representing 34% of all FDA approved drug targets (Hauser, Chavali et al. 2018). Tandem Liquid-Chromatography Mass-Spectrometry (LC-MS) is a powerful tool for quantifying the expression of thousands of proteins across many samples (Wolters, Washburn et al. 2001, Schweppe, Eng et al. 2020). Direct protein measurement can more accurately predict protein abundance levels as compared to using transcriptome data to infer protein levels. Factors influencing the correlation of mRNA to protein levels include translation rates, and both protein and transcript stability (Liu, Beyer et al. 2016). Furthermore, the direct analysis of proteins can reveal information about post-translational modification (PTM) levels. PTMs such as phosphorylation, ubiquitination, acetylation, and methylation, among others, are involved in regulating protein function (Deribe, Pawson et al. 2010). PTMs are often key mediating events in various cellular signaling processes and are differentially regulated in the context of many diseases. For example, in diseases such as cancer, hyper protein phosphorylation due to mutational events, or through aberrant kinase activation, can lead to constitutive cell activation and growth (Sefton, Hunter et al. 1980, Rikova, Guo et al. 2007, Hornbeck, Kornhauser et al. 2019). As a result of its impact on the biology the cell, protein phosphorylation represents one of the most studied PTMs by LC-MS based experiments. Significant technical progress has been made in both the sample preparation, and instrumentation required for high-throughput LC-MS characterization of the phosphoproteome as

demonstrated by recent publications which have characterized as many as 50,000 unique phosphorylated peptides from mammalian cell lines (Sharma, D'Souza et al. 2014, Giansanti, Aye et al. 2015).

Sample Preparation Techniques for LC-MS Phosphoproteomics

Much like standard bottom-up proteomics, sample preparation techniques involve the digestion of proteins into smaller peptide fragments for analysis. However, due to their typically lower abundance than their unphosphorylated counterparts, phosphopeptides must be enriched to reduce sample complexity on the instrument and enable deeper profiling of the phosphoproteome. Techniques to enrich for phosphopeptides include Strong Cation Exchange, immobilized metal-ion affinity chromatography (IMAC) with metal ions commonly including iron or titanium, immunoprecipitation with motif specific phosphorylation antibodies (Rikova, Guo et al. 2007, Villen and Gygi 2008, Thingholm and Larsen 2016). These techniques have been shown to have a degree of orthogonality in their selectivity for different phosphopeptides (Possemato, Paulo et al. 2017). An alternative approach to increase the depth of phosphoproteome profiling is to utilize multiple enzymes to increase sequence coverage and recover somewhat orthogonal phosphorylation sites (Giansanti, Aye et al. 2015).

In addition to sample enrichment prior to analysis by LC-MS, improvements have been made to workflows to increase sample throughput. Many researchers utilize chemical labeling techniques to enable multiplexing of samples on the instrument and relative quantitation. Alternatively, label free quantitation (LFQ) has also been widely adopted and aided by improvements in retention time alignment algorithms to match

peaks between runs (Cox, Hein et al. 2014, Lim, Paulo et al. 2019). Amongst the sample multiplexing approaches, common labeling techniques include stable isotope labeling with amino acids in cell culture (SILAC), reductive demethylation, and the usage of isobaric labeling techniques such as isobaric tag for relative and absolute quantitation (iTRAQ) and tandem mass tags (TMT) (Thompson, Schafer et al. 2003, Wiese, Reidegeld et al. 2007, Zanivan, Krueger et al. 2012, Tolonen and Haas 2014). Recent improvements to TMT tags (Thermo Fisher) now enable the simultaneous profiling of up to 18 samples at a time (Li, Cai et al. 2021).

Phosphorylation Localization Algorithms

Many search algorithms have emerged for assigning experimental spectra to database sequences (Eng, McCormack et al. 1994, Kim, Gupta et al. 2009, Eng, Jahan et al. 2013). To help control the false discovery rate (FDR) of assignments, a target-decoy databased approach is commonly used (Elias and Gygi 2007, Elias and Gygi 2010, Savitski, Wilhelm et al. 2015). LC-MS based analysis of PTMs presents the additional challenge of having to localize modifications to the correct residue. Understanding which residue on a protein is phosphorylated can be critical to understanding the effects of the modification. For example, two different phosphorylation on the Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) can lead to either activation of inhibition of enzymatic activity (Bhattacharyya, Lee et al. 2020). Often, a peptide spectrum match (PSM) will contain multiple potentially modified residues within the peptide sequence. The goal of a dedicated localization algorithm is to distinguish which of the residues was modified, and to provide a metric for reliability of the localization assignment. Analogous to how search algorithms can misassign PSMs, PTMs may also be

misassigned and controlling the false-localization rate (FLR) of PTMs is an important issue.

In order to control FDR for PSM identifications, utilizing a target-decoy approach has become ubiquitous. However, an analogous widely adopted standard to control the FLR within a given experiment has not been established. The idea of using 'decoy' phosphoacceptors to help control the FLR in phosphoproteomics has been proposed (Chalkley and Clauser 2012). Most phosphoproteomics experiments instead however rely on a localization score cutoff to categorize observed phosphorylation sites as either confidently localized or not confidently localized sites. This is somewhat problematic, as the validation of phosphorylation localization algorithms typically is done on specific peptide libraries, with specific sample workup conditions, and specific instrument parameters. Further complicating the situation, there are a diversity of localization algorithms used for phosphorylation localization analysis with different behaviors. Studies comparing the performance and reproducibility across researchers have found that the same PSMs are often assigned to differing residues within a given peptide due to differences in scoring metrics (Chalkley and Clauser 2012).

Some of the most popular site localization algorithms include A-score (Beausoleil, Villen et al. 2006), PTM Score (Olsen, Blagoev et al. 2006), PhosphoRS (Taus, Kocher et al. 2011), Phosphorylation Localization Score (PLS) (Albuquerque, Smolka et al. 2008), SLoMo (Bailey, Sweet et al. 2009), Phosphinator (Phanstiel, Brumbaugh et al. 2011), Mascot Delta Score (Savitski, Lemeer et al. 2011), SLIP score (Baker, Trinidad et al. 2011), pSite (Yang, Chi et al. 2018), data independent acquisition (DIA) compatible methods for Spectronaut (Bekker-Jensen, Bernhardt et al. 2020), and

methods based on predicted or experimental spectral matching (Yang, Horvatovich et al. 2021). Many of these methods rely on site determining ions, which are fragment ions that distinguish one phosphoisomer from another. There are commonalities in some of these algorithms, such as the utilization of the cumulative binomial probability to estimate the liklihood that a localized site is assigned by chance. However, some of these algorithms behave differently altogether. Some localization algorithms rely on differences between search engine scores to discriminate between potential phosphorylation sites (Savitski, Lemeer et al. 2011). Others utilize diagnostic ion species to aid in phosphate localization, such as the unique ability of only phosphotyrosine to form an immonium ion (Steen, Kuster et al. 2001, Johnson and White 2012). Regardless of the algorithm, a confounding problem in the localization of phosphate to a particular residue is the lability of the protein phosphate bond.

Neutral Loss Species

The chemical bond between phosphate and Serine, Threonine, and or Tyrosine residues is inherently unstable during phosphoproteomics LC-MS analysis. During analysis on a mass spectrometer, phosphopeptides readily undergo neutral loss reactions which can predominantly result in either the loss of a neutral (non-charged) HPO_3, or H_3PO_4 from a phosphorylated side chain (Potel, Lemeer et al. 2019). This reaction can proceed in the gas phase through multiple potential mechanisms shown in Figure 1. The resulting product of metaphosphoric acid (HPO_3) neutral loss reactions is a ‘native’ Serine, Threonine, or Tyrosine which resembles an unmodified side chain. The resulting product from the neutral loss of phosphoric acid is a ‘native’ Serine, Threonine, or Tyrosine which has undergone an additional loss of water.

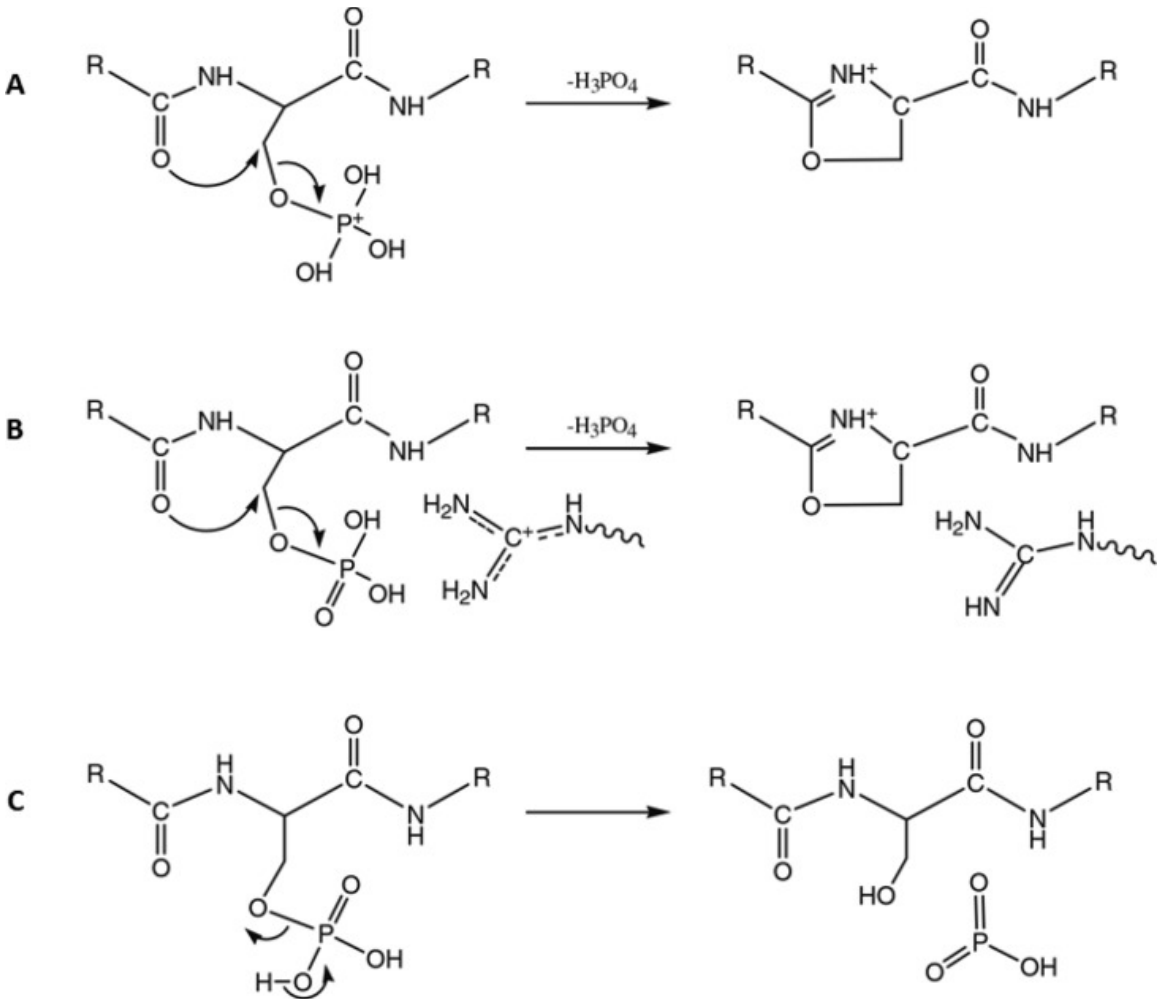

Fig 1. Mechanisms of Neutral Loss Reactions.

This figure demonstrates the predominant mechanisms by which phosphopeptides undergo neutral loss reactions. Panels A, and B depict H_3PO_4 neutral loss mechanisms, whereas panel C shows HPO_3 neutral loss. This figure was obtained from a recent review by Potel et al (Potel, Lemeer et al. 2019).

Of the possible neutral loss reactions that phosphopeptides undergo in the gas phase during LC-MS experiments, the loss of H_3PO_4 is thought to be the most common for phosphorylated Serine (pSer) and phosphorylated Threonine (pThr) peptides, and less favorable for phosphorylated Tyrosine (pTyr) (Tholey, Reed et al. 1999). The neutral loss of HPO_3 was recently shown to be a common occurrence for TMT-labeled pTyr

peptides (Everley, Huttlin et al. 2017). Precursor neutral loss peaks have been shown to be a dominant species in MS2 scans with collisional induced dissociation (CID) (Martin, Eng et al. 2005). Factors believed to affect the propensity for a given phosphopeptide to undergo different neutral loss include proton mobility among others (Boersema, Mohammed et al. 2009, Palumbo, Smith et al. 2011).

Neutral Loss Reactions Effect on Site Localization Algorithms

The neutral loss of HPO_3, or metaphosphate, poses a particular risk for many phosphorylation site localization algorithms. Site localization algorithms such as A-score, and others, interpret fragment ions which contain a potential phosphoacceptor (Ser, Thr, or Tyr), and no phosphate modification, as site determining ions that support the phosphate modification being localized to an alternative phosphoacceptor not present in that fragment. In other words, the observation of a peptide fragment without phosphate is used as positive evidence in support of the complementary fragment ion harboring the phosphate. Figure 2 shows a representation of this problem with a theoretical peptide of length 12 amino acids. This peptide is given as an example where a fragmentation event causes the formation of a b6 fragment with phosphate and a y6 fragment without phosphate. For site localization purposes, both fragments are considered by most algorithms to be evidence in support of the phosphate being on Serine at position 4. However, the observation of a y6 fragment without phosphate does not exclusively imply the phosphate must have resided on the alternative Serine on position 4. It is possible to generate a non-phosphate harboring y6 fragment through concurrent HPO_3 neutral loss and peptide backbone fragmentation events.

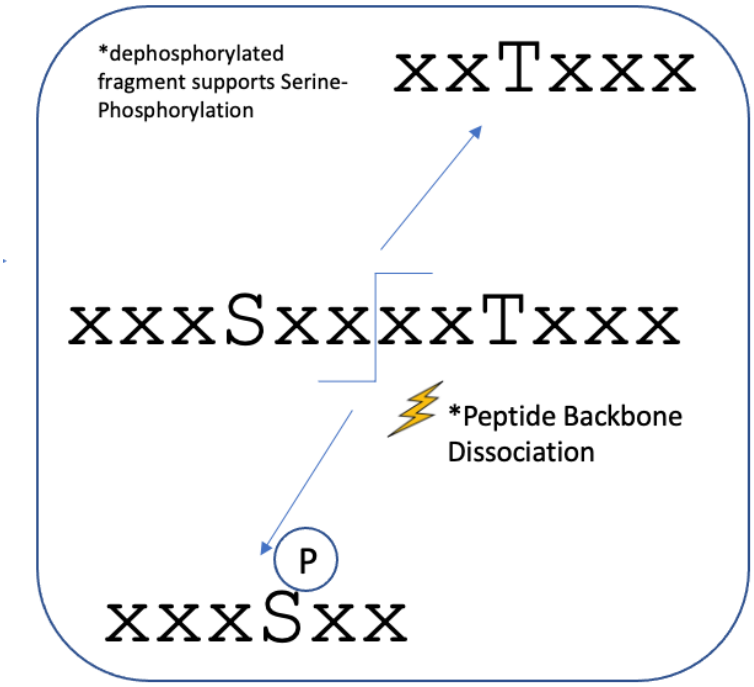

Fig 2. Schematic of a Site Determining Ion.

This figure shows a theoretical example where a phosphorylated peptide is fragmented into a b6 and y6 ion. For site localization purposes, the b6 ion shown with phosphate, and the y6 ion shown without phosphate are site determining ions in support of the phosphate being localized to the Serine residue at position 4 of this peptide. However, it is possible the y6 without phosphate could have been generated by successive HPO_3 neutral loss and peptide backbone fragmentation events.

The possibility of these concurrent fragmentation events occurring and being problematic for site localization algorithms has been previously noted (Cui and Reid 2013). However, to our knowledge, an extensive characterization of the incidence of metaphosphate neutral loss with peptide backbone fragmentation has not previously been reported. Alternative potentially problematic gas-phase reactions such as intramolecular phosphate transfer with concurrent peptide backbone fragmentation may confound the

ability to algorithmically localize phosphate to the correct residue. However, the incidence of intramolecular phosphate transfer has been described as inconsequential for site localization algorithms (Aguiar, Haas et al. 2010). An important factor which influences the propensity of different fragment ions to form, in addition to chemical labeling, is the type a dissociation deployed in an LC-MS experiment.

