DISCOVERING MECHANISMS IN THE GAS PHASE CHEMISTRY OF PHOSPHORYLATED PEPTIDES THROUGH STATISTICAL LEARNING

by

ROBERT ALEXANDER BROWN

B.S., Yale University, 2006

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy

Department of Chemistry and Biochemistry

2014
This thesis entitled:

Discovering mechanisms in the gas phase chemistry of phosphorylated peptides through statistical learning.

written by Robert A Brown

has been approved for the Department of Chemistry and Biochemistry

__________________________________________
Natalie Ahn

__________________________________________
William Old

Date ______________

The final copy of this thesis has been examined by the signatories, and we

Find that both the content and the form meet acceptable presentation standards

Of scholarly work in the above mentioned discipline.
Brown, Robert Alexander (Ph.D., Chemistry and Biochemistry)

Discovering mechanisms in the gas phase chemistry of phosphorylated peptides through statistical learning.

Thesis directed by Professor of Distinction Natalie G. Ahn

Mass spectrometry has become the preeminent method in proteomics, making identification of peptides from their pattern of gas phase fragmentation of interest. Current state-of-the-art peptide identification discards intensity information in peptide spectra; however, improved understanding of the gas phase chemistry allows prediction of ion intensities. Improving the prediction of ion intensities requires uncovering novel fragmentation mechanisms and incorporating them algorithmically into programs designed to computationally model spectra. One such mechanism, discussed here, involves unexpectedly avid cleavage when proline is at the second position. Modeling this mechanism indicates that the gas phase chemistry is not catalyzed by charge. No such mechanism has been previously identified. Incorporating the predicted intensities improves peptide identification; however, certain classes of reaction, such as those exhibited in phosphopeptides continue to confound identification.

Reversible protein phosphorylation modulates almost all aspects of cellular function, making understanding of this post-translational modification essential in modern proteomics. Unfortunately, the phosphoester bond is very reactive, leading to spectra that are convoluted by the neutral loss of \( \text{H}_3\text{PO}_4 \). By using a dataset tenfold larger than previous studies, I am able to rigorously analyze how sequence affects neutral loss. In contrast to cleavage of peptide bonds, I find that neutral loss of \( \text{H}_3\text{PO}_4 \) is affected significantly by distal sites, most notably the basic residues and N-terminus. Previous studies have suggested that basic residues directly catalyze neutral loss and initial analysis shows enhanced neutral loss near bases. However, in an example of Simpson’s paradox, when we stratify the spectra by charge-mobility, we find evidence for the converse, that nearby bases inhibit neutral loss regardless of mobility class. In mobile proton spectra, the N-terminus is the strongest predictor of high neutral loss, with proximity to the N-terminus essential for peptides to exhibit the highest levels of neutral loss. These observations suggest a model in which the phosphate is always in complex with protons immobilized at basic sites though the adoption of secondary structure, in contrast to the prevailing model, which suggests direct protonation of the phosphate. Further evidence suggests that this mechanism may be general to all neutral loss reactions from peptide side chains.
In memory of

KATHERYN A. H. RESING

(1944-2009)

whose passion and intellect started me on this path,

who left too soon,

and

whose council is still missed.
I would like to acknowledge the following people for their assistance in this project:

Katheryn Resing for starting the journey.

Natalie Ahn and William Old for advising me after she passed away.

Chia-yu Yen, Stephane Houel, and Shaojun Sun for collaborating with me and teaching me so much.

My family for putting up with me through the good times and the bad.
Contents

Chapter 1: Shotgun proteomics and the problem of peptide identification ........................................... 1

 Section 1: ‘Bottom-up’ proteomics ........................................................................................................... 2

 Section 2: Peptide identification ........................................................................................................... 4

 The basics of peptide fragmentation ..................................................................................................... 4

 Methods of identifying peptides from their MS/MS spectra .................................................................. 6

 Statistical models ..................................................................................................................................... 8

 Interpretative models ............................................................................................................................... 9

 Stochastic models .................................................................................................................................. 10

 Descriptive models .................................................................................................................................. 11

 Current limitations of search algorithms ............................................................................................. 12

 Section 3: Peptide fragmentation under the mobile proton model ......................................................... 12

 The mobile proton model ....................................................................................................................... 12

 Pathways in competition model ......................................................................................................... 15

 Section 4: An algorithm for the kinetic prediction of peptide fragmentation ...................................... 16

 Generation of theoretical spectra using the pathways in competition model ..................................... 16

 Section 5: The importance and fragmentation of phosphopeptides .................................................... 19

 Chapter 2: Implementing knowledge of chemistry to predict the fragmentation pattern of peptides ................................................................................................................................. 23

 Section 1: Uncovering novel fragmentation mechanisms ..................................................................... 23
Section 2: Using fragmentation intensities to improve identification ................................. 29

Chapter 3: Using phosphopeptide libraries to investigate the neutral loss of phosphate ........ 34

Section 1: The development of combinatorial phosphopeptide libraries .................................... 34

Initial tests with simple combinatorial libraries ......................................................................... 34

Scaling up to more diverse phosphopeptide libraries .................................................................. 39

Section 2: Analyzing the fragmentation patterns in phosphopeptide libraries .......................... 42

Creating a data structure for fragmentation analysis .................................................................... 42

Section 3: Variability in neutral loss patterns based on instrumentation used ............................. 46

Section 4: Assessing determinants of neutral loss ....................................................................... 50

Chapter 4: Understanding the variations in neutral loss from phosphopeptides ...................... 54

Section 1: Evaluation of the types and amounts of neutral loss .................................................. 55

Characterization of types of neutral loss ..................................................................................... 55

Variation in neutral loss between peptides .................................................................................. 59

Section 2: Effects of local sequence on neutral loss ..................................................................... 60

Visualizing neutral loss patterns ................................................................................................. 62

Simpson’s paradox: controlling for charge mobility, proximal basic residues suppress H₃PO₄ neutral loss ........................................................................................................................................ 66

The N-Terminus is the primary driver of neutral loss in mobile proton peptides ....................... 69
Proline provides a competitive pathway and reduces backbone flexibility required for distal basic interactions .................................................................................................................. 73

Aspartic and glutamic acid ........................................................................................................ 75

Section 3: Neutral loss of water and ammonia from threonine and other amino acids
interferes with the loss of H₃PO₄ .............................................................................................. 76

Chapter 5: Future work suggested by these studies .................................................................. 81

Section 1: Statistically robust scoring for intensity-based spectral matching......................... 81

Section 2: Predicting the intensities of neutral losses through prediction of secondary structures .................................................................................................................................. 87

Chapter 6: Conclusions ................................................................................................................. 91

Appendix A: Methods for chromatography and mass spectrometry ........................................... 97

WM239A dataset: Samples from cellular extract ........................................................................ 97

Mass spectrometry ....................................................................................................................... 98

Phosphopeptide identification ..................................................................................................... 99

Citations .......................................................................................................................................... 100
Table of Figures

Figure 1. Typical Workflow for Shotgun Proteomics ............................................................... 3

Figure 2. The nomenclature of peptide sequence cleavages. .................................................. 5

Figure 3. An example of an annotated mass spectrum. ............................................................ 5

Figure 4. The categories of database search algorithms along with several commercially
available programs that implement them. ................................................................. 8

Figure 5. The mechanism of mobile proton-directed sequence cleavage into b or y-ions .......... 14

Figure 6. The mechanism of charge-directed cleavage adjacent to basic residues ............... 14

Figure 7. A hierarchical scheme of peptide fragmentation pathways. ................................... 16

Figure 8. The protonation sites assumed in Mass Analyzer. .................................................. 17

Figure 9. The schema of fragmentation in Mass Analyzer. .................................................... 18

Figure 10. Proposed mechanisms for the neutral loss of phosphoric acid. ............................ 22

Figure 11. Error in prediction of N2 cleavage by amino acid at the second position. ............. 24

Figure 12. Mechanism of charge directed diketopiperazine formation ................................ 25

Figure 13. Charge remote mechanism for diketopiperazine formation .................................. 26

Figure 14. Comparison of charge remote and charge directed mechanisms of N2 cleavage ... 27
Figure 15. Correction of predicted intensity after addition of N2 mechanism. ......................... 28

Figure 16. Examples of spectral prediction with and without the N2 mechanism. .................... 29

Figure 17. Comparison of similarity score to Mascot search engine....................................... 30

Figure 18. Comparison of ranked similarity score to Mascot search engine. ......................... 31

Figure 19. Performance of ranked similarity score when ignoring neutral loss ions.................. 32

Figure 20. Combinatorial synthesis yields diverse peptide libraries. ..................................... 35

Figure 21. The sequences of the first four combinatorial libraries used in this paper. ............ 36

Figure 22. Contour plot showing identifications from the library phosl1............................... 37

Figure 23. The effects of phosphatase treatment on phosphopeptide libraries. ..................... 38

Figure 24. The sequences of the complex combinatorial libraries used in this paper.............. 41

Figure 25. Contour plot of a combinatorial peptide library with 720 expected phosphopeptides.
............................................................................................................................................... 42

Figure 26. Schema of the relational database used for the storage of spectral information and
annotations. .................................................................................................................................. 44

Figure 27. Correlation if neutral loss in the Qtrap and LTQ mass spectrometers....................... 47

Figure 28. Cumulative distribution of neutral loss from the parent in the LTQ and Qtrap........ 48
Figure 29. Comparison of the amount of intact and neutral loss sequence ions in the Qtrap and LTQ. ................................................................. 49

Figure 30. The quality of classification of neutral loss by naive Bayesian classifier ............... 51

Figure 31. Distal amino acids can affect neutral loss more dramatically than proximal amino acids. ........................................................................................................... 53

Figure 32. The distribution of neutral loss ........................................................................ 56

Figure 33. Assessment of neutral loss types in non-tryptic peptides................................. 58

Figure 34. Spectral databases compress the observed range of intensities ...................... 62

Figure 35. The effects of local sequence on neutral loss ................................................... 66

Figure 36. The effects of flanking sequence on neutral loss from immobile proton spectra. ..... 67

Figure 37. The effects of the N-terminus on neutral loss .................................................. 70

Figure 38. The effects of proline on neutral loss .............................................................. 74

Figure 39. The interaction between the neutral loss of phosphoric acid and the neutral loss of water ............................................................................................................. 78
Chapter 1: Shotgun proteomics and the problem of peptide identification

Since the completion of the Human Genome Project, 'omics'-based biology has become increasingly focused on how the now-known genome sequence leads to cellular and organismal phenotype. Because the majority of the information contained within the genome ultimately expresses through the regulated translation and modification of the proteins encoded, comprehensive protein analysis, or proteomics, is an essential intermediary understanding molecular biology.

The field of proteomics has expanded rapidly since the late 1980s, largely due to the creation and adoption of several new technologies to facilitate its development. The most important breakthroughs in this context are the invention of electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). Each of these techniques introduces the possibility of ionizing even large peptides into the gas-phase without disturbing their structure and makes mass spectroscopic (MS) analysis of proteins possible. The use of MS-based proteomics, instead of previous chemical sequencing techniques, has the dual effect of reducing the amount of protein and time needed for proteomic analyses. More fundamentally, it allows many studies to be conducted which were not previously possible. Though most current applications use only a fraction of the available data, mass spectroscopic techniques give more information than chemical sequencing and have already been used to assay primary, secondary, tertiary, and even quaternary structures of proteins. While most of the work to date focuses exclusively on the sequencing of primary structures, there is great promise for the
mass spectroscopic proteomics to bridge the gap between understanding the sequence and structure of proteins.

Section 1: ‘Bottom-up' proteomics

Shotgun proteomics, in contrast to targeted proteomics, represents the attempt to fully capture the protein state of cells without any prior assumptions about the proteins of interest. Consequently, rapid, unbiased sampling of proteins is of paramount importance. The theoretically simplest and most powerful method of shotgun proteomics is the direct analysis of intact proteins – the so-called ‘top-down' approach to proteomics. While this method preserves information about the intact protein and leaves a relatively simple mixture of proteins, intact proteins behave wildly variably in solution and are difficult to identify in the gas-phase. Consequently, the ‘top-down' approach has been relegated to a niche, though growing, role in shotgun proteomics. Instead, experimenters have favored a ‘bottom-up' approach, in which proteins are protease-digested into more manageable peptides prior to analysis.