Effects of Dissociation Methods on Phosphopeptide Fragmentation

A diversity of sample preparation, data collection techniques, and analysis pipelines are utilized to collect and analyze phosphoproteomics data. Fragmentation techniques commonly used for MS2 identification scans include electron transfer dissociation (ETD), collisional induced dissociation (CID), collisional induced dissociation with multi-stage activation (CID/MSA), and higher energy collisional dissociation (HCD) (Potel, Lemeer et al. 2019). Experimental schemes are largely divided into label free and multiplexed approaches. Sample preparation and instrument parameters utilized depend upon the experimental design and lab preferences. Thermo Fisher recommended fragmentation modes for quantifying TMT reporter ion fragments include HCD and ETD fragmentation (Viner, Zhang et al. 2009, Ting, Rad et al. 2011). HCD based TMT quantification can be achieved with an MS2 based method, whereby concurrent peptide identification and reporter ion quantification are performed in the same MS2 scans. Alternatively, an MS3 based routine can use CID or CID/MSA for peptide identification on the MS2 scan, and synchronous precursor selection (SPS) to send selected peptide fragments from the MS2 scan for HCD fragmentation and quantitation on the orbitrap (Ting, Rad et al. 2011, McAlister, Nusinow et al. 2014).

Some groups favor HCD MS2 based TMT methods to increase the throughput and sensitivity of proteome and phosphoproteome profiling (Mertins, Tang et al. 2018). They have also described similar workflows for the detection of Ubiquitination, in addition to phosphorylation and protein levels (Udeshi, Mani et al. 2020). Alternatively, others utilize MS3 based methods for TMT-phosphoproteomics in order to help mitigate ratio compression which is often observed in TMT based experiments due to co-isolation of co-eluting peptide precursors (Li, Paulo et al. 2019). Yet another fragmentation technique which has been recently described for TMT multiplexed proteomics and phosphoproteomics is Electron-transfer/higher-energy collision dissociation (EThcD) (Yu, Shi et al. 2017).

There is not a clear consensus on the best fragmentation techniques for analyzing the phosphoproteome (Jedrychowski, Huttlin et al. 2011). When using HCD based MS2 methods for TMT phosphoproteomics experiments the collisional energy utilized must be optimized in order to efficiently fragment reporter ions, and peptide backbones. The higher collisional energies utilized to fragment reporter ions can lead to over-fragmentation of peptide backbones. One solution could be to use a 'Stepped' HCD collisional energy, whereby fragment ions are collected at a lower, intermediate, and higher collisional energies, and simultaneously analyzed on the Orbitrap (Diedrich, Pinto et al. 2013). Regardless of the methods utilized, the neutral loss of phosphate can confound the analysis. It has been reported that considering H_3PO_4 neutral loss as a site determining ion can lead to an increase the FLR (Suni, Imanishi et al. 2015).

Using a synthetic phosphopeptides library, we aimed to determine whether there were appreciable differences in the propensity for phosphopeptides to undergo concurrent

neutral loss, either H_3PO_4 or HPO_3, and peptide backbone fragmentation. We compared a variety of fragmentation techniques with and without TMT labeling. We also aimed to understand how these different neutral loss containing fragment ions affect the ability to accurately localize phosphate modifications.

II. Materials and Methods .

Generation of a Phosphopeptide Library for LC-MS Analysis

In order to evaluate peptide fragmentation patterns, and the accuracy of localizing phosphate signals to the correct residue, a synthetic phosphopeptide library was analyzed. The seed sequence for the peptide library analyzed in subsequent experiments is shown in figure 3, and was based on one of the phosphopeptide 96 libraries recently described (Marx, Lemeer et al. 2013). Each peptide within this library is of length 14 amino acids, contains a fixed phosphorylated residue at position 7, a fixed alternative phosphoacceptor (threonine) at position three, redundant residues at positions 6 and 8, and either a c-terminal lysine or arginine.

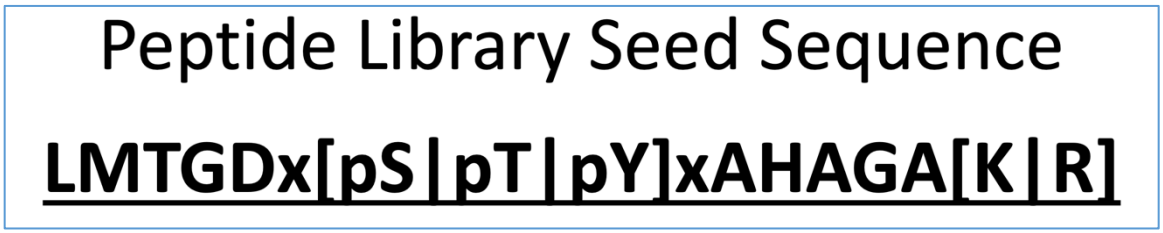

Figure 3: Phosphopeptide Library sequence.

The library sequence analyzed within this thesis is shown. A central phosphorylated Serine, Threonine or Tyrosine is flanked by positions of redundancy denoted with an 'x', which include a mixture of all 20 amino acids. C-terminal residues are composed of Lysine or Arginine.

Peptide Library Synthesis

Peptide library pools were prepared by the peptide lab at Cell Signaling Technology. Peptides were synthesized on a CEM Multipep Peptide Synthesizer utilizing standard FMOC peptide synthesis chemistry. Peptides were cleaved off the solid support and deprotected with a cocktail composed of Trifluoroacetic Acid, Water, Triisopropylsilane, Ethanedithiol and Dimethyl sulfide. Peptides were precipitated from the cleavage cocktail with Diethyl Ether two times. The resulting precipitated peptide was dried, reconstituted in water and lyophilized. Analytical data was acquired using a Waters UPLC and a Bruker Daltonics Maldi-TOF.

Peptide Cleanup and TMT-Labeling

Chemical labeling of peptides is known to affect fragmentation patterns, we therefore analyzed this peptide library both with and without TMTzero labeling (Thermo Fisher Scientific). Crude peptide synthesis material was desalted over a C18 SepPak cartridge, then dried in a vaccufuge. Peptide library samples were then resuspended for labeling in a 200mM solution of HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) at pH 8.5, in 30% acetonitrile. Labeled and unlabeled library aliquots were subjected to a final desalting prior to LC-MS analysis using stop and go extraction tips (STAGE-tips) (Rappsilber, Ishihama et al. 2003).

LC-MS Data Collection

Figure 4 shows a schematic of the experimental outline. STAGE-tip cleaned and dried peptide library samples were resuspended in 10uL of buffer A (5% acetonitrile in 0.1% formic acid). 100ng of peptide library material was loaded onto a 40cm column packed with 2.5micrometer diameter C18 Accucore resin and separated with a gradient of

buffer A and buffer B (93% acetonitrile in 0.1% FA), starting with 0% buffer B and increasing to 30% buffer B over the course of 30 minutes. The liquid chromatography was performed with a ThermoFisher Scientific EASY-nLC 1200. Tandem mass spectrometry data was collected using a ThermoFisher Scientific Orbitrap Fusion Lumos Tribrid Mass Spectrometer.

To understand the effect of different types of fragmentation techniques on peptide identification, fragmentation, and phosphate localization, several different instrument routines were run on each sample. In all cases, the MS1 survey scans were collected at a resolution of 60,000 with a scan range of 300-1500m/z and a maximum ion injection time of 25ms. Precursors with a charge state between 2-7 and an intensity above 5.0 e3 were selected for MS2 scans. Dynamic exclusion was utilized to mitigate oversampling of species, after one ms2 scan, each precursor was excluded from MS2 scans for 60s. Precursors were isolated using the quadrupole with an isolation window of 1.2 m/z. MS2 scans were all performed at a resolution of 30,000 in the Orbitrap in centroid mode. The first m/z mass was set to 110 m/z, with a normalized AGC target of 200%, with a maximum ion accumulation time of 120 ms. Multiple dissociation techniques, including HCD (using a variety of normalized collisional energies), stepped HCD (with normalized collisional energies of 15, 30, and 45), CID with normalized collisional energy of 35, and CID with a normalized collisional energy of 35 with multi-staged activation (MSA) targeting a neutral loss mass of 97.9763. Table 1 shows a summary of the various fragmentation instrument parameters.

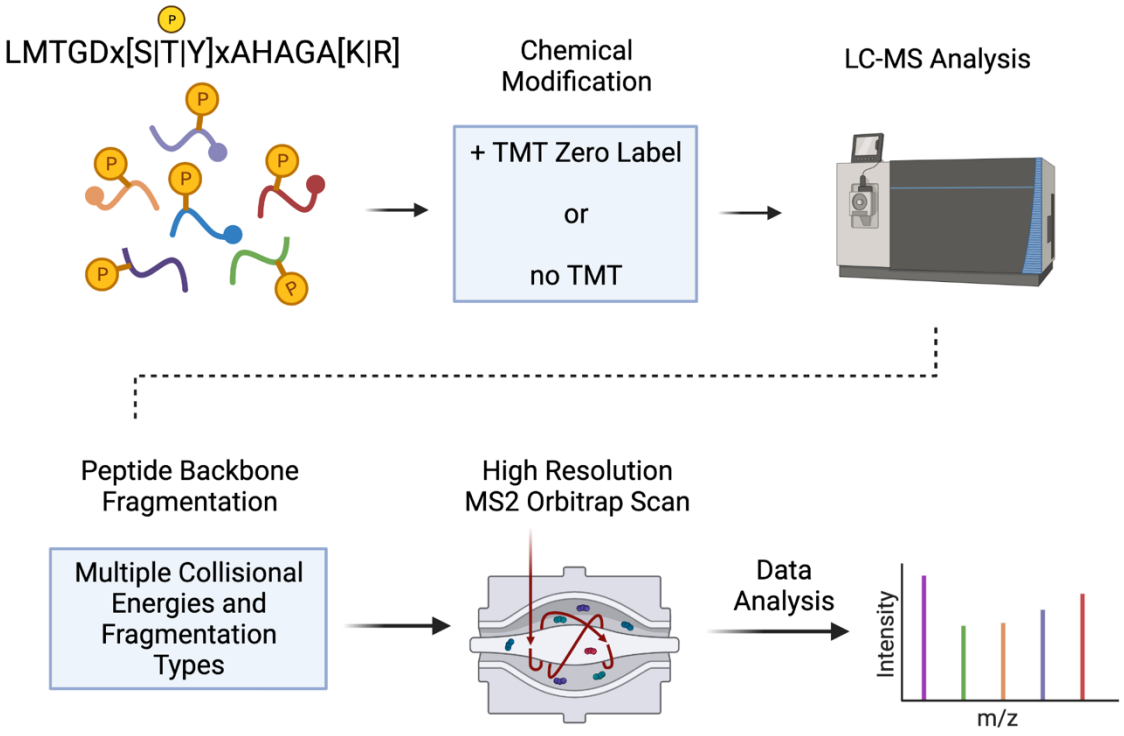

Figure 4. Experimental Scheme.

A graphical depiction of the experimental flow. Synthetic phosphopeptides libraries were analyzed with and without TMT-zero labeling using a variety a fragmentation conditions, and high resolution MS2 scans in the orbitrap detector of a ThermoFisher Orbitrap Fusion Lumos. Figure created with BioRender.com

Table 1. Fragmentation Methods Summary.

Fragmentation Type	Normalized Collisional Energy	Detector	Resolution
CID	35	Orbitrap	30,000
CID/MSA	35	Orbitrap	30,000
HCD	35	Orbitrap	30,000
HCD	30	Orbitrap	30,000
HCD	35	Orbitrap	30,000
HCD	40	Orbitrap	30,000
HCD	45	Orbitrap	30,000
HCD-stepped	15, 30, 45	Orbitrap	30,000

A summary table of the fragmentation types, collisional energies, detector, and resolution utilized for MS2 scans.

LC-MS Data Analysis

All runs were searched using a pipeline based around Comet, version 2019.01 rev. 5 (Huttlin, Jedrychowski et al. 2010, Eng, Jahan et al. 2013). Data were searched with a custom FASTA reference database containing the 2,400 peptide library sequences, common contaminants, and human Swissprot entries downloaded from Uniprot during February 2020 for further distraction. For all Comet searches, the maximum precursor ion matching tolerance was set to 20 ppm. While no enzymes were physically added to digest the libraries, searches were run with fully tryptic enzyme specificity, and allowing up to two missed cleavages. Variable phosphorylation (+79.96633) was considered on Serine, Threonine, and Tyrosine residues, allowing a maximum of up to 5 modification

per peptide. Variable oxidation (+15.99491) was considered on Methionine residues, allowing up to a maximum of 3 modifications per peptide. For runs with TMT-labeled samples, fixed TMT-zero (+224.15248) addition to Lysine side chains and peptide N-terminus was included. Fragment ion tolerance was set to 0.02 m/z for Comet searches. Neutral loss peaks from NH_3 or H_2O loss were also considered by Comet.

To selectively filter identifications for further analysis, a target-decoy approach was applied on the peptide level (Elias and Gygi 2007). Peptide spectral matches (PSMs) were filtered using a linear discriminant analysis (LDA) approach (Huttlin, Jedrychowski et al. 2010), considering the Comet log expect score, unique diff seq delta score, peptide length, number of missed cleavages, adjusted precursor ppm, fraction of ions matched, and charge state. Peptide identifications were filtered to a less than 1% false discovery rate (FDR).

As a benchmark for localization scoring, PSMs with a phosphate were analyzed using a modified version of the original A-Score algorithm, referred to as ModScore (Beausoleil, Villen et al. 2006). Fragment ion error tolerance for matching site determining ions was set to 0.02 m/z, the width of the bins for choosing the top n peaks was set to a peak window of 100 m/z, and the top 5 most abundant peaks within each 100 m/z bin were considered for site localization. To calculate a FLR from the ModScore outputs, only PSMs corresponding to full length library sequences, with a single phosphate modification were considered.

In addition to searching with a Comet and A-score based pipeline, RAW files were also analyzed through the MaxQuant pipeline (v2.0.2.0). In MaxQuant, the phosphopeptides runs were searched with default parameters. The default

phosphorylation modification which was input by Neuhouser in 2014 was added a variable STY modification. The default oxidized Methionine modification was also considered. For the TMT-zero labeled runs, custom TMT-zero modifications (+224.15248) were added to MaxQuant for n-terminal modification on any residue, and lysine modification at any position. Fixed peptide n-terminal and lysine side chain TMT-label modifications were added to the search parameters for TMT-labeled runs. Datasets were filtered to a target FDR of 1%. To analyze the false localization rates for MaxQuant at different localization thresholds, the 'Phospho (STY)Sites.txt' file from each run was analyzed. To subset singly phosphorylated identifications, only rows where "Phospho (STY) Probabilities" summed to 1 were considered.

Python Script to Search MS2 Scans

To search for and quantify the abundance and frequency of different fragment ions, a custom Python script was written which takes as inputs pepxml and mzxml files. pepxml and mzxml files were merged based on scan ID numbers. To read these file types, the Python package Pyetomics was used (Goloborodko, Levitsky et al. 2013, Levitsky, Klein et al. 2019). All python scripts were written in Python 3. To aid in calculations, NumPy, Scipy, and scikit learn were all used (Pedregosa, Varoquaux et al. 2011, Harris, Millman et al. 2020, Virtanen, Gommers et al. 2020). For visualization, Matplotlib was used in Python, and ggplot2 in RStudio (Hunter 2007, Wickham 2016). Plots in R were produced with R version R version 4.1.1,

III. Results

Instrument Parameters and Peptide Labeling Affects False Localization Rates

Using the phosphopeptide libraries, instrument parameters, and algorithms described above, we aimed to understand whether the diversity of conditions tested produced differences in FLR at commonly utilized threshold scores for localization. We analyzed the same phosphopeptides library with and without TMT labeling, across different fragmentation parameters scanned at the same resolution on the Orbitrap. First, we used the previously described Comet-based pipeline, and an algorithm called ModScore which is based on A-Score. When running ModScore, we used a fragment ion tolerance of 0.02 m/z, peak windows were set to 100 m/z, and the peak depth was fixed at n = 5. A commonly used threshold for A-Score based methods is 13, which, based on the cumulative binomial probability, corresponds to a p-value of < 0.05 that the observed site localizing ions were observed by chance (Beausoleil, Villen et al. 2006, Huttlin, Jedrychowski et al. 2010). However, as shown in figure 5, sites with ModScore values greater 13 resulted in a large range in empirically calculated FLRs. For this library, at ModScore cutoffs greater than 13, HCD fragmentation techniques produced the highest observed FLR, particularly the runs without TMT labeling and at higher collisional energies.