The workflow of a bottom-up proteomics experiment (Figure 1.) typically involves harvesting proteins from cell culture, tissue, or biological fluid of interest. These proteins may optionally be put through one or more rounds of fractionation, both to simplify the mixture and to isolate the relevant proteins. Once proteins are adequately processed, they are exposed to protease. This step permanently destroys information about the intact protein; however, the resulting peptides are much easier to deal with and more impervious to experimental contamination. The typical protease for this analysis is trypsin, which yields peptides which typically contain only one lysine or arginine. The peptides are then fractionated one or more
times. The most common method of fractionation is reverse phase C18 chromatography, which separates based on the hydrophobicity of the peptide.

![Diagram of Shotgun Proteomics Workflow](image)

**Figure 1. Typical Workflow for Shotgun Proteomics**

The previous separations allow a greatly simplified mixture of peptides to be introduced into the mass spectrometer and ionized for gas-phase analysis. In the gas phase, peptides can be further separated by mass through the commission of an MS\(^1\) scan and isolation of the observed ions by a quadrupole mass filter. Provided the peptides are sufficiently separated, a single peptide, of known mass, is now contained in the quadrupole. This peptide can be fragmented by collisional heating (CID), yielding ion patterns that are characteristic of its
sequence. The correlation of this fragmentation pattern to the sequence of the peptide that generated is the subject of the remainder of this work. Once the peptide sequence is known, it is possible to infer the protein that provides that sequence.

**Section 2: Peptide identification**

**The basics of peptide fragmentation**

Peptides fragmenting under low-energy CID preferentially dissociate along the amide bond. Thus, all backbone cleavages result in two complementary fragments, one stretching from the N-terminus to the cleavage site and the other from the cleavage site to the C-terminus. Often only one is visible, as the other may not retain any charge. The N-terminal ions are traditionally labeled a, b, or c depending on which bond was broken, while the complementary C-terminal ions are labeled x, y, or z, respectively (Figure 2). Furthermore, each sequence ion numbered corresponding to the number of α-carbons it contains. By far the most common of these sequence ions are the b- and y-ions, which correspond to cleavages of the peptide bonds. These sequence ions, however, are only one of many fragmentations that occur in the peptide, and the experimental spectra can be very complex. One notable class of convoluting ions result from the neutral loss of amino acid side chains. Figure 3 is an example of a very good spectrum from a doubly-charged tryptic phosphopeptide. The annotations show the vast amount of information that is available in a spectrum, but also that even in high-quality spectra many expected ions may not be present. Note here that the b-ion series is only present in dehydrated form. Most spectra that are of significant concern in database-searching have far lower signal-to-noise ratios than are present here.
Figure 2. The nomenclature of peptide sequence cleavages. C-terminal ions are named x, y, or z, while N-terminal ions are named a, b, or c. Each is further numbered according to the number of α-carbons contained in the ion. (From Paizs & Suhai, 2005)

Figure 3. An example of an annotated mass spectrum. The parent is doubly charged KGpSTDQGkpTSAPKKEEGK from Xenopus Aurora A (Δ indicates dehydration). All but one sequence cleavage is present in at least one of the ion series.
Methods of identifying peptides from their MS/MS spectra

The complexity of peptide spectra ensures that increased computational power, while less-heralded than improvements in peptide ionization, is equally essential to the growth of proteomics, as it is required to facilitate the interpretation and analysis of spectra. Even the seemingly simple determination of peptide sequence from tandem MS (MS²) spectra is a nontrivial and computationally intensive problem. Experienced individuals can manually determine a sequence from a mass spectrum with relative ease. However, with protein mixtures as complex as whole cell extracts being analyzed by these techniques, high-throughput runs typically yield greater than a million spectra. Manual analysis on this scale is clearly unfeasible. As a result, powerful automated analysis of spectra is an important requirement of large-scale proteomics. Many computer programs have been introduced to solve this problem. The two most popular programs for these analyses are Matrix Science’s MASCOT and SEQUEST. Both programs work by comparing an input spectrum with theoretical spectra that they generate for sequences in a protein database.

The fundamental difficulty for database search algorithms is that whole-cell proteomes can contain many millions of possible distinct peptides, especially if wide ranges of posttranslational modifications or enzymatic cleavages are expected. The sheer number of peptides in cells simultaneously reduces the quality of spectra obtained and increases the need for precise analysis. The increase in the number of possible matches means that, in order to maintain a reasonable number of expected matches, the probability threshold for acceptance
must be much lower than 1%. At this acceptability threshold, computationally feasible methods of determining matches between spectra and sequences often strain the amount of information available for identification. Consequently, experimenters have three undesirable options. They can set a very stringent acceptance threshold and face the likelihood of false negatives, or they can reduce the stringency of their threshold and face the possibility of false positives and ambiguous matches, or they can manually reanalyze all ambiguous data. The first two options lead to incomplete or incorrect data; while the third is extremely time consuming, since as many as a fifth of spectra are likely to fall into the critical range.\textsuperscript{12} One goal of proteomics, therefore, must be to create algorithms that better utilize available spectral information in peptide identification.

Algorithms used to search mass spectra against proteomic databases can be classified into four general types: statistical, interpretative, stochastic, and descriptive (Figure 4).\textsuperscript{24} At the most basic level, all of these methods attempt to determine what identifying characteristics the mass spectrum of a given sequence will contain and produce a theoretical spectrum from that sequence. Statistical and stochastic methods may be described as top-down methods, focusing on statistical inference from previously curated data. Descriptive and interpretative methods, meanwhile, are bottom-up, basing their actions on a theoretical understanding of the way peptides fragment. Descriptive and statistical methods are generally concerned with the presence or absence of specific peaks, while interpretative and stochastic methods assess the relationships between the fragments that are present.
<table>
<thead>
<tr>
<th></th>
<th>Attributes Considered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Identity of peaks</strong></td>
</tr>
<tr>
<td><strong>Peak Prediction</strong></td>
<td><strong>Top-down (statistical)</strong></td>
</tr>
<tr>
<td></td>
<td>Mascot, Gygi tree</td>
</tr>
<tr>
<td><strong>Peak Prediction</strong></td>
<td><strong>Bottom-up (theoretical)</strong></td>
</tr>
<tr>
<td></td>
<td>SEQUEST, SONAR, SALSA, MassAnalyzer</td>
</tr>
<tr>
<td></td>
<td>PeptideSearch, MS-Seq, GutenTag</td>
</tr>
</tbody>
</table>

Figure 4. The categories of database search algorithms along with several commercially available programs that implement them. The methods are organized along two axes: whether the absolute mass of peaks or the difference between peaks is considered and whether the presence of peaks is predicted according to statistical analysis of previous data or by theories of peptide fragmentation.

**Statistical models**

Statistical models use empirical data to generate the statistical likelihood of a given fragment appearing in a sequence. Due to the stringent demands of database searching, this implies computationally and data intensive methods of machine learning to deduce characteristic fragmentation patterns from large pools of previously analyzed data. The tantalizing strength of such methods is that they should be capable of predicting the gas-phase chemistry of peptides regardless of the understanding of that chemistry. On the other hand, any cases that are not encountered in the training set may be predicted effectively randomly. Spectral quality is also of paramount importance, because without *a priori* assumptions, sequence peaks cannot be distinguished from noise. This implies that large data sets must be used for training. The commitment to the equipment, laboratory techniques, and instrument
settings used in the training set severely reduces the flexibility of laboratories that use this technique and the portability of any models.

**Interpretative models**

Interpretative models are a direct implementation of the logic that is usually used in the manual *de novo* sequencing. These techniques rely on the existence of “ladders” of ions with mass differences between peaks corresponding to the mass of an amino acid. From a group of these peaks a partial sequence tag may be determined. In practice this means that there needs to be a sequence of adjacent y-ions or b-ions from which to create a tag to query for in the database. For example, in Figure 2 the y-ions from $y_2$ to $y_7$ are visible. As such, interpretative models would deduce from the gaps between these peaks that either the sequence VPKEEG or GEEKPV (in the counterfactual case where the sequence corresponded to b-ions) was present in the peptide. The program can then search for this string in a protein database.

The fundamental problem with this technique is that in some cases peptides will not yield usable tags of sufficient length, while in others there may be ambiguous or even incorrect tags that may be recognized by the algorithm. One significant advantage of these methods is that they are error-tolerant. As these methods do not consider the actual masses of fragments but instead the differences in masses, incorrectly entered, mutated, or unexpectedly modified residues do not prevent the program from finding surrounding tags and identifying the peptide. PeptideSearch, for example, considers two distinct matches in a region confirmation of its identity, even if the parent mass does not agree with the assignment. Interpretative models at their most basic do not require any form of knowledge of peptides or mass spectroscopy other
than the masses of the amino acids. More recent and complex methods, such as GutenTag\textsuperscript{21} show the promise of using interpretative models coupled to empirical predictors of fragment intensity. Indeed, it is possible to very cheaply integrate any other method into interpretative models to predict the expected intensities of the fragments in the tag better. Such hybrid models could dramatically reduce the number of false positives and time spent searching for nonexistent tags by effectively and cheaply emulating the logic of trained observers.

**Stochastic models**

Stochastic models are statistical methods based on the development of conditional probabilities for the appearance of a fragmentation pattern, given the other fragments observed. SCOPE\textsuperscript{18} is probably the most famous stochastic model. It finds the probability that a given spectrum arises from a given peptide in a two-step process. Firstly, it estimates the likelihood of each possible fragmentation pattern of the peptide, based on a collection of previously annotated spectra. Then it determines the probability that the given fragmentation pattern gives rise to the observed spectrum. Finally, it takes the product of the two previous probabilities and sums across all possible fragmentation patterns. When \(p\) is a peptide sequence, \(F\) is a characteristic fragmentation pattern, and \(S\) is a spectrum this is formally written as:

\[
\psi(S, p) = \sum_{F \subseteq S(p)} \psi(S|F, p) \Pr(F|p)
\]

More recent models, such as OLAV\textsuperscript{17} and MassLynx, use Markov chains and hidden Markov models as more powerful methods of determining the probabilities of sequences emitting a
given spectrum. These models provide a conceptually powerful method of predicting whether a fragment should appear given that other fragments have appeared. Though, like in statistical models, high-quality spectra in the training set are imperative, stochastic methods have the advantage of being relatively impervious to noisy data in test sets. Due to this ability to ignore spurious data, these techniques may also be ideal for deconvoluting chimerical spectra. It seems that this method would also be very well suited for the prediction of the creation of internal fragment ions and of neutral losses of water and ammonia; however, such methods have not been implemented.

Unfortunately both stochastic and interpretive models, generally, lack the statistical power necessary for whole-proteome searches. They are consequently mostly used for pre-filtering of candidate lists prior to additional analysis. In doing so, they have the power to allow computationally expensive searches on unusually modified or otherwise unexpected species and extend the chances of novel discoveries in proteomics experiments.

**Descriptive models**

Descriptive algorithms are the workhorses of modern proteomics. They are based on models of the fragmentation chemistry of peptides. Theoretical spectra are generated for peptides in the database and then assayed for similarity with the experimental spectrum. Due to the computational costs of performing these calculations and recently emergence of understanding of peptide gas-phase chemistry, the models that are used today are generally very crude. For example, SEQUEST, the most commonly used descriptive algorithm, simply assigns a magnitude of 50 to all b- and y-ions, 25 to the masses +/- 1 Da from the b- and y-ions
(to catch slight mass inaccuracies), and 10 to masses corresponding to water and ammonia losses from the b and y-ions. Given how little these theoretical spectra resemble experimental spectra, they do a remarkable job of unambiguously identifying their target. The development of more accurate models of peptide fragmentation and theoretical spectra is the overriding goal of the development of more powerful descriptive models.

**Current limitations of search algorithms**

Other than certain statistical models, the current state of the art in peptide identification entirely ignores the intensity information contained within peptide spectra. The intensity information is more difficult to predict than the expected masses of ions. However, discoveries about the gas-phase chemistry of peptides and their resulting fragmentations have suggested a possible framework for the prediction of fragmentation patterns. Given the need for improved discrimination in peptide identification, leveraging this information will be advantageous.