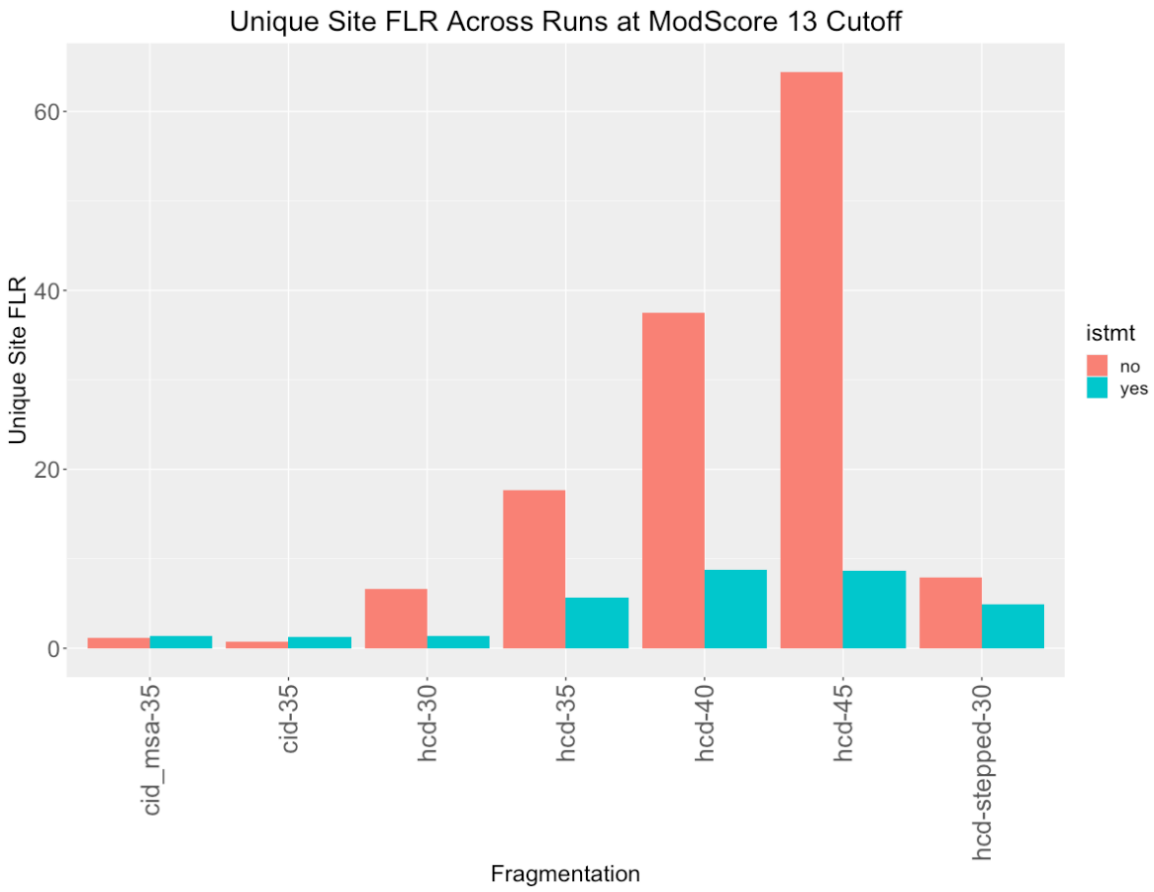

Figure 5. False Localization Rate Across Runs at ModScore 13 Cutoff.

The FLR at the unique site level is shown across different fragmentation modes and colored by whether the phosphopeptide library was TMT labeled. All peptides which produced a ModScore value greater than or equal to 13 were considered in this barplot. Unique Site FLR was calculated as the percentage of incorrectly localized phosphorylation sites relative to the total number of localized phosphorylation sites.

Next, we applied a similar analysis using the MaxQuant pipeline. For the MaxQuant analysis, a localization probability cutoff of 0.75 was utilized, which defines class I sites, or those with increased probability of being correctly localized (Olsen, Blagoev et al. 2006, Tyanova, Temu et al. 2016). This 0.75 cutoff has been reported to produce a FLR of around 3% on phosphopeptide libraries (Bekker-Jensen, Bernhardt et

al. 2020). In our data, however, there was a large degree of variability in the empirically calculated FLR at a localization probability cutoff of 0.75 across runs. As shown in figure 6, MaxQuant FLR values were worst among the higher collisional energy HCD runs, and highest on those runs without TMT labeling.

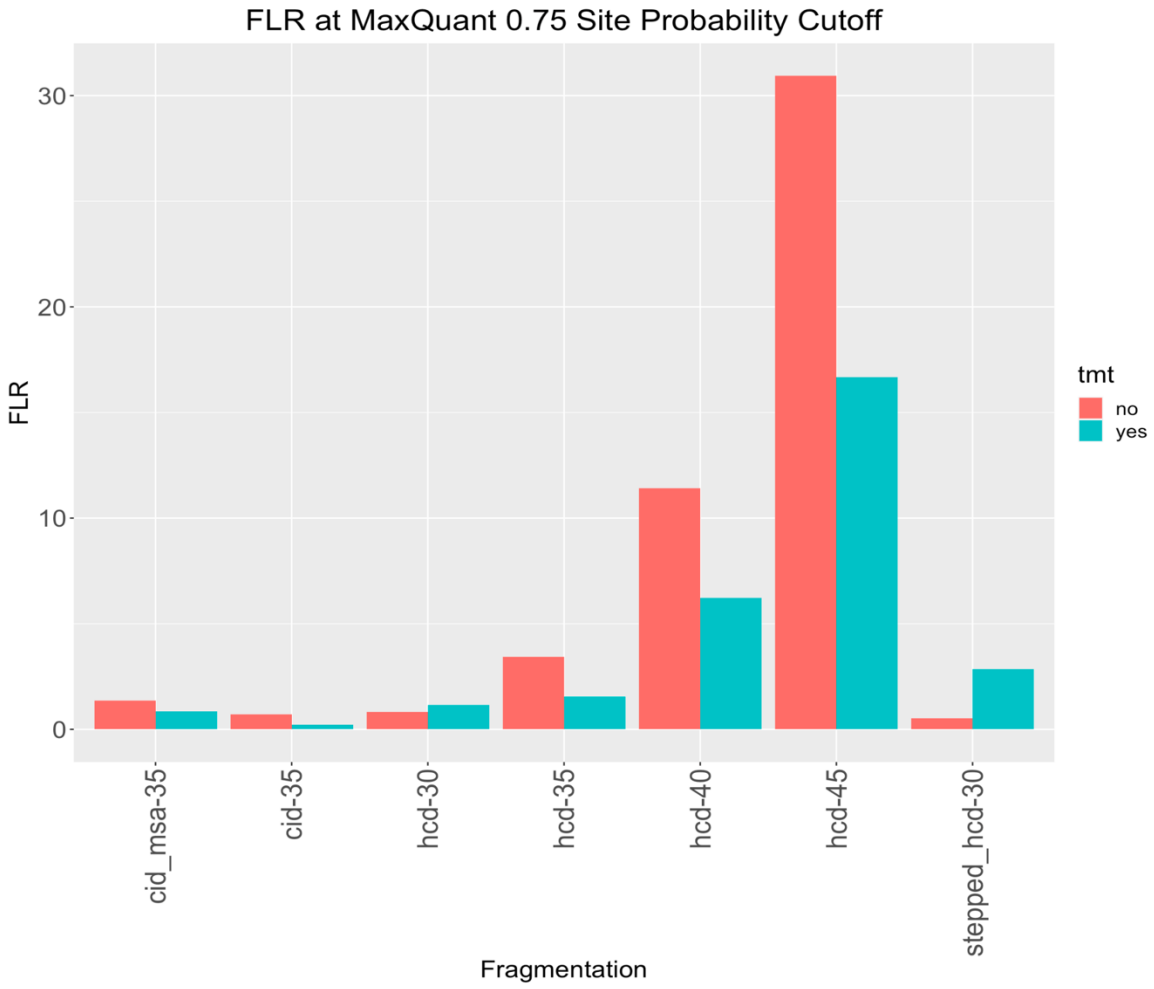

Figure 6. False Localization Rates at MaxQuant 0.75 Localization Probability Cutoff

The FLR across the same runs was calculated across runs as the fraction of incorrectly localized sites with a localization probability greater than or equal to 0.75. Phosphorylation sites derived from multiply phosphorylated identifications, reverse hits, and non-full length peptide library identifications were omitted from the analysis. FLR

was calculated as the percentage of incorrectly assigned phosphorylation sites relative to the total number of phosphorylation sites.

This variability in FLR at static cutoffs of localization scores is problematic. We hypothesized that the variability in FLR might be partially due to the confounding effects of concurrent peptide backbone fragmentation and phosphate neutral loss. To better understand if these observed differences in FLR correlate with differences in neutral loss fragmentation, we characterized the frequency with which fragment ion species were observed across these same runs. Among the fragment ions we searched for included the complete set of a, b, c, x, y, and z ions for this peptide library, neutral loss containing versions of those fragment ions, in addition to precursor neutral loss fragmentation (Steen and Mann 2004).

Observation of Neutral Loss containing Fragment Ions

To effectively annotate fragment ions from MS2 scans in a high-throughput fashion, a custom python script was used. This script utilizes mzxml and pepxml files as inputs. The mzxml file contains essential information included observed peaks, and their corresponding intensities, whereas the pepxml files contain peptide identification information. These files were merged by scan identifying numbers. First, only MS2 scans which passed the following criteria were considered for fragment ion mapping: the PSM must pass LDA filtering to a 1% FDR, the identification must map to a singly phosphorylated entry from the phosphopeptide library and be of length 14 amino acids. After these criteria were met, each peak within the mzxml file was mapped against the possible fragment ions that could be produced by the identified sequence. In addition to the regular b and y-ion series, we also considered concurrent neutral loss of HPO_3 or

H_3PO_4 and peptide backbone fragmentation. Figure 7 shows annotation of b and y ions from this phosphopeptide library for reference.

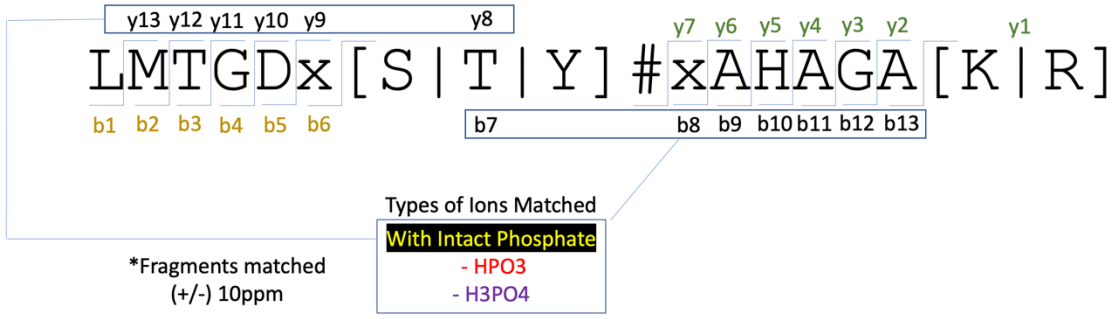

Figure 7. Annotation of Phosphopeptide Library Fragments.

The numbering of b and y fragment ions from this phosphopeptides library is displayed above. In addition to the b and y ions from this peptide, fragment ions containing the central phosphorylated residue were considered with additional HPO_3 or H_3PO_4 neutral loss were considered.

To mitigate the potential for incorrectly annotating any peaks as the result of concurrent neutral loss and peptide backbone fragmentation events, every peak was first considered against a variety of other fragment ion species. First, any fragments that matched the precursor m/z, or a precursor H_3PO_4 or HPO_3 neutral loss were identified and removed from further consideration for mapping to other fragment ion types. Next, fragment ions mapping to the terminal ends of the phosphopeptide, which do not include the central phosphoacceptor at position 7, were identified and removed from further consideration. The remaining m/z values in the peak list which did not map to any of the previous ion species were considered for further mapping of fragments ions. Next, a, b, c, x, y, or z ions which contain the central phosphorylated residue were searched for (a7-a13, b7-b13, c7-c13, x8-x13, y8-y13, and z8-z13). Each of these fragment ion species

were considered with an intact phosphate moiety, H_3PO_4 containing neutral loss, or HPO_3 neutral loss.

When searching for each of these fragment ions, zero, one, or two c13 peaks were considered in the calculations. For a given peptide precursor at a charge state n, fragment ion charge states were considered up to n-1. When searching for peptide precursor neutral loss species, only charge state n was considered, but up to three c13 peaks were considered. When calculating the theoretical masses of these different fragment ions to search for, the University of Washington's Proteomics Resource was used to collect monoisotopic masses for all amino acids. Numpy was used for most calculations in the custom python script (Harris, Millman et al. 2020). Representative annotated spectra from the TMT labeled phosphopeptides library are provided in figures 8-10 from the CID, CID-MSA, and HCD40 runs. The coloring scheme for labeling from figure 7 is utilized to aid in distinguishing different fragment ion species. As expected, the CID run has a predominant precursor neutral loss peak, as does the CID-MSA. This precursor neutral loss peak is less predominant in the HCD40 run. In all cases however, there are multiple fragment ions identified in each scan (shown in red) with concurrent HPO_3 neutral loss and peptide backbone fragmentation.

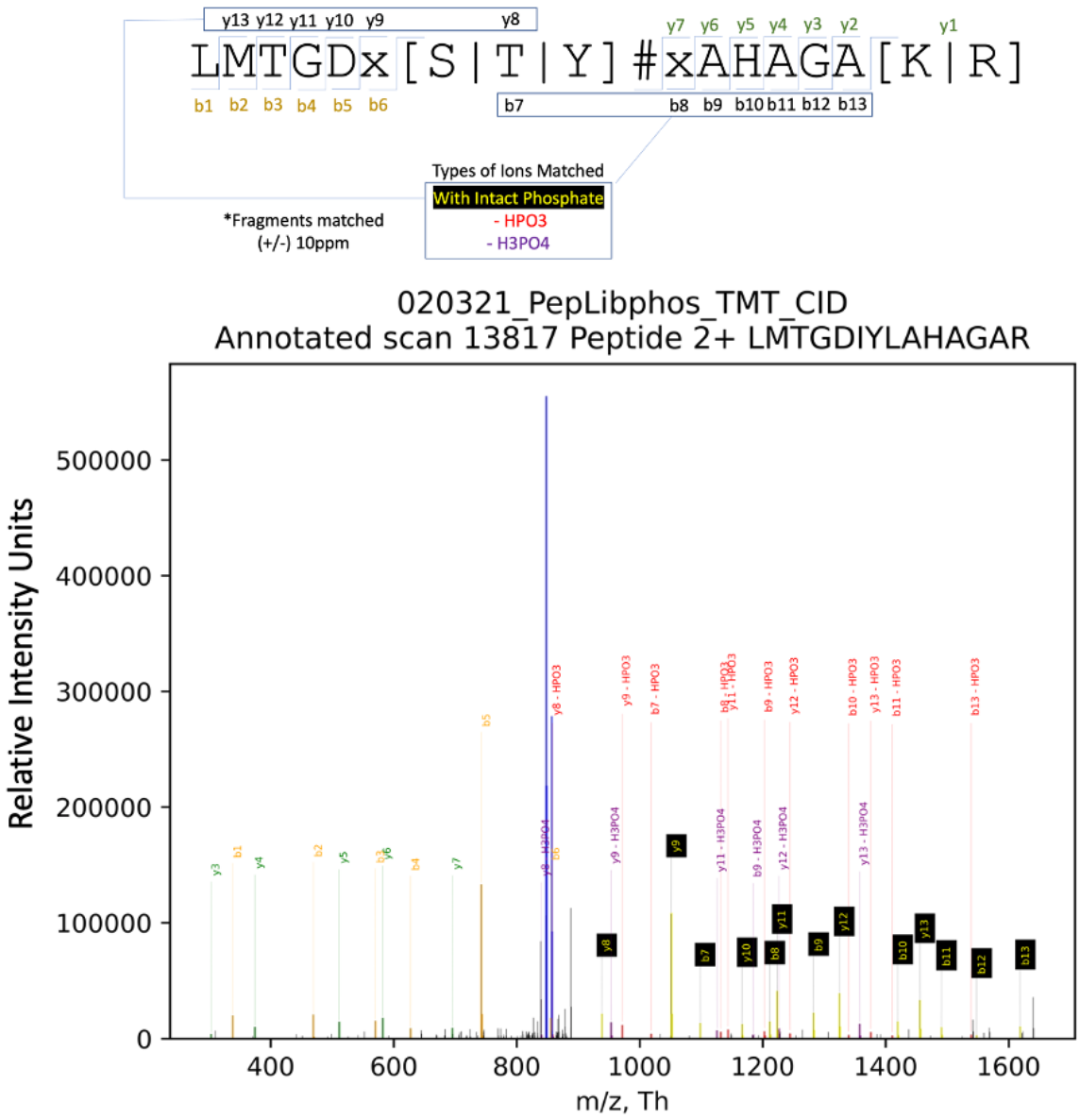

Figure 8. Annotation of Fragment Ions from a CID Scan

Shown above are annotated peaks from a 2+ precursor with amino acid sequence LMTGDIYLAHAGAR. Shown in blue are precursor neutral loss fragments, brown are b ions which do not contain the central phosphorylated residue, green are y ions which do not contain the central phosphorylated residue, yellow are b or y ions with the central phosphoacceptor and an intact phosphate, purple are b or y ions with the central

phosphoacceptor and H_3PO_4 neutral loss, and red are b or y ions with the central phosphoacceptor and HPO_3 neutral loss.