**Section 3: Peptide fragmentation under the mobile proton model**

**The mobile proton model**

The mobile proton model is the synthesis of a large number of studies into the gas-phase chemistry of peptides. Several groups have shown that most canonical fragmentation under low-energy CID proceeds by charge-directed mechanisms. Peptides contain a variety of basic sites that can accept protons during ionization. The most important of these are the N-terminus and the amide nitrogens and oxygens along the backbone, as well as the side chains of basic residues. For the purposes of the mobile proton model, these sites fall into two distinct
categories. The first contains those sites that have gas-phase basicity high enough to effectively sequester a proton from the sites on the backbone, even when the peptide is energetically excited by collision. These sites are generally limited to arginine, lysine, and histidine side chains. The second variety of protonation sites has basicities similar to those of amide nitrogens. These sites will share any excess protons, when the peptide is excited by collision. As such, the energetically less-favored amide nitrogen may become protonated and activate the adjacent peptide bond for charge-directed cleavage (Figure 5). For all practical purposes, peptides undergoing positive ion MS have charge states of +1, +2, or +3 and a corresponding number of protons available to the protonation sites. If the number of basic residues in the peptide is greater or equal to the charge-state of the peptide, there are said to be no mobile protons. As a result, no reactions that require an adjacent proton can occur, except at sites adjacent to the protonated residues (Figure 6). Charge-remote pathways for backbone cleavage do exist, most notably C-terminal to acid residues, but they are sequence specific and generally less-favored than charge-directed mechanisms.
Figure 5. The mechanism of mobile proton-directed sequence cleavage into b or y-ions. A mobile proton activates the amide nitrogen of a peptide bond, an oxazolonium intermediate is formed, and is then resolved into two fragments. If the charge ends up on the N-terminal fragment a b-ion is formed, if it ends on the C-terminal fragment a y-ion is formed. (From Paizs & Suhai, 2005)

Figure 6. The mechanism of charge-directed cleavage adjacent to basic residues. The charge from a basic residue is transferred to an adjacent amide and then the basic side chain cyclizes to the peptide carbonyl in a nucleophilic substitution and cleaves the C-terminal fragment. (From Paizs & Suhai, 2005)
Pathways in competition model

The pathways in competition model has been proposed by Paizs as a method of generating a comprehensive model of peptide fragmentation from the mobile proton model.\textsuperscript{22} His argument is that if all peptide fragmentation pathways (PFPs) are understood, it should become possible to accurately predict the spectra that peptides will produce. Numerous major fragmentation pathways are known already; Paizs describes twelve classes of pathways (Figure 7). In order to create a synthesized model, however, complete understanding of the energetics and the kinetics of the fragmentation pathways must be fully elucidated. The essence of this model is that those sites that can more effectively attract protons and those fragmentations with lower energies will compete better than their more acidic and stable counterparts when mobile protons are present, and those sites adjacent to residues capable of charge-remote cleavage will fragment preferentially when all protons are sequestered. With an understanding of the rates and energies of all the reactions, it presumably is possible to predict quantitatively the final distribution of products under any given conditions. To this end there have recently been great efforts to elucidate structures and energies of gas-phase peptides both experimentally and computationally.\textsuperscript{22-27}
Figure 7. A hierarchical scheme of peptide fragmentation pathways. The pathways separate on charge-directed and charge-remote then upon sequence and nonsequence fragmentations. Note the relative paucity of specificity in the charge-remote and nonsequence segments.

Section 4: An algorithm for the kinetic prediction of peptide fragmentation

Generation of theoretical spectra using the pathways in competition model

Zhongqi Zhang has created the first systematized implementation of the pathways in competition model into the software package Mass Analyzer. This program attempts to mechanistically predict spectra, including intensities, from peptide sequences. The process is rooted in basic physical principals and is conceptually quite easy to understand. First, the program estimates the ‘effective temperature’ of the peptide after collision. Each peptide terminus, amide bond and side chain is treated as a distinct protonation site with a gas-phase basicity that is estimated to be a function of only the local sequence (Figure 8). Because protons are assumed to be in thermodynamic equilibrium, given the temperature and basicities, the Boltzmann distribution reveals the average proton density at each site. Each possible protonation state can then fragment in a kinetically controlled process (Figure 9). The rate
constants of all fragmentation pathways are simultaneously calculated using the Arrhenius
equation and empirically derived activation energies. Once these constants are calculated, the
expected products after Δt are found. The effective temperature is then reduced and the
proton distributions and rate constants are calculated for the product ions and another round
of fragmentation is simulated. This process repeats until either the allotted fragmentation time
has passed or the effective temperature becomes too low for further fragmentation. At this
point, the intensities of the product ions are calculated and a theoretical spectrum is produced.
This spectrum may then be compared with experimental according to the equation:

$$ Sim = \frac{\sum \sqrt{I_{m}^{\text{theory}} I_{m}^{\text{exp}}}}{\sqrt{\left(\sum I_{m}^{\text{theory}}\right)\left(\sum I_{m}^{\text{exp}}\right)}}. $$

Figure 8. The protonation sites assumed in Mass Analyzer. Sites on the peptide that are protonatable are circled in
red. Each basic site is assumed to be strictly independent of all others, ignoring the possibility of charge sharing.
The gas-phase basicity (GB) is approximated as a function of the side chain alone for amino acid side chains. The
basicity of the backbone amide is assumed to be a function of the identities of the two neighboring amino acids.
Figure 9. The schema of fragmentation in Mass Analyzer. Protons are assumed to be in equilibrium across all basic sites. Protonation at each basic site introduces the possibility of charge-directed fragmentation reactions, under kinetic control. These fragmentation devices lead to the diversity of fragments that are observed in spectra.