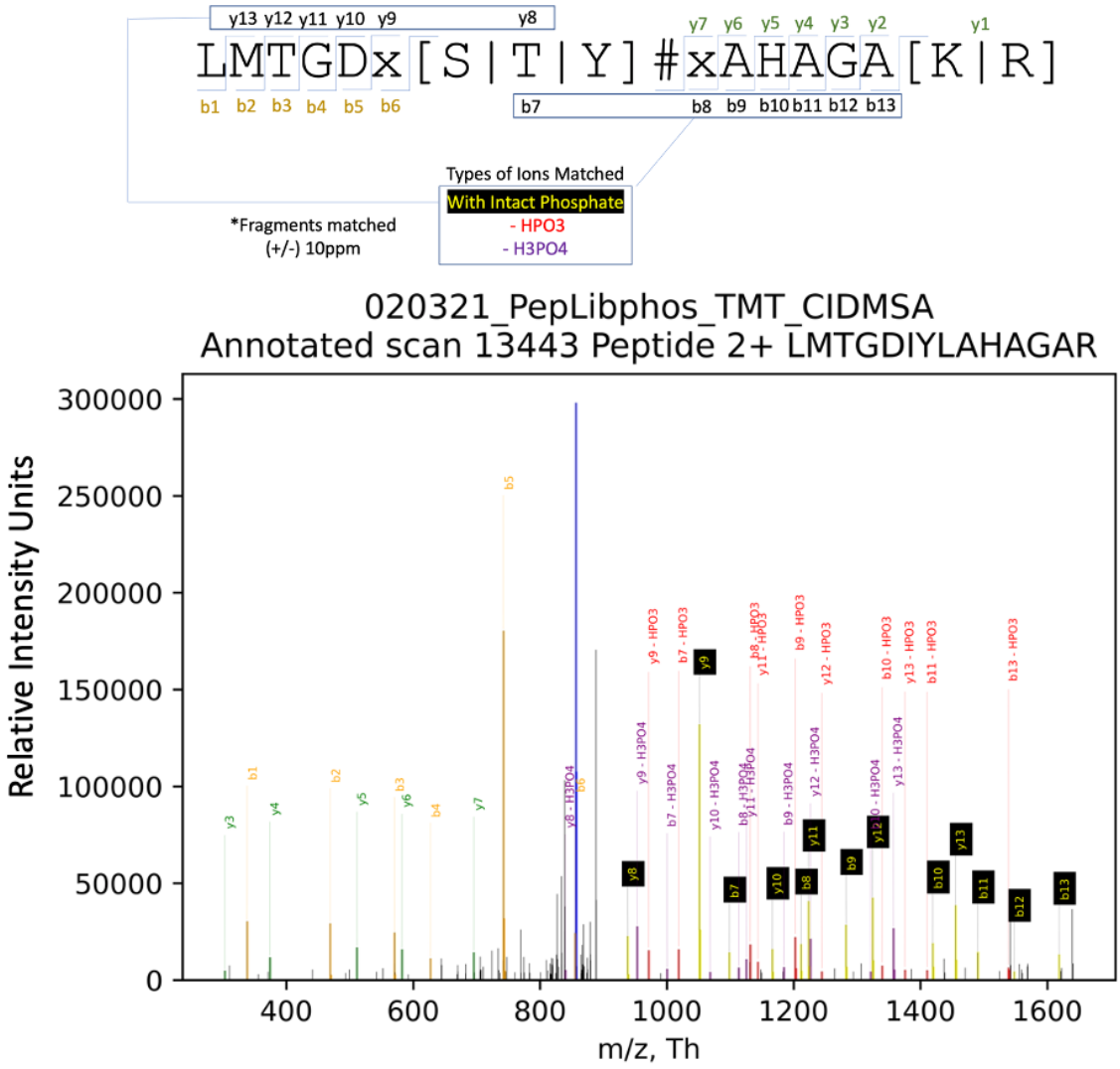

Figure 9. Annotation of Fragment Ions from a CID-MSA Scan

This figure shows an analogous MS2-scan for the same precursor presented in figure 8 for the CID-MSA fragmentation run.

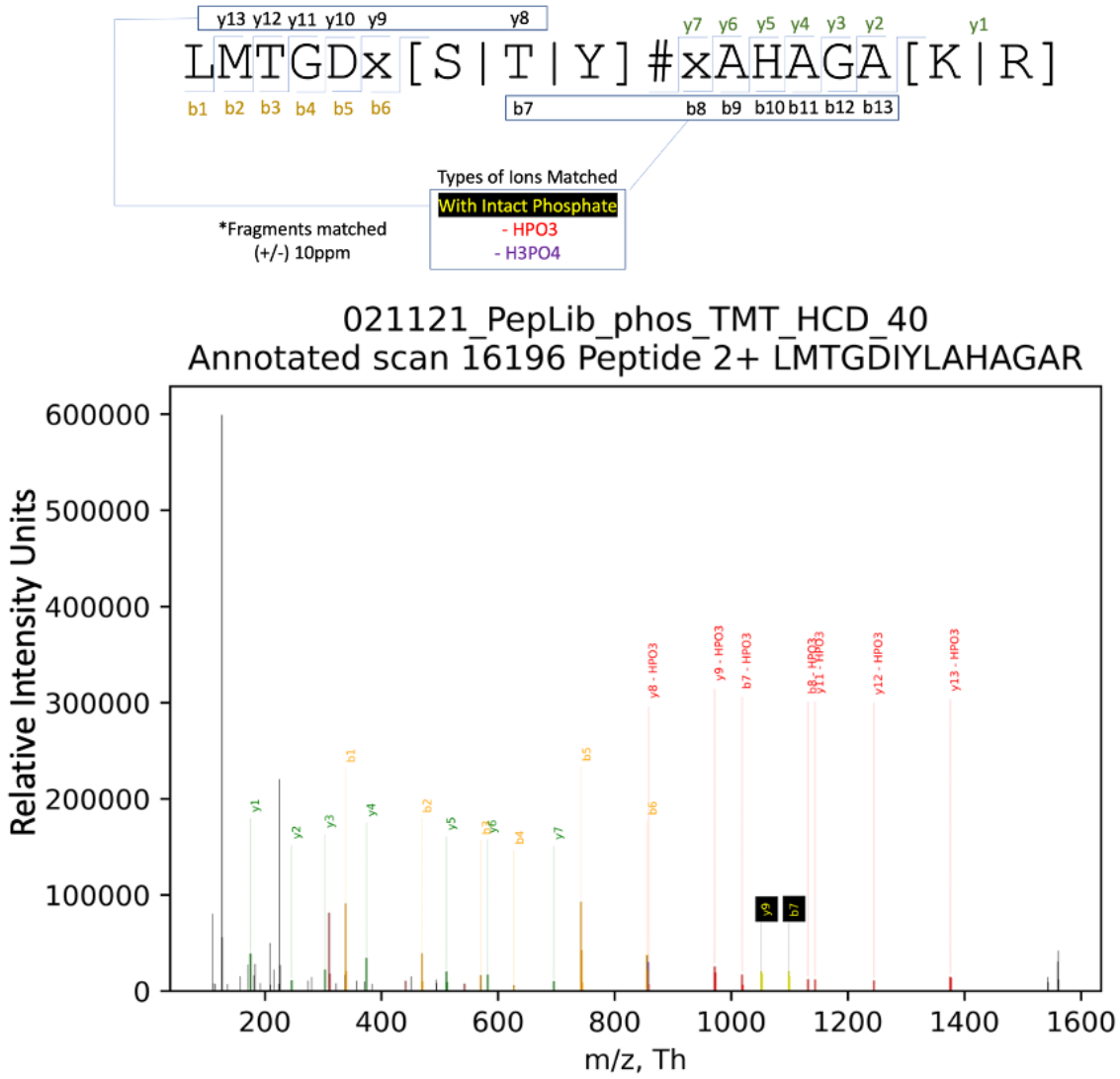

Figure 10. Annotation of Fragment Ions from a HCD Scan with Collisional Energy 40

This figure shows an analogous MS2-scan for the same precursor presented in figure 8 for the HCD fragmentation run with collisional energy of 40.

Using this algorithm designed to distinguish and annotate the different types of fragment ions, we next wanted to determine how TMT-labeling, and instrument fragmentation parameters, affects both the propensity to form different fragment ions, and the ability to accurately localize phosphate.

B and Y Ion Frequency with and Without TMT

Correlation Between HPO_3 Neutral Loss and FLR

The previously described method for annotating peaks for each MS2 scan was applied across all the runs. Among the fragment ions which contain the central phosphorylated residue in this phosphopeptides library, there are three types of ions we tracked: fragment ions with an intact phosphate, fragment ions with HPO_3 neutral loss, and fragments with H_3PO_4 neutral loss. To estimate the relative abundance of these different species, we counted the number fragment ions detected and the relative ion current for each fragment ion type. As shown in figure 11, with or without TMT labeling, there was a decrease in the number of fragment ions identified with an intact phosphate in the HCD collisional energy of 30 compared to either CID or CID-MSA fragmentation. Across these runs, there was a significant number of fragment ions identified with H_3PO_4 neutral loss for phosphoserine and phosphothreonine containing peptides. The favorability of H_3PO_4 neutral loss on Serine and Threonine is consistent with previous reports. Previous publications have shown that peptide precursor HPO_3 neutral loss peaks are increased on phosphotyrosine peptides with the addition of TMT: here we also see that the concurrent neutral loss and peptide backbone fragmentation is increased with the addition of TMT to phosphotyrosine containing peptides.

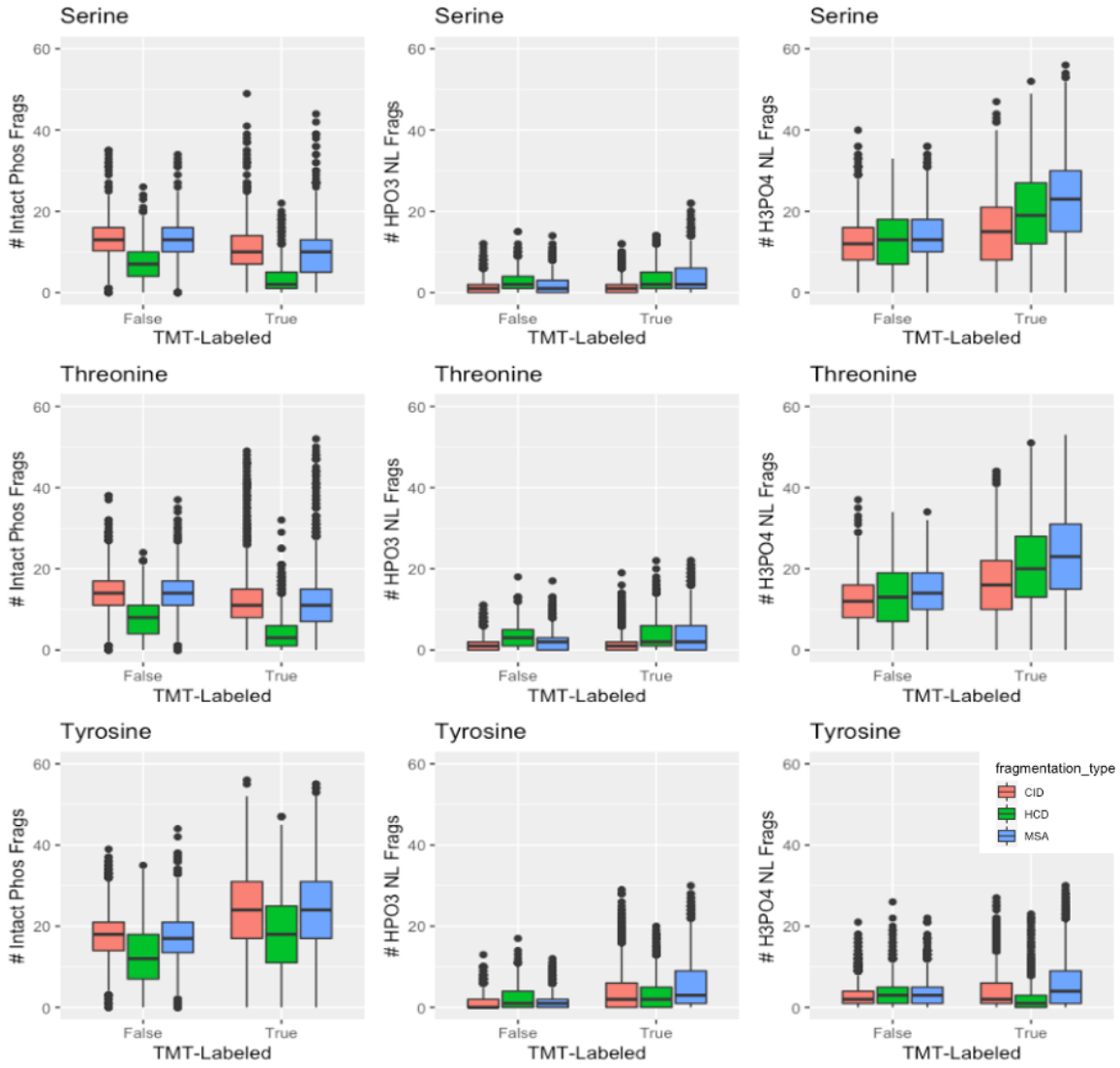

Figure 11. Number and Type of Ions Detected Across CID, MSA, and HCD Runs.

Across the figures from left to right, the number of identified fragments harboring the phosphorylated residue in this library with intact phosphate, HPO_3 neutral loss (resulting in a 'native' Serine Threonine or Tyrosine), and H_3PO_4 neutral loss are shown. From top to bottom, identified peptide sequences containing phosphoserine, phosphothreonine, and phosphotyrosine are shown. The HCD runs shown were at collisional energy of 30. Number of fragments identified includes all charge states and up to 2 different C13 peaks per fragment.

With the capability to track these different fragment ion types across scans and runs, we next aimed to correlate the incidence of fragment ion species with the ability of localization algorithms to correctly localize the phosphate. With increased HCD collisional energy on this peptide library, there was a significant decrease in b and y ion sequence coverage, due to propensity of HCD fragmentation to generate smaller fragment ions at higher collisional energies. This results in the recovery of fewer fragment ions which contain the centrally located phosphorylated residue. This is shown in figure 12, where the overall number of identified ions containing the central phosphorylated residue decreases at higher HCD collisional energies.