The theory behind Zhang’s algorithm is rooted in the most basic physical chemistry. However, while it has aspirations of being purely descriptive, in execution, Mass Analyzer is highly statistical. Zhang used nearly 9000 previously-annotated CID spectra from Amgen’s proprietary database to train numerous variables which were later pruned to 22832 numerical factors which appeared to be critical to the model. These primarily include the gas-phase basicities of side chains and backbone of each amino acid and the contributions of the each residue to the energy of a peptide bond when it is N-terminal, C-terminal, or two residues N-terminal of the bond. Zhang briefly mentions that due to oversimplification of his model some of the physical factors must be viewed with skepticism. I would note that this model gives a single value for the backbone basicities at each bond, whereas, in reality there are protonation
sites at both the amide nitrogen and the amide oxygen, the former of which is by far the weaker of the two and the site which activates the bond for cleavage. The empirically derived basicity of these sites must, therefore, be some function of the ratio of the basicities of the oxygen and the nitrogen and cannot be theoretically correct to use in both the context of activating the bond and in the Boltzmann distribution. Looking at the changes in Zhang’s optimized values between his original and later models shows that the backbone basicities are the factors that are, indeed, the values most changed due to alterations to the model. Furthermore, all basic sites are assumed to be independent, negating any chance of modelling the ways in which multiple sites solvate protons. The more effectively these deviations from reality can be eliminated, the more realistic the kinetic basis of the model becomes.

**Section 5: The importance and fragmentation of phosphopeptides**

Reversible protein phosphorylation is involved in the regulation of virtually all aspects of cellular function. At least a third of the human proteome is thought to be phosphorylated. Phosphorylation networks, comprised of complex pathways of kinases, phosphatases and interacting regulatory proteins enable the cell to adapt to diverse environmental stimuli and transduce extracellular signals to the nucleus. Mass spectrometry has become the primary technology for large-scale phosphoproteomics, allowing thousands of phosphorylation sites to be monitored simultaneously and providing a global snapshot of complex signaling network responses to diverse biological stimuli. Unfortunately, the unusual chemistry of the phosphoester bond ensures that only a small percentage of these phosphopeptides can be identified by the predominant ion trap dissociation method, collisionally activated dissociation
Furthermore, the exact amino acid position of the phosphate can be localized in only 44% of identified peptides\textsuperscript{38}. Phosphoester bonds in phosphoserine and phosphothreonine are highly labile relative to other bonds in peptides. As a result, phosphopeptide CAD spectra are frequently dominated by a single peak corresponding to the neutral loss of phosphoric acid, H\textsubscript{3}PO\textsubscript{4}. Neutral losses from fragment ions further complicate the spectra by adding additional peaks to the spectrum, which confound commonly used search engines for phosphopeptide identification.

MS\textsuperscript{3} sequencing of the dominant H\textsubscript{3}PO\textsubscript{4} parent neutral loss ion yields a higher proportion of sequence-specific fragments. Several groups have found, however, that performing MS\textsuperscript{3} may lower identification rates\textsuperscript{39}, likely due to the reduction in sequencing depth. Consequently, most identifications come from MS\textsuperscript{2} scans or multistage activation scans that combine information from both levels of fragmentation.\textsuperscript{40} Electron transfer dissociation (ETD) is a complementary fragmentation method that tends to preserve the phosphoester bond and results in greater fragmentation across peptide backbone sites, but works well only for a subset of cationic phosphopeptides with high charge density\textsuperscript{41}. Incidentally, CAD performs well for localizing these same types of peptides, limiting the complementarity of ETD.

Loss of 98 Da in an MS\textsuperscript{2} spectrum is a positive indicator of the presence of phosphorylation, barring a few caveats\textsuperscript{42}. In practice however, the degree of neutral loss is highly variable and in some cases not present at all, depending on the peptide sequence\textsuperscript{43}. The inability to predict neutral loss levels in CAD MS\textsuperscript{2} from sequence alone reduces the effectiveness of phosphopeptide identification by common search engines by at least 40%\textsuperscript{44}. 
Accurate models to predict the fragmentation of phosphopeptides would enable more sophisticated and discriminative methods for phosphopeptide identification. However, the gas phase chemistry of phosphopeptides is poorly understood. Peptide fragmentation is generally explained according the mobile proton model \(^2\), which assumes that fragmentation occurs either by charge directed mechanisms that require a proton at the site of cleavage, or charge remote mechanisms that do not require protonation but may involve participation of neighboring side chains. Thus, the pattern of fragmentation is strongly influenced by whether the peptide contains “mobile” protons that are not sequestered by basic residues and free to migrate around the peptide.

Recent studies have noted inverse correlation between neutral loss of \(\text{H}_3\text{PO}_4\) and charge mobility in phosphopeptide MS\(^2\), \(^4\), which suggests that neutral loss proceeds by a charge remote mechanism. Tholey, et al. proposed a charge remote beta-elimination of phosphate from phosphoserine or phosphothreonine to form dehydrobutyric acid or dehydroalanine, respectively (Figure 10a)\(^4\). More recent work has shown strong evidence that neutral loss proceeds by a charge directed mechanism involving nucleophilic attack by a backbone carbonyl oxygen to form an oxazoline ring (Figure 10b)\(^4\), which would require protonation of the phosphate. This would predict higher neutral loss in mobile proton phosphopeptides, contrary to previously observed trends. Formation of a stable hydrogen bonded structure between the phosphate and a protonated basic residue, usually arginine, in place of the mobile proton, has been proposed for conferring electropositive character to the phosphate when mobile protons are not present.
Figure 10. Proposed mechanisms for the neutral loss of phosphoric acid. a) Charge remote beta-elimination of phosphoric acid to form dehydroalanine. b) Charge directed nucleophilic substitution to form an oxazoline.
Chapter 2: Implementing knowledge of chemistry to predict the fragmentation pattern of peptides

We suspected that knowledge of fragment ion intensities could remarkably improve spectral identification. Mass Analyzer proved a useful tool for predicting fragmentation spectra; however, there were problems, notably that the optimization algorithm that it uses, Brent’s Method,\textsuperscript{48} which is fairly slow and inefficient at minimizing the objective function. Furthermore, Zhonqi Zhang works at Amgen, and thus, the code for Mass Analyzer is proprietary. In order to make the improvements that we wanted to Mass Analyzer, we had to produce our own algorithm for kinetic fragmentation, which we call S3.