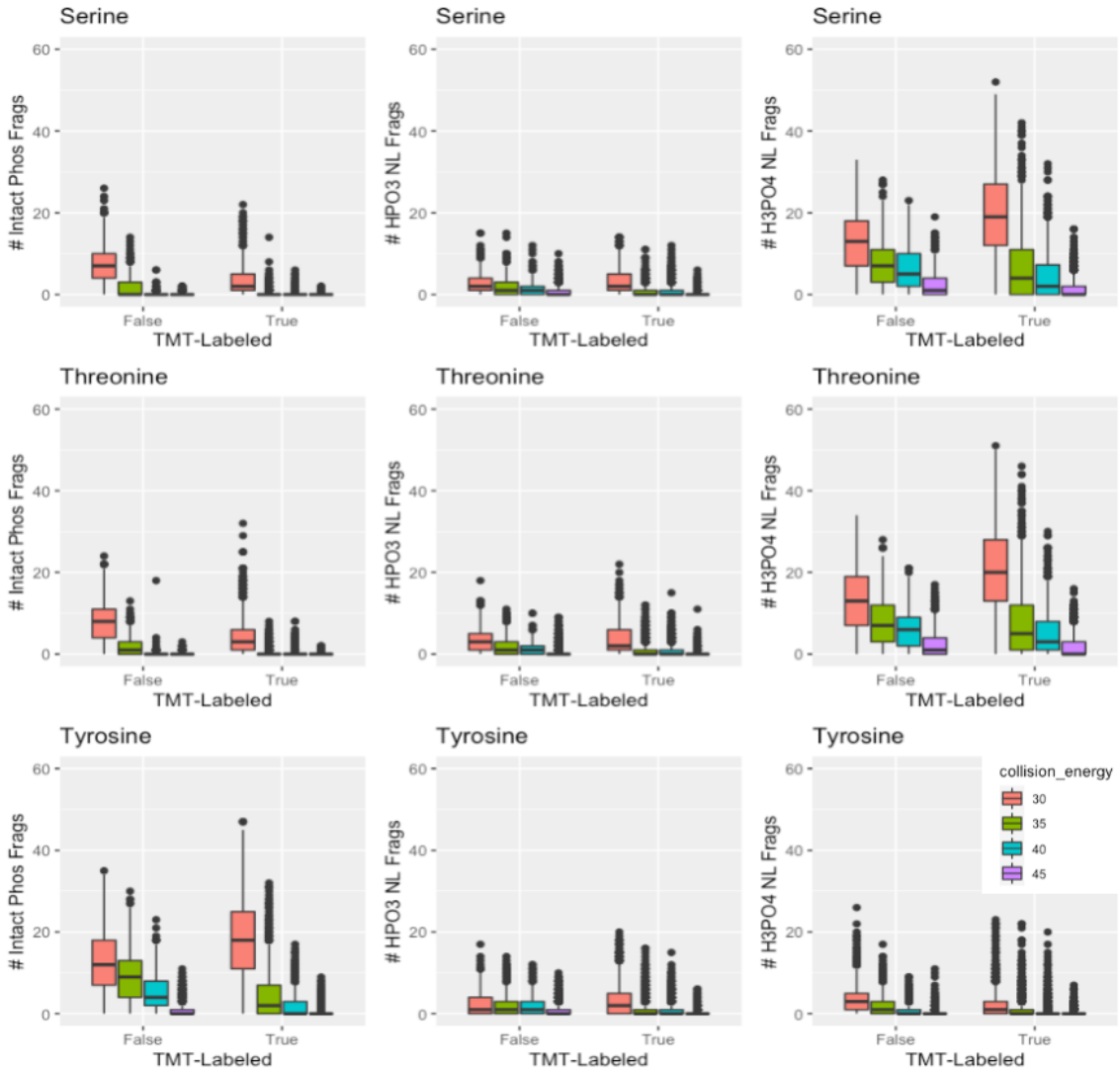

Figure 12. Number and Type of Ions Detected at Various HCD Collisional Energies

This figure shows the number of fragment ions identified across different HCD collisional energies using the same criteria as described for figure 11.

The overall number of identified ions containing the central phosphorylated residue in the higher energy HCD runs are reduced. However, when we subset on MS2 scans where phosphate containing or HPO_3 neutral loss containing fragments are found, we see the ratio an increase in the ratio of ion current derived from HPO_3 neutral loss compared to fragments with an intact phosphate. Figure 13 shows the relative increase in

ion current generated from concurrent HPO_3 neutral loss and peptide backbone fragmentation as HCD collisional energy increases. In the runs with increased signal of HPO_3 neutral loss, we also observe a corresponding increase in the FLR.

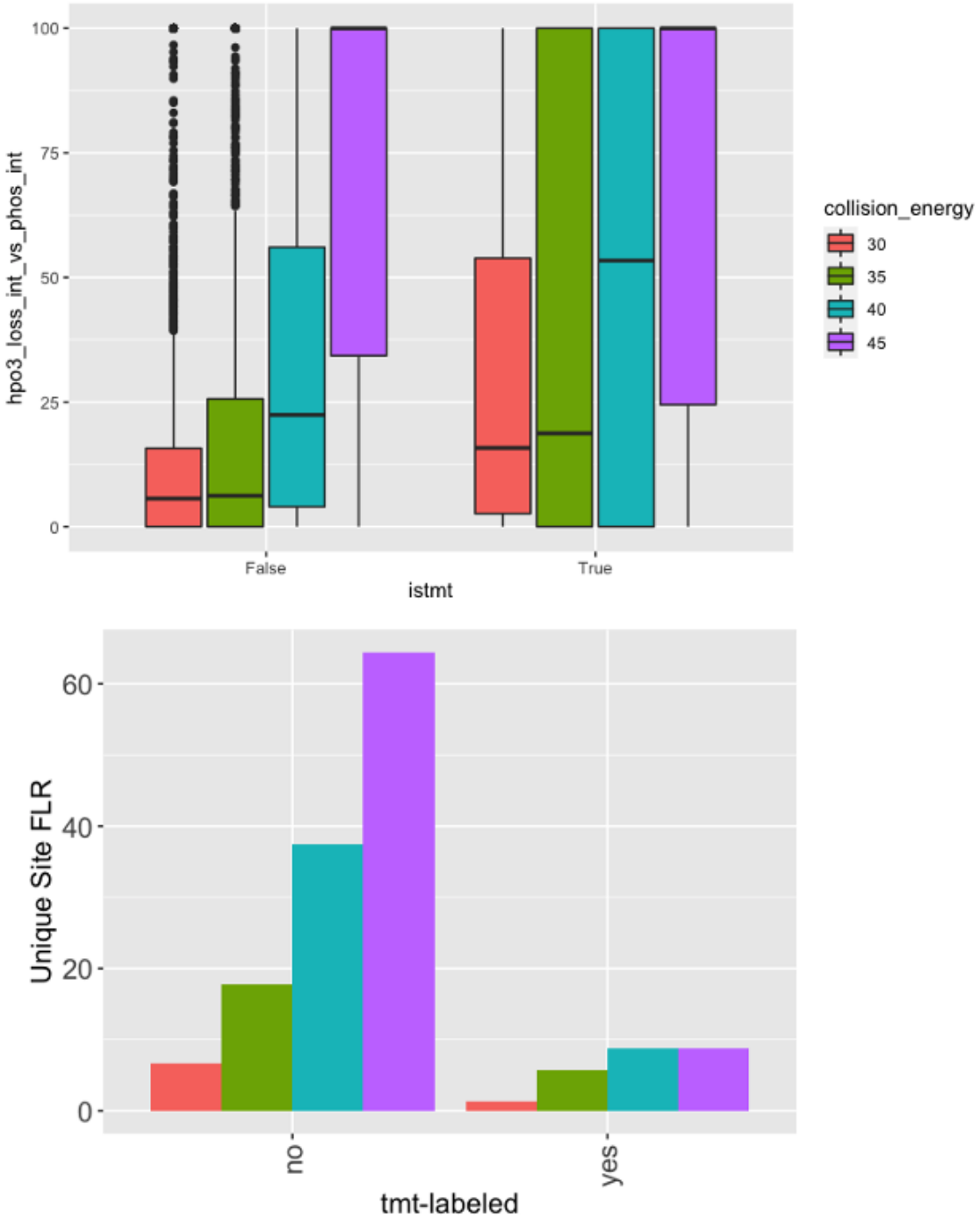

Figure 13. Relative HPO_3 NL Signal Correlates to Run FLR

On the top row, the fraction of ion current containing HPO_3 neutral loss in addition to peptide backbone fragmentation is expressed relative to the total ion current of intact phosphate or HPO_3 neutral loss containing fragments. On the bottom panel, the corresponding FLR is shown for comparison.

The correlation between the runs with a higher relative signal for HPO_3 neutral loss and increased FLR, is suggestive that these fragment ions could be causative for incorrect localization. An alternative way to implicate HPO_3 neutral loss with false localization is to compare HPO_3 neutral loss and localization scores within a run. To do this, we subset on MS2 scans where fragments were recovered that contained the centrally phosphorylated residue. When analyzing scans with ModScore values greater than 13, we see that incorrectly localized assignments are associated with a relatively higher fraction of ion current derived from HPO_3 neutral loss. As shown in figure 14, for all the different HCD fragmentation energies tested, with and without TMT-labeling, we see more HPO_3 neutral loss in incorrectly localized identifications. While the degree of HPO_3 neutral loss is correlated with incorrect localization, there are still significant differences in the FLR between the higher energy HCD runs with and without TMT.

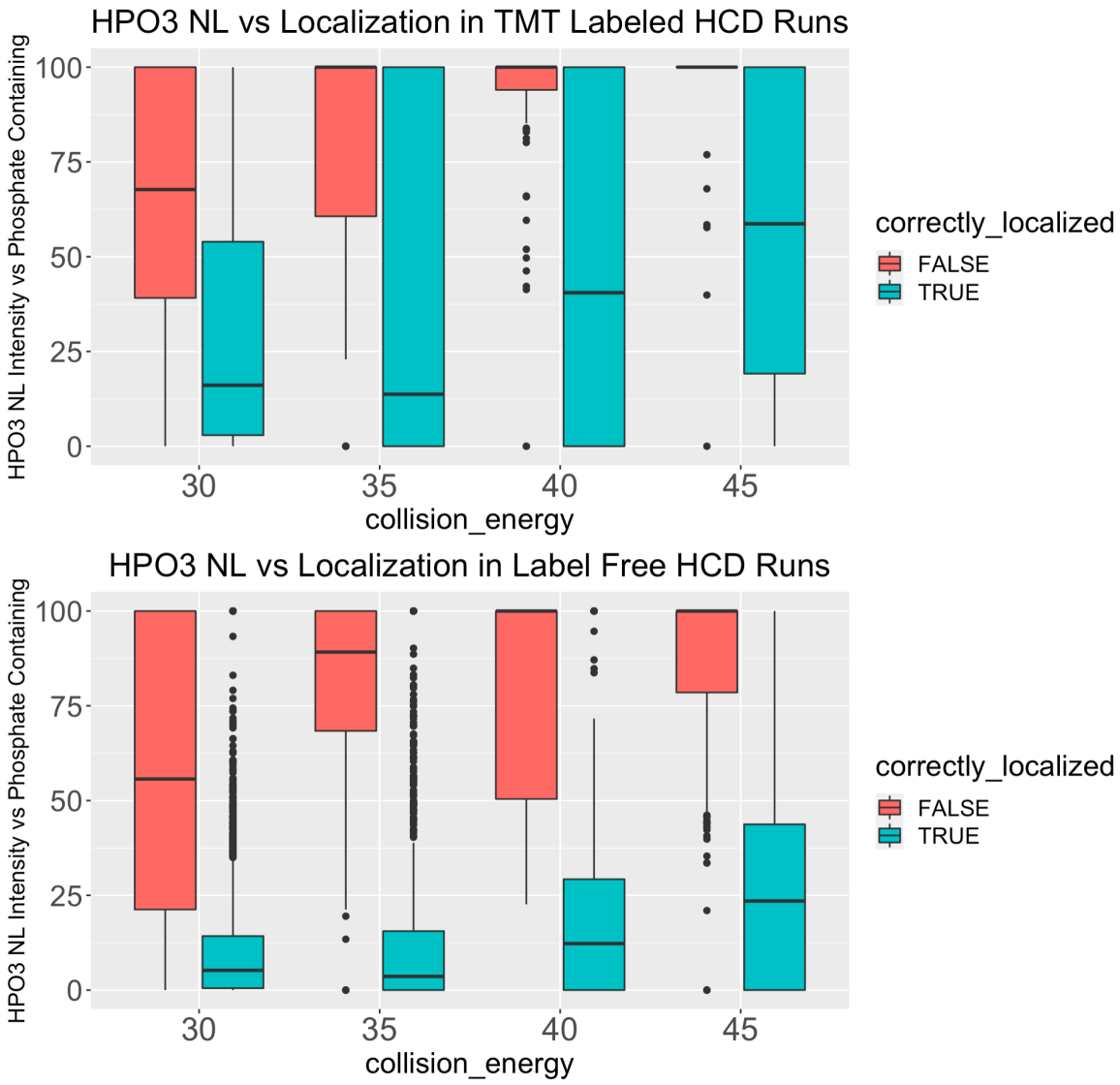

Figure 14. Correlation of HPO_3 Intensity to Localization.

This figure shows the fraction of intensity for b and y ions which contain the phosphorylated residue with concurrent HPO_3 neutral loss, expressed as a percentage of the sum of all fragment ions intensities containing the phosphorylated residue or HPO_3 neutral loss. For this library, correctly localized assignments had the phosphate assigned to position 7. Assignments with ModScore cutoffs >13 are shown. Scans without phosphate or HPO_3 neutral loss containing fragments are omitted from plots.

One difference between the TMT and label free HCD runs, which could explain the observed difference in FLR, is the ability to recover b and y ions. As shown in figure 15, we observed a significant decrease in the number of identified b-ion fragments from the unlabeled runs, as compared to the y-ion fragments. Since there is a fixed alternative phosphoacceptor at position three in this library, the inability to efficiently recover a b-ion series could be driving the higher FLR in the unlabeled libraries.

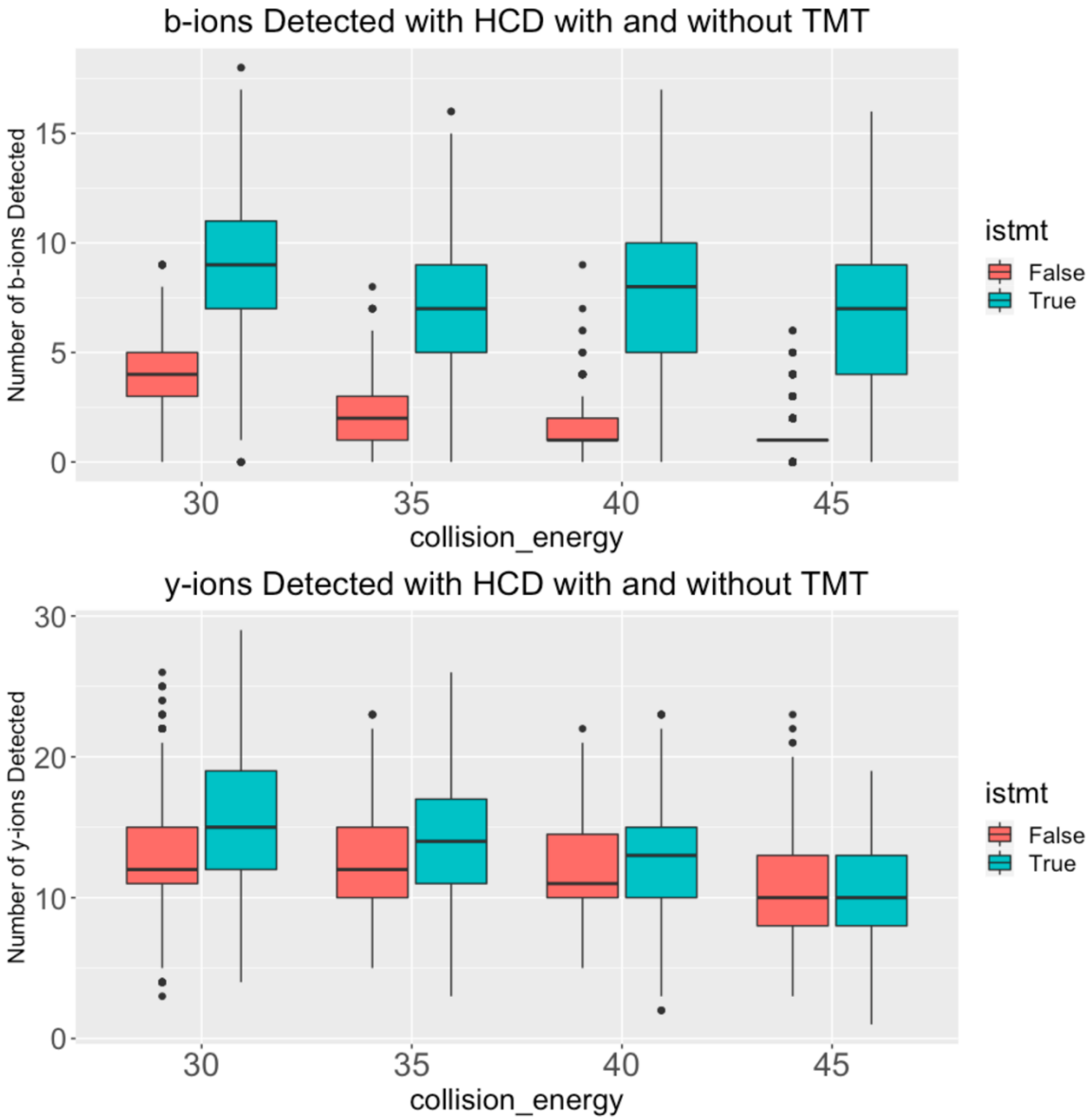

Figure 15. b- and y-ion Detection With and Without TMT

Shown above are the number of b and y fragments ions detected across the HCD runs at various collisional energies, which do not contain the phosphorylated residue at position 7. Identified ions include all charge states and up to 2 different C13 peaks per fragment.