Shaojun Sun spearheaded the programming of the S3 model, while I provided insight into the gas phase chemistry underlying the model. His most significant improvement to the Mass Analyzer model was the implementation of Levenberg-Marquardt minimization.\textsuperscript{49} This allowed much more efficient fitting of parameters and more room for expanding the model. With this new algorithmic power, I studied predicted spectra and proposed and parameterized novel mechanisms to improve the model. The most interesting of these mechanisms was the addition of a special fragmentation pathway when proline is at the second position.

Section 1: Uncovering novel fragmentation mechanisms

It has long been known that peptides undergo dramatically reduced backbone cleavage C-terminal to proline residues.\textsuperscript{50,51} Furthermore, it is known that peptides preferentially cleave at the second peptide bond.\textsuperscript{52} The kinetic model accounts for the first of these effects well, and Zhang took care of the second by including the possibility of neutral loss of N-terminal
residues. Upon fitting our version of the model, we found that these two mechanisms interacted in unexpected ways and that the inhibition of cleavage C-terminal to proline was not present when proline is at the second position. Consequently, we significantly under predicted the intensity of backbone cleavages at the second peptide bond when proline is at the second position (Figure 11).

![Figure 11](image)

**Figure 11.** Error in prediction of N2 cleavage by amino acid at the second position. The log error in intensity prediction of fragments requiring cleavage at the second peptide bond as a function of the amino acid at the second positron is presented. Parameterizations with initial guesses are presented in black and those after fitting are in blue. There is clear under prediction when Gly or Pro are in the second position. It is must severe for proline which shows and average of 300-fold under prediction of the intensity of the ions.

The under prediction of b2 ions when proline is at the second position suggested the existence of a mechanism of cleavage that is not accounted for in the model. There had previously been speculation that b2 ions arose from nucleophilic attack by the N-terminus, leading to the formation of diketopiperazine structures (Figure 12). In formulating this mechanism, I assumed that this reaction was charge directed in a manner analogous to oxazoline formation. Using this assumption, we could fit partial activation energies for each
amino acid. Leading to an overall activation energy for the reaction:

\[ E_{a}^{N2} = E_{a}^{N2L} (R_2) + \Delta E_{a}^{N2R} (R_3). \]

All that remained to calculate the rate of reaction was fitting a pre-exponential factor. Unfortunately, this moved the average of our prediction to the correct value; however, the variance was still extremely high.

![Diketopiperazine](image)

Figure 12. Mechanism of charge directed diketopiperazine formation. Charge directed formation of diketopiperazines is analogous to oxazoline formation, except instead to the amide oxygen engaging in nucleophilic attack on the next charged amide, the N-terminus attacks to create a diketopiperazine.

In analyzing the problems with this model, Shaojun Sun and I found that peptides with a Lys at the C-terminus were universally underpredicted, while peptides with an Arg at the C-terminus were universally overpredicted. I deduced that this could be explained by a mechanism that was charge remote and did not require protonation at the site of reaction (Figure 13). To the best of my knowledge, this is the first reaction to be suggested which requires neither protonation nor acid catalysis to cleave the backbone of cationic peptides in the gas phase. However, this is justifiable, because the N-terminus is a primary amine, which is 30-40 kJ/mol more basic than the amide oxygens that perform the nucleophilic attack in more canonical backbone cleavages. Consequently, the N-terminus could carry out the reaction
despite the lack of protonation on the peptide bond. Zwitterions are generally unfavored in gas phase reactions; however, the resulting zwitterion would quickly resolve through protonation of the anionic amine.

Figure 13. Charge remote mechanism for diketopiperazine formation. Proceeds similarly to the charge directed mechanism, except that it does not require protonation at the second peptide bond for reaction to occur. This would be the first observed backbone cleavage in peptides that requires neither protonation nor acid catalysis. This is justifiable because the primary amine of the N-terminus is approximately 40 kJ/mol more basic than the amide oxygens that perform most charge directed reactions.

To investigate the likelihood of a charge remote mechanism of backbone cleavage, I instructed Shaojun Sun to assess the quality of the kinetic model, if a second, charge remote, fragmentation pathway were included. I rationalized that the activation energies of all reactions would be offset by a near constant value from the charge directed mechanism. This could be closely approximated with modifications to only the pre-exponential term of the Arrhenius equation. Thus, the relative importance of the two mechanisms could be assessed by fitting only the pre-exponential factors of both reactions. Models in which the charge remote mechanism dominates yielded much better fits to the observed data (Figure 14). Further testing is required to definitively demonstrate the importance of the charge remote mechanism.
Figure 14. Comparison of charge remote and charge directed mechanisms of N2 cleavage. Shown is the goodness of fit in terms of deviation from similarity as a function of the log-magnitude of the preexponentials of charge remote and charge directed N2 cleavage. Fits in which the charge remote process dominates the cleavage show nearly three-fold greater increased accuracy compared to those dominated by the charge directed process.
By the implementation of this new mechanism within the S3 model of peptide fragmentation, we removed the underprediction of fragmentation when proline is at the second position (Figure 15). Theoretical spectra with proline at the second position tend to be, obviously, more similar to the observed, largely due to better prediction of the intensity of the $\gamma_{N-2}$ ion and the $b_2$ ion (Figure 16). The addition of the N2 mechanism reduces the overall deviation in similarity scores by 2.88%. The addition of the N2 mechanism is the most interesting story of a mechanism uncovered in the pursuit of improving S3, but many others were found including: modified N-terminal basicities for the secondary amine when Pro is N-terminal, novel water loss from Glu, and side chain loss from N-terminal Thr.

![Figure 15](image-url)  
**Figure 15.** Correction of predicted intensity after addition of N2 mechanism. Plots of the deviation in predictions by amino acid at the second position are shown before fitting the new mechanism (black) and after (red). Proline is emphasized and shows near perfect improvement after fitting the N2 mechanism.
Figure 16. Examples of spectral prediction with and without the N2 mechanism. The observed and theoretical spectra of NPDSQYGELIEK are shown. The simulation with hate N2 mechanism added differs primarily in that it predicts the same large y10 ion as is seen in the experimental spectrum.