Taken together, these previous data provide circumstantial evidence that concurrent peptide backbone fragmentation and HPO_3 neutral loss can influence localization algorithms to incorrectly assign the phosphate modification. To further support this, we implemented a localization algorithm to demonstrate that the removal of these ion types from consideration improves the ability to control the FLR.

Site Localization Algorithm Testing

Finally, a custom Python script was written to evaluate whether FLR can be better controlled by only relying on phosphate harboring site determining ions. Similar to the original A-Score algorithm, and MaxQuant site localization algorithm, the underlying statistics are built around the cumulative binomial probability. For reference, the cumulative binomial probability function is provided in figure 16, as shown in Beausoleil et al 2006 (Beausoleil, Villen et al. 2006).

$$P(x) = \sum_{k=n}^{N} \left[\begin{matrix} N \\ k \end{matrix} \right] p^k (1-p)^{N-k}$$

Figure 16. Cumulative Binomial Probability.

The cumulative binomial probability equation shown here is utilized to provide a probability-based score for phosphorylation localization. This figure was taken from the

original A-score paper (Beausoleil, Villen et al. 2006). Trials, which are denoted as 'N' are the number of site determining ions considered. The variable 'p' is a function of the width of the m/z range surveyed, and resolution of the ms2 scan. The calculated score is the cumulative probability of observing k or more site determining ions.

There are several differences between the site localization algorithm tested here and the original A-score algorithm. Rather than requiring fixed bin widths, peak depth, or ion matching tolerances, an algorithm was implemented which does not require any of these parameters to be pre-defined. Instead of evaluating multiple bins, the difference between the smallest and largest m/z observed within each MS2 scan was used as a single bin. Additionally, every observed peak in the scan was considered, rather than the top n most abundant peaks. Therefore, p is a function of the m/z range of peaks detected in each scan, the number of peaks in the scan, and the fragment matching tolerance.

Instead of specifying a fragment ion tolerance cutoff for matching site determining ions, a z-score based method was used. To enable a z-score based cutoff for fragment ion matching, all ms2 scans passing LDA filtering were first searched for fragment ion matches within a +/- 10 ppm window. Of the matched fragment ions, the difference between predicted and observed fragment ion m/z was often found to be affected by both the retention time, and the predicted m/z. Figure 17 shows an inverse relationship between the measured fragment ion m/z error and retention time for this label free MSA run. Figure 18 also shows the inverse relationship in this same run between the error in measurement of fragment ion m/z, and the predicted fragment ion m/z. To account for this relationship, a multiple regression was performed on the matched fragment ions while considering retention time and the predicted m/z. From this

linear model, a predicted m/z value was generated which is a function of the retention time, predicted fragment m/z, and an additional constant. After regression, the resulting difference between measured m/z and adjusted predicted m/z are centered over zero, and normally distributed. Figure 19 shows a histogram of adjusted ppm errors. A z-score cutoff of 2 from this model was used for ion matching in our site localization algorithm: in the data shown here, a z-score cutoff of 2 equates to (+/-) 4.1 ppm.

The algorithm tested here also differs in how potential phosphorylation sites are scored against each other. Instead of scoring only the two most likely phosphorylation sites against each other, the algorithm tested here evaluated all potential phosphorylation sites in a pairwise method. Like the A-score paper, the difference in negative log10 of the p-values from the binomial were used to compare each site against each other. However, this logic was extended to compare each potential phosphorylated residue against all other potential phosphorylated residues. To combine the pairwise scores for each site, a geometric mean was computed. Additionally, when defining a list of site determining ions as trials for our site localization algorithm, only the c12 version of each potential site determining ion was considered. Several peaks were removed from consideration prior to searching for site determining ions. Since this algorithm was developed and tested on high resolution ms2 data, fragment ions which appeared to be a c13 peak from an isotopic cluster were removed from consideration. This algorithm was used to further understand the relationship between fragment ions with HPO_3 neutral loss and incorrect localization.

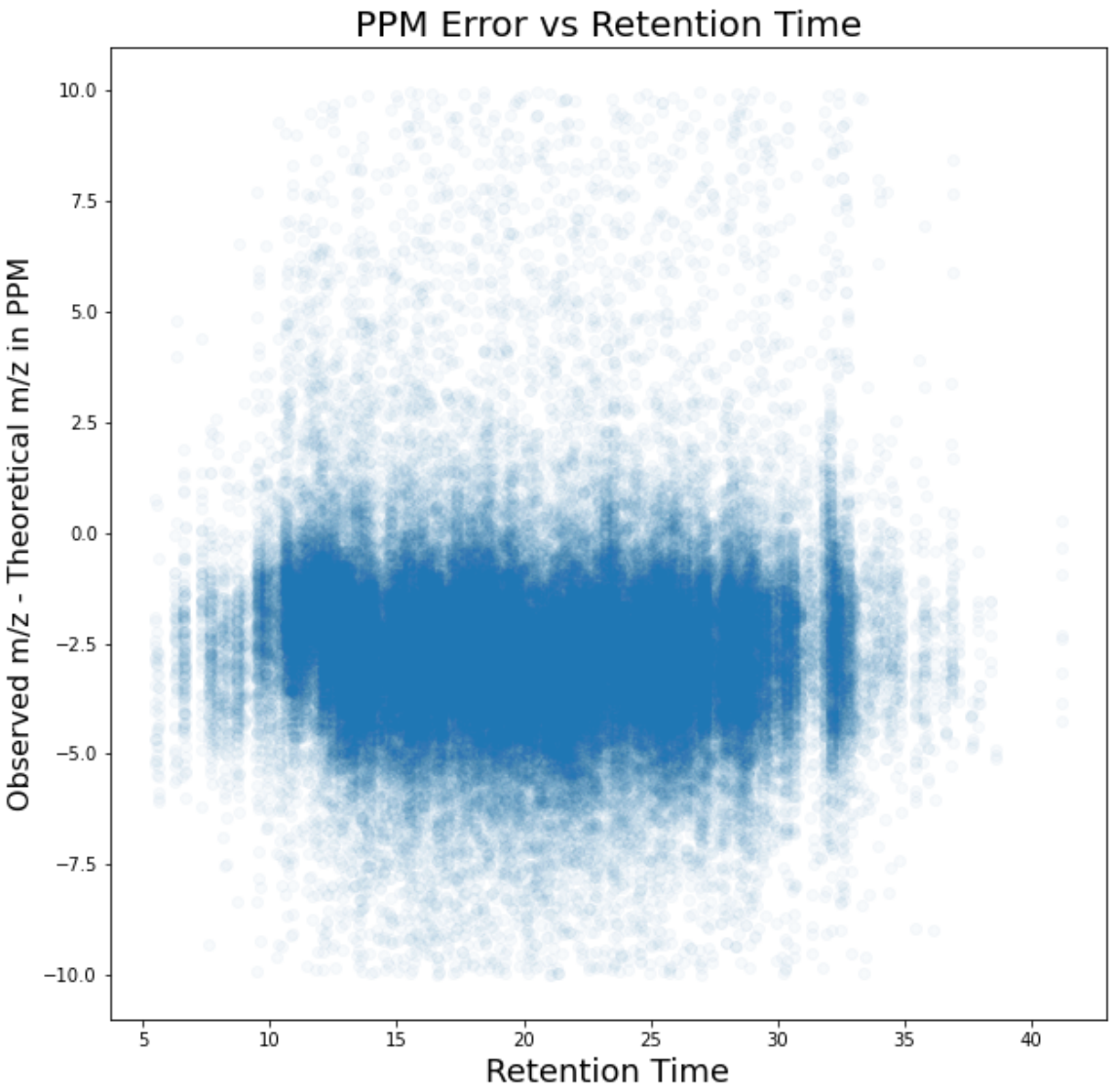

Figure 17. Matching Fragment Ion Ppm Errors by Retention Time.

Theoretical fragments from the identified peptide were searched within each ms2 scan. Fragment ion matches +/- 10ppm are plotted. In this label free CID-MSA run, there is a slight downward trend in the matched fragment ion ppm error over chromatographic time.

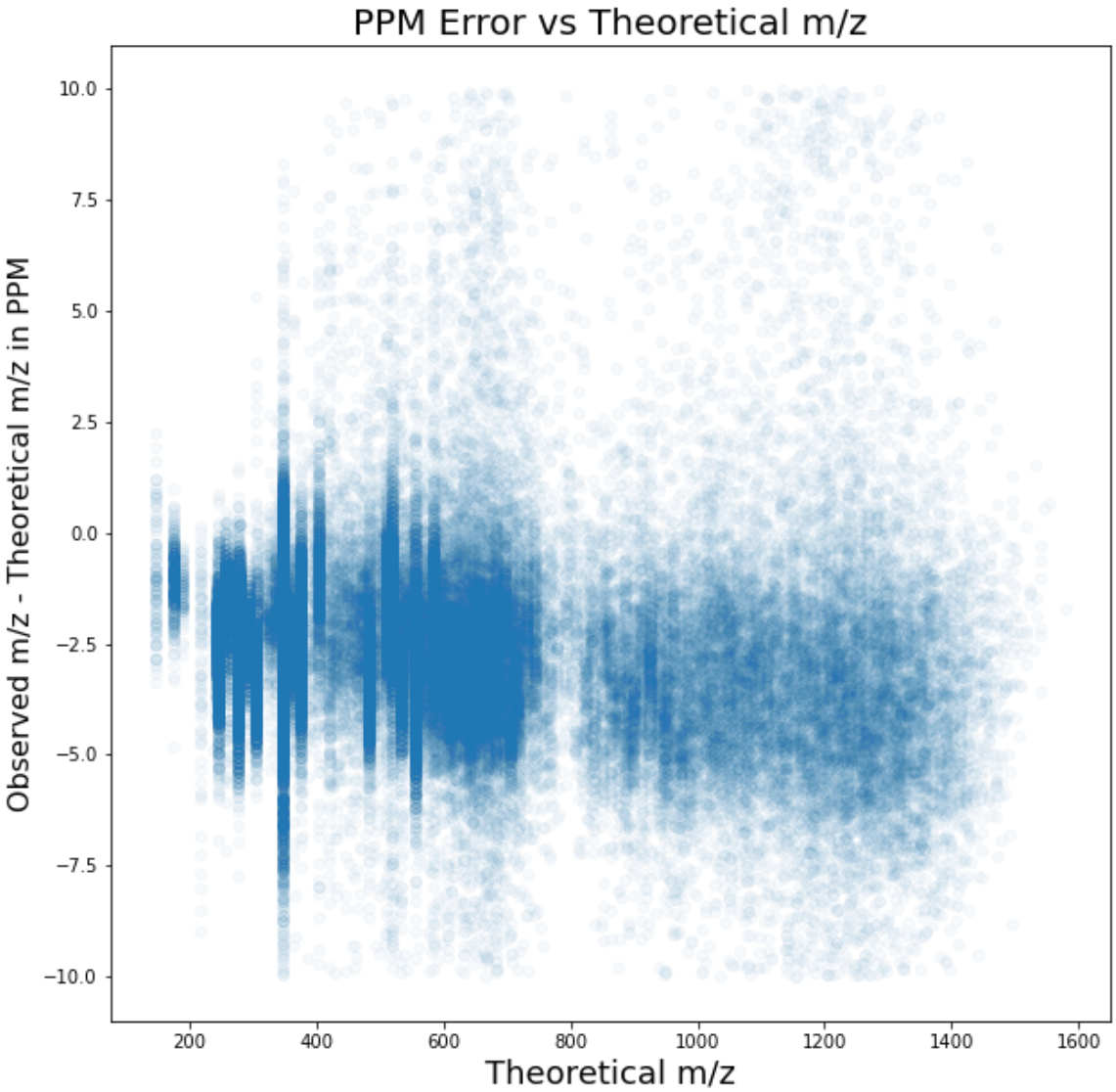

Figure 18. Matching Fragment Ion Ppm Errors by m/z.

Theoretical fragments from the identified peptide were searched within each ms2 scan. Any fragment which matched +/- 10ppm are plotted here. In this label free CID-MSA run there, is a downward trend in the ppm error as the theoretical fragment m/z increases.

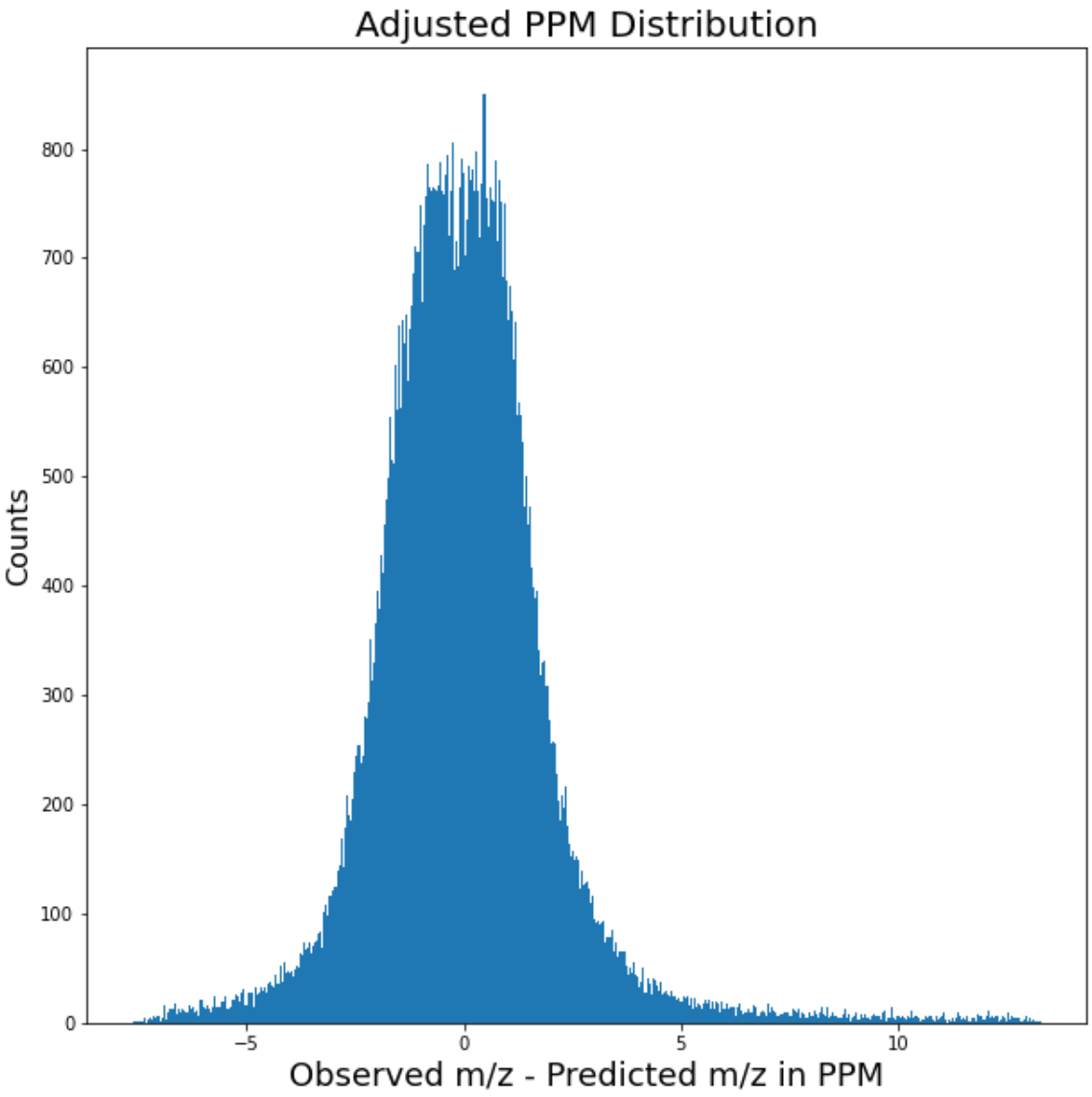

Figure 19. Adjusted PPM Histogram

This histogram shows the distribution of adjusted PPM errors for matching fragment ions after applying multiple linear regression to the label free CID-MSA run.

Site Localization with and Without Including

Metaphosphate Neutral Loss Site Determining Ions

With a new site localization algorithm described above, we sought to compare the effect of restricting the site determining ion list to only include phosphate harboring ions. This new phosphorylation localization is referred to as NewScore. When running this algorithm while considering all site determining ions, or only those with intact phosphate modifications, we see drastically different results. As shown in figure 20, when considering all site determining ions (SDI) the FLR at equivalent score cutoffs was substantially higher across all runs compared to only considering phosphate harboring site determining ions. For this library, the effect is most pronounced in the label free higher energy HCD runs. As compared with figures 5 and 6, running this site localization algorithm with only phosphate harboring site determining ions can more reproducibly control the FLR at the same score across runs.

However, restricting the site determining ion list to only phosphate harboring ions cuts the potential list of site determining ion list in half. This has an adverse effect on the ability to recover localized sites. As shown in figure 21, including only phosphate harboring site determining ions results in lower FLR values, but also reduces the number of localized phosphorylation sites. These data suggest that restricting the site determining ion list to consider only phosphate harboring ions may be necessary to achieve a consistently low FLR at a static score cutoff across a diversity of instrument and sample preparation techniques. Fragmentation modes, collisional energies, and algorithms used in data analysis should be carefully optimized to recover the greatest number of confidently localized sites. In addition to the FLR, it is also important to control the FDR on phosphoproteomics datasets.

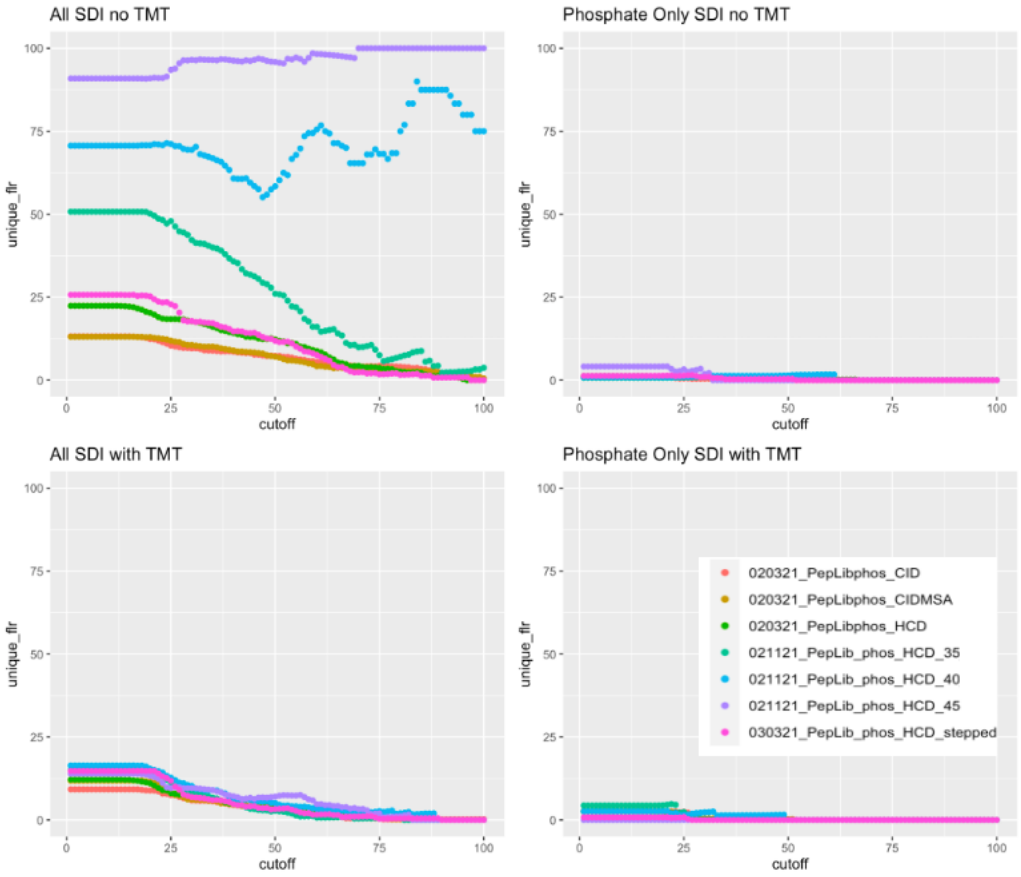

Figure 20. False Localization Rate with Site Determining Ion Restriction.

In all panels, the unique site FLR was calculated as the fraction of incorrectly localized sites with site localization scores greater than or equal to the cutoff. The top panels are from label free runs, and the bottom panels are from TMT labeled runs. The left panels are generated while considered all site determining ions, whereas the right panels are generated with only phosphate harboring site determining ions.

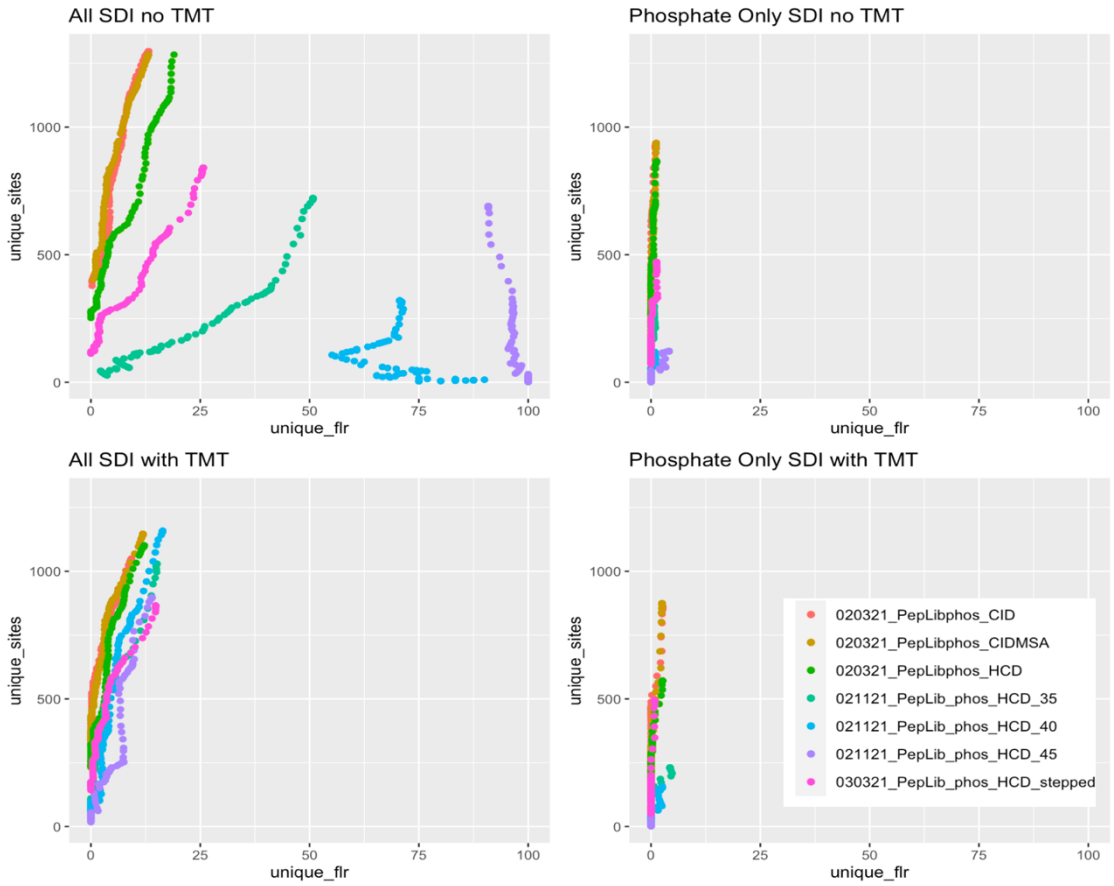

Figure 21. Recovery vs FLR Across Runs with and without site Determining Ions.

Shown above are the number of unique sites at a particular FLR. Score cutoffs of 1 to 100 are shown in increments of 1. Panel arrangement is the same as in figure 20.

False Discovery Rate

Using the datasets described previously, we aimed to assess the FDR further. These libraries were all searched and filtered with a target-decoy approach, to achieve a peptide level FDR of 1%. As described above, this library should only contain monophosphopeptides. However, we found that multiply phosphorylated identifications were often passing our LDA filtering at a 1% FDR. Additionally, there was significant

variability between runs in the fraction of multiply phosphorylated peptide IDs which passed LDA filtering at a 1% threshold. By comparing the number of multiply phosphorylated to the number of singly phosphorylated peptide IDs, we can generate a pseudo-FDR. In theory the pseudo-FDR should reflect the intended FDR from the target decoy approach, however figure 22 shows this is not the case for many of the runs.

Figure 22. Pseudo-FDR Variability Across Runs

The Pseudo-FDR is plotted here across all runs. Pseudo-FDR was calculated as the percentage of identifications which map to multiply phosphorylated full length identifications from the library database. PSMs were filtered to a 1% FDR using a target-decoy LDA approach.

Incorrect peptide identification will confound the ability of localization algorithms to correctly localize phosphate modifications. This variability in pseudo-FDR suggests additional filtering steps might be necessary to control dataset-level FDR and FLR. In addition to controlling the FDR and FLR, TMT-based phosphoproteomics datasets also require sufficient reporter ion signal intensity to be able to accurately quantify relative expression levels across samples.

TMT-Reporter Ion Signal

Another consideration when optimizing parameters for TMT-based phosphoproteomics experiments is how to best balance the instrument conditions for peptide identifications, localization, and generation of reporter ions for quantification purposes. Increasing the collisional energy of the HCD fragmentation will generate more reporter ion signal, however it can cause over fragmentation of peptide backbones. While we see a relatively higher fraction of confounding neutral loss containing fragments in the higher energy HCD runs, we also see that those runs tend to produce more reporter ion signal as shown in Figure 23.

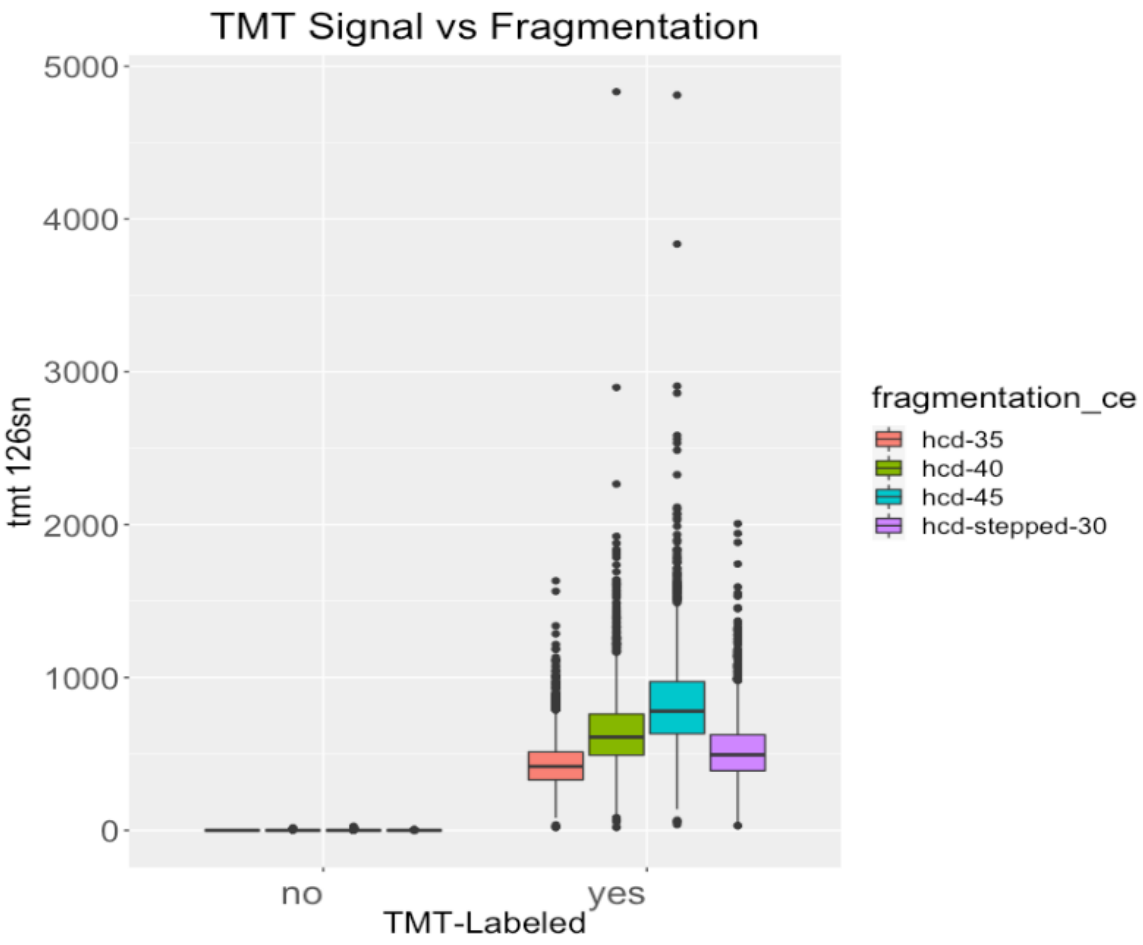

Figure 23. TMT Reporter Ion Signal Intensity by Method

Shown in this plot are the distribution of reporter ion signal to noise values extracted from the HCD fragmentation runs. The HCD30 run is excluded.

Taken together, these data suggest that when performing HCD, MS2 based TMT-phosphoproteomics analysis, the effects modulating the collisional energy has on the FDR, FLR, and reporter ion intensities should all be carefully considered.

IV. Discussion

Given the capability of phosphoprotoeomic LC-MS experiments to observe tens of thousands of phosphorylation sites derived from thousands of proteins in a single experiment, it is critical for computational tools to evolve with advances in sample preparation and instrumentation. Some of the previously developed algorithms for site localization assume that tandem HPO_3 neutral loss, and peptide backbone fragmentation do not co-occur (Beausoleil, Villen et al. 2006, Olsen, Blagoev et al. 2006). However, these fragment ions can adversely affect the ability to accurately localize the phosphate modification. Through the analysis of a library with over 2,000 phosphopeptides, with and without TMT-labeling, using a variety of fragmentation techniques, we have shown concurrent HPO_3 neutral loss and peptide backbone fragmentation are associated with an increased FLR using widely adopted localization algorithms.

Across the variety of instrumentation techniques we analyzed, CID, CID-MSA, and lower collisional energy HCD fragmentation techniques produced the least amount of problematic concurrent HPO_3 neutral loss and peptide backbone fragmentation. When applying higher energy HCD fragmentation, we observed an increase in the signal for fragment ions with HPO_3 neutral loss compared to fragment ions with intact phosphate. This relative increase in signal for HPO_3 neutral loss containing fragment ions correlated with elevated rates of incorrectly localized assignments at commonly used A-Score and MaxQuant localization score thresholds.

As a proof of principle, we developed and applied a localization algorithm to further assess the confounding effects of HPO_3 neutral loss. When only considering phosphate harboring ions as site determining ions, this algorithm was able to more reproducibly control the FLR. However, this approach resulted in a decrease in the number of recovered localized phosphorylation sites. Taken together, our results suggest that concurrent HPO_3 neutral loss and peptide backbone fragmentation can be problematic for commonly utilized data collection and analysis methods.

Next Steps

As the size of phosphoproteomics datasets grow, it becomes increasingly important to accurately control the dataset level FDR and FLR. In the analysis of a phosphopeptides library, the FLR can be empirically determined. However, most LC-MS experiments are not focused on analyzing synthetic standards. It is instead common to instead apply a localization score cutoff to control the FLR. Our data suggests that static localization score cutoffs do not produce consistent FLR's across a diversity of conditions. Instead, using a decoy residue approach for tracking and controlling the FLR on individual datasets may be necessary (Chalkley and Clauser 2012). These datasets also highlight the complexity and variability in the fragmentation patterns of phosphopeptides.