Section 2: Using fragmentation intensities to improve identification

While Shaojun Sun and I were improving the quality of spectral prediction in the S3 model, Chia-yu Yen was spearheading efforts to use those predictions to improve the identification of peptides.\textsuperscript{53} Zhongqi Zhang, who originally developed Mass Analyzer, has told me that his goal was solely to predict spectra as accurately as possible. He never intended to use the spectra for identification. Consequently, the similarity (SIM) score used in fitting turns out to be very poor for discriminating between correct and incorrect peptide matches. The
problem is that the similarity score is very sensitive to a small number of intense matched ions. Given the fact that peptide identification requires picking the correct match out of thousands, or even millions, of incorrect matches, the chance of spurious, high-scoring matches is high. Chia-yu Yen has found that this means that, at any high level of sensitivity, SIM-score based identification performs worse than the Mascot search engine, despite taking into account intensity information (Figure 17). SIM does use the additional information available to it to perform better at lower selectivities, but this is less useful experimentally.

Figure 17. Comparison of similarity score to Mascot search engine. Shown is a ROC plot for peptide identifications based on the UPS1 peptide standard. Similarity score (blue) performs worse than commercially-available Mascot at high selectivity. The even more permissive dot product score (red) performs extremely poorly throughout the range.
To improve discrimination, Chia-yu Yen developed more robust similarity metrics in intensity-based peptide identification.\textsuperscript{53} These techniques, generally, work by reducing sensitivity to outliers and residual errors in spectral prediction. The conceptually simplest manner that he used to do this was to strip the absolute intensities from the observed and theoretical spectra, preserving only the order of intensities. The implementation of similarity score on the ranked spectra proves much more effective than using SIM-score on absolute intensities (Figure 18). These improved scores demonstrate the effectiveness of using intensity information in peptide identification; however, the fact that the intensity information has to be softened prior to its usage implies the need for better spectral prediction.

Figure 18. Comparison of ranked similarity score to Mascot search engine. Shown is a ROC plot for peptide identifications based on the UPS1 peptide standard. Ranked similarity score (blue) performs only slightly worse than commercially-available Mascot at the very highest selectivity and significantly better throughout the rest of the range. The equivalent ranked dot product score (red) performs extremely similarly to ranked SIM score.
In studying the fragmentations predicted by S3, I found that many of the ions that differed most markedly from the observed were those that resulted from neutral loss of amino acid side chains. These ions are often intense and occur at similar mass shifts in all spectra. Thus, these ions are very important in discrimination during peptide identification. Spuriously high matches are often the result of these ions. I suggested that Chia-yu Yen should ignore neutral loss ions near the parent mass to help mitigate these problems. The result was significantly increased accuracy in ranked SIM score identification. For a given sensitivity, rSIM is twice as selective when neutral loss ions are ignored (Figure 19).

Figure 19. Performance of ranked similarity score when ignoring neutral loss ions. Shown is a ROC plot for peptide identifications based on the UPS1 peptide standard. Ranked similarity score is twice as selective when neutral loss ions are ignored (blue) than when they are considered (red).
Neutral loss is very sparsely represented by mechanisms in S3, because very little is understood about the underlying mechanisms in the fragmentation of side chains. The problems that these extraneous ions introduce in peptide identification make them an important target for future research. For the remainder of this work I will investigate the nature of neutral loss reactions, using pSer and pThr as my model residues. Through better understanding of the nature of neutral loss reactions, it may be possible to improve the fragmentation models that we have developed and ultimately increase the sensitivity of peptide identification.
Chapter 3: Using phosphopeptide libraries to investigate the neutral loss of phosphate

The neutral loss of phosphoric acid from pSer and pThr is ideally suited as a test reaction for the study of neutral loss events both because of the biological importance of protein phosphorylation and because it is a well-defined, but problematic, event in peptide identification. The neutral loss of $\text{H}_3\text{PO}_4$ is a well-known fragmentation event that often dominates the spectra of phosphopeptides. The stereotypical phosphopeptide fragmentation introduces a very large neutral loss peak 98 Da less massive than the parent (the M-98 peak); however, there is also neutral loss from sequence ions, leading to more complexity throughout the spectrum. Despite the additional complexity, the loss is of a known and distinct value and arises from a single side chain with known localization. Consequently, the neutral loss of $\text{H}_3\text{PO}_4$ is an effective system to probe the variability of neutral loss events.

Section 1: The development of combinatorial phosphopeptide libraries

Initial tests with simple combinatorial libraries

Given that there are difficulties identifying phosphopeptides from their MS$^2$ spectra, I initially attempted to create synthetic phosphopeptide libraries that could be identified \textit{a priori}. It is affordable and convenient to produce libraries in a single synthesis by introducing degenerate sites in the synthesis. These sites are introduced by adding a mixture of protected amino acids in one or more rounds of solid phase synthesis. As more degenerate sites are added, the diversity of peptides in the library increases exponentially (Figure 20). By designing
libraries that contain no isobaric peptides, I was able to identify the phosphopeptides from mass in the MS scan alone.

<table>
<thead>
<tr>
<th>Library</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAsEFTVVEK</td>
<td>YlSEFTVVEK</td>
</tr>
<tr>
<td>YAsSFTVVEK</td>
<td>YlSFFTVVEK</td>
</tr>
<tr>
<td>YAsNFTVVEK</td>
<td>YlNFNTVVEK</td>
</tr>
<tr>
<td>YAsNPTVVEK</td>
<td>YlNPNTVVEK</td>
</tr>
<tr>
<td>YAsEFYVEK</td>
<td>YlEFYVEK</td>
</tr>
<tr>
<td>YAsFYTVVEK</td>
<td>YlFYTVVEK</td>
</tr>
<tr>
<td>60 unique sequences</td>
<td></td>
</tr>
</tbody>
</table>

Figure 20. Combinatorial synthesis yields diverse peptide libraries. An example of how the combinatorial library phos1 can yield (4x3x5)=60 distinct sequences from a single sequence. The blue trace shows a path yielding the sequence YAsEFTVVEK; the green trace yields YDtnFTVVEK. The 60 resulting sequences are identical except at the sites of combination.

I ordered the initial libraries from Genscript (Piscataway, NJ). Phosphopeptide libraries were synthesized by solid phase synthesis on Wang resin with N-terminal f-Moc protection and phosphoresidues protected by t-Boc. Degenerate sites in library peptides were generated by adding a mixture of amino