The complex fragmentation patterns pose a challenge for both the identification and localization of phosphopeptides. Recent advances in deep learning algorithms have been applied to predict peptide fragmentation patterns. Practically, these predicted peptide fragmentation patterns are useful to compare against experimentally observed fragmentation patterns to provide further support for identifications. This approach has

been applied to non-phosphorylated peptides and has been demonstrated to greatly reduce the dataset-level FDR (Gessulat, Schmidt et al. 2019). More recently this has been extended to the analysis of the phosphoproteome (Yang, Horvatovich et al. 2021). Our data suggests that when applying such peptide fragmentation prediction tools to phosphopeptides, a diversity of fragment ions should be included, such as the generation of peptide precursor neutral loss, and tandem neutral loss and peptide backbone fragmentation: including H_3PO_4 and HPO_3 neutral losses. Our data also shows the ability of a given phosphopeptide to generate these different fragment ions is related to the peptide sequence, chemical labeling, and dissociation method. A deep-learning model which can effectively predict the relative intensities of these species could be useful to more effectively control the FDR and FLR in phosphoproteomics datasets.

Limitations

While a variety of conditions were tested in this thesis, the synthetic library was relatively small, and both the fragmentation techniques and the algorithms tested were not comprehensive. Surveying a more complex synthetic library would enable a more informed selection of ideal fragmentation conditions and provide a more thorough comparison of localization algorithms. While our library analyzed here did contain over 2,000 phosphopeptides, there was only one seed sequence analyzed. The peptides here were all singly phosphorylated: a similar analysis of multiply phosphorylated peptides should be performed. Additionally, the newer TMTpro reagents may affect peptide fragmentation patterns, and warrants further analysis. Likewise, fragmentation techniques such as ETD could be similarly analyzed in future experiments.

References

Aguiar, M., W. Haas, S. A. Beausoleil, J. Rush and S. P. Gygi (2010). "Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets." J Proteome Res **9**(6): 3103-3107.

Albuquerque, C. P., M. B. Smolka, S. H. Payne, V. Bafna, J. Eng and H. Zhou (2008). "A multidimensional chromatography technology for in-depth phosphoproteome analysis." Mol Cell Proteomics **7**(7): 1389-1396.

Bailey, C. M., S. M. Sweet, D. L. Cunningham, M. Zeller, J. K. Heath and H. J. Cooper (2009). "SLoMo: automated site localization of modifications from ETD/ECD mass spectra." J Proteome Res **8**(4): 1965-1971.

Baker, P. R., J. C. Trinidad and R. J. Chalkley (2011). "Modification site localization scoring integrated into a search engine." Mol Cell Proteomics **10**(7): M111 008078.

Beausoleil, S. A., J. Villen, S. A. Gerber, J. Rush and S. P. Gygi (2006). "A probability-based approach for high-throughput protein phosphorylation analysis and site localization." Nat Biotechnol **24**(10): 1285-1292.

Bekker-Jensen, D. B., O. M. Bernhardt, A. Hogrebe, A. Martinez-Val, L. Verbeke, T. Gandhi, C. D. Kelstrup, L. Reiter and J. V. Olsen (2020). "Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries." Nat Commun **11**(1): 787.

Bhattacharyya, M., Y. K. Lee, S. Muratcioglu, B. Qiu, P. Nyayapati, H. Schulman, J. T. Groves and J. Kuriyan (2020). "Flexible linkers in CaMKII control the balance between activating and inhibitory autophosphorylation." Elife **9**.

Boersema, P. J., S. Mohammed and A. J. Heck (2009). "Phosphopeptide fragmentation and analysis by mass spectrometry." J Mass Spectrom **44**(6): 861-878.

Chalkley, R. J. and K. R. Clauser (2012). "Modification site localization scoring: strategies and performance." Mol Cell Proteomics **11**(5): 3-14.

Cox, J., M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj and M. Mann (2014). "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ." Mol Cell Proteomics **13**(9): 2513-2526.

Cui, L. and G. E. Reid (2013). "Examining factors that influence erroneous phosphorylation site localization via competing fragmentation and rearrangement reactions during ion trap CID-MS/MS and -MS(3.)." Proteomics **13**(6): 964-973.

Deribe, Y. L., T. Pawson and I. Dikic (2010). "Post-translational modifications in signal integration." Nat Struct Mol Biol **17**(6): 666-672.

Diedrich, J. K., A. F. Pinto and J. R. Yates, 3rd (2013). "Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot." J Am Soc Mass Spectrom **24**(11): 1690-1699.

Elias, J. E. and S. P. Gygi (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." Nat Methods **4**(3): 207-214.

Elias, J. E. and S. P. Gygi (2010). "Target-decoy search strategy for mass spectrometry-based proteomics." Methods Mol Biol **604**: 55-71.

Eng, J. K., T. A. Jahan and M. R. Hoopmann (2013). "Comet: an open-source MS/MS sequence database search tool." Proteomics **13**(1): 22-24.

Eng, J. K., A. L. McCormack and J. R. Yates (1994). "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." J Am Soc Mass Spectrom **5**(11): 976-989.

Everley, R. A., E. L. Huttlin, A. R. Erickson, S. A. Beausoleil and S. P. Gygi (2017). "Neutral Loss Is a Very Common Occurrence in Phosphotyrosine-Containing Peptides Labeled with Isobaric Tags." J Proteome Res **16**(2): 1069-1076.

Gessulat, S., T. Schmidt, D. P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, U. Reimer, H. C. Ehrlich, S. Aiche, B. Kuster and M. Wilhelm (2019). "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning." Nat Methods **16**(6): 509-518.

Giansanti, P., T. T. Aye, H. van den Toorn, M. Peng, B. van Breukelen and A. J. Heck (2015). "An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas." Cell Rep **11**(11): 1834-1843.

Goloborodko, A. A., L. I. Levitsky, M. V. Ivanov and M. V. Gorshkov (2013). "Pyteomics--a Python framework for exploratory data analysis and rapid software prototyping in proteomics." J Am Soc Mass Spectrom **24**(2): 301-304.

Harris, C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Rio, M. Wiebe, P. Peterson, P. Gerard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant (2020). "Array programming with NumPy." Nature **585**(7825): 357-362.

Hauser, A. S., S. Chavali, I. Masuho, L. J. Jahn, K. A. Martemyanov, D. E. Gloriam and M. M. Babu (2018). "Pharmacogenomics of GPCR Drug Targets." Cell **172**(1-2): 41-54 e19.

Hornbeck, P. V., J. M. Kornhauser, V. Latham, B. Murray, V. Nandhikonda, A. Nord, E. Skrzypek, T. Wheeler, B. Zhang and F. Gnad (2019). "15 years of PhosphoSitePlus(R): integrating post-translationally modified sites, disease variants and isoforms." Nucleic Acids Res **47**(D1): D433-D441.

Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering **9**(3): 90-95.

Huttlin, E. L., M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausoleil, J. Villen, W. Haas, M. E. Sowa and S. P. Gygi (2010). "A tissue-specific atlas of mouse protein phosphorylation and expression." Cell **143**(7): 1174-1189.

Jedrychowski, M. P., E. L. Huttlin, W. Haas, M. E. Sowa, R. Rad and S. P. Gygi (2011). "Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics." Mol Cell Proteomics **10**(12): M111 009910.

Johnson, H. and F. M. White (2012). "Toward quantitative phosphotyrosine profiling in vivo." Semin Cell Dev Biol **23**(8): 854-862.

Kim, S., N. Gupta, N. Bandeira and P. A. Pevzner (2009). "Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra." Mol Cell Proteomics **8**(1): 53-69.

Levitsky, L. I., J. A. Klein, M. V. Ivanov and M. V. Gorshkov (2019). "Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework." J Proteome Res **18**(2): 709-714.

Li, J., Z. Cai, R. D. Bomgarden, I. Pike, K. Kuhn, J. C. Rogers, T. M. Roberts, S. P. Gygi and J. A. Paulo (2021). "TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing." J Proteome Res **20**(5): 2964-2972.

Li, J., J. A. Paulo, D. P. Nusinow, E. L. Huttlin and S. P. Gygi (2019). "Investigation of Proteomic and Phosphoproteomic Responses to Signaling Network Perturbations Reveals Functional Pathway Organizations in Yeast." Cell Rep **29**(7): 2092-2104 e2094.

Lim, M. Y., J. A. Paulo and S. P. Gygi (2019). "Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model." J Proteome Res **18**(11): 4020-4026.

Liu, Y., A. Beyer and R. Aebersold (2016). "On the Dependency of Cellular Protein Levels on mRNA Abundance." Cell **165**(3): 535-550.

Martin, D. B., J. K. Eng, A. I. Nesvizhskii, A. Gemmill and R. Aebersold (2005). "Investigation of neutral loss during collision-induced dissociation of peptide ions." Anal Chem **77**(15): 4870-4882.

Marx, H., S. Lemeer, J. E. Schliep, L. Matheron, S. Mohammed, J. Cox, M. Mann, A. J. Heck and B. Kuster (2013). "A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics." Nat Biotechnol **31**(6): 557-564.

McAlister, G. C., D. P. Nusinow, M. P. Jedrychowski, M. Wuhr, E. L. Huttlin, B. K. Erickson, R. Rad, W. Haas and S. P. Gygi (2014). "MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes." Anal Chem **86**(14): 7150-7158.

Mertins, P., L. C. Tang, K. Krug, D. J. Clark, M. A. Gritsenko, L. Chen, K. R. Clauser, T. R. Clauss, P. Shah, M. A. Gillette, V. A. Petyuk, S. N. Thomas, D. R. Mani, F. Mundt, R. J. Moore, Y. Hu, R. Zhao, M. Schnaubelt, H. Keshishian, M. E. Monroe, Z. Zhang, N. D. Udeshi, D. Mani, S. R. Davies, R. R. Townsend, D. W. Chan, R. D. Smith, H. Zhang, T. Liu and S. A. Carr (2018). "Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry." Nat Protoc **13**(7): 1632-1661.

Olsen, J. V., B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen and M. Mann (2006). "Global, in vivo, and site-specific phosphorylation dynamics in signaling networks." Cell **127**(3): 635-648.

Palumbo, A. M., S. A. Smith, C. L. Kalcic, M. Dantus, P. M. Stemmer and G. E. Reid (2011). "Tandem mass spectrometry strategies for phosphoproteome analysis." Mass Spectrom Rev **30**(4): 600-625.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research **12**: 2825--2830.

Phanstiel, D. H., J. Brumbaugh, C. D. Wenger, S. Tian, M. D. Probasco, D. J. Bailey, D. L. Swaney, M. A. Tervo, J. M. Bolin, V. Ruotti, R. Stewart, J. A. Thomson and J. J. Coon (2011). "Proteomic and phosphoproteomic comparison of human ES and iPS cells." Nat Methods **8**(10): 821-827.

Possemato, A. P., J. A. Paulo, D. Mulhern, A. Guo, S. P. Gygi and S. A. Beausoleil (2017). "Multiplexed Phosphoproteomic Profiling Using Titanium Dioxide and Immunoaffinity Enrichments Reveals Complementary Phosphorylation Events." J Proteome Res **16**(4): 1506-1514.

Potel, C. M., S. Lemeer and A. J. R. Heck (2019). "Phosphopeptide Fragmentation and Site Localization by Mass Spectrometry: An Update." Anal Chem **91**(1): 126-141.

Rappsilber, J., Y. Ishihama and M. Mann (2003). "Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics." Anal Chem **75**(3): 663-670.

Rikova, K., A. Guo, Q. Zeng, A. Possemato, J. Yu, H. Haack, J. Nardone, K. Lee, C. Reeves, Y. Li, Y. Hu, Z. Tan, M. Stokes, L. Sullivan, J. Mitchell, R. Wetzel, J. Macneill, J. M. Ren, J. Yuan, C. E. Bakalarski, J. Villen, J. M. Kornhauser, B. Smith, D. Li, X. Zhou, S. P. Gygi, T. L. Gu, R. D. Polakiewicz, J. Rush and M. J. Comb (2007). "Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer." Cell **131**(6): 1190-1203.

Savitski, M. M., S. Lemeer, M. Boesche, M. Lang, T. Mathieson, M. Bantscheff and B. Kuster (2011). "Confident phosphorylation site localization using the Mascot Delta Score." Mol Cell Proteomics **10**(2): M110 003830.

Savitski, M. M., M. Wilhelm, H. Hahne, B. Kuster and M. Bantscheff (2015). "A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets." Mol Cell Proteomics **14**(9): 2394-2404.

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

Sefton, B. M., T. Hunter, K. Beemon and W. Eckhart (1980). "Evidence that the phosphorylation of tyrosine is essential for cellular transformation by Rous sarcoma virus." Cell **20**(3): 807-816.

Sharma, K., R. C. D'Souza, S. Tyanova, C. Schaab, J. R. Wisniewski, J. Cox and M. Mann (2014). "Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling." Cell Rep **8**(5): 1583-1594.

Steen, H., B. Kuster, M. Fernandez, A. Pandey and M. Mann (2001). "Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode." Anal Chem **73**(7): 1440-1448.

Steen, H. and M. Mann (2004). "The ABC's (and XYZ's) of peptide sequencing." Nat Rev Mol Cell Biol **5**(9): 699-711.

Suni, V., S. Y. Imanishi, A. Maiolica, R. Aebersold and G. L. Corthals (2015). "Confident site localization using a simulated phosphopeptide spectral library." J Proteome Res **14**(5): 2348-2359.

Taus, T., T. Kocher, P. Pichler, C. Paschke, A. Schmidt, C. Henrich and K. Mechtler (2011). "Universal and confident phosphorylation site localization using phosphoRS." J Proteome Res **10**(12): 5354-5362.

Thingholm, T. E. and M. R. Larsen (2016). "Phosphopeptide Enrichment by Immobilized Metal Affinity Chromatography." Methods Mol Biol **1355**: 123-133.

Tholey, A., J. Reed and W. D. Lehmann (1999). "Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues." J Mass Spectrom **34**(2): 117-123.

Thompson, A., J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. Mohammed and C. Hamon (2003). "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS." Anal Chem **75**(8): 1895-1904.

Ting, L., R. Rad, S. P. Gygi and W. Haas (2011). "MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics." Nat Methods **8**(11): 937-940.

Tolonen, A. C. and W. Haas (2014). "Quantitative proteomics using reductive dimethylation for stable isotope labeling." J Vis Exp(89).

Tyanova, S., T. Temu and J. Cox (2016). "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics." Nat Protoc **11**(12): 2301-2319.

Udeshi, N. D., D. C. Mani, S. Satpathy, S. Fereshetian, J. A. Gasser, T. Svinkina, M. E. Olive, B. L. Ebert, P. Mertins and S. A. Carr (2020). "Rapid and deep-scale ubiquitylation profiling for biology and translational research." Nat Commun **11**(1): 359.

Villen, J. and S. P. Gygi (2008). "The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry." Nat Protoc **3**(10): 1630-1638.

Viner, R. I., T. Zhang, T. Second and V. Zabrouskov (2009). "Quantification of post-translationally modified peptides of bovine alpha-crystallin using tandem mass tags and electron transfer dissociation." J Proteomics **72**(5): 874-885.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and C. SciPy (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python." Nat Methods **17**(3): 261-272.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York.

Wiese, S., K. A. Reidegeld, H. E. Meyer and B. Warscheid (2007). "Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research." Proteomics **7**(3): 340-350.

Wolters, D. A., M. P. Washburn and J. R. Yates, 3rd (2001). "An automated multidimensional protein identification technology for shotgun proteomics." Anal Chem **73**(23): 5683-5690.

Yang, H., H. Chi, W. J. Zhou, W. F. Zeng, C. Liu, R. M. Wang, Z. W. Wang, X. N. Niu, Z. L. Chen and S. M. He (2018). "pSite: Amino Acid Confidence Evaluation for Quality Control of De Novo Peptide Sequencing and Modification Site Localization." J Proteome Res **17**(1): 119-128.

Yang, Y., P. Horvatovich and L. Qiao (2021). "Fragment Mass Spectrum Prediction Facilitates Site Localization of Phosphorylation." J Proteome Res **20**(1): 634-644.

Yu, Q., X. Shi, Y. Feng, K. C. Kent and L. Li (2017). "Improving data quality and preserving HCD-generated reporter ions with EThcD for isobaric tag-based quantitative proteomics and proteome-wide PTM studies." Anal Chim Acta **968**: 40-49.

Zanivan, S., M. Krueger and M. Mann (2012). "In vivo quantitative proteomics: the SILAC mouse." Methods Mol Biol **757**: 435-450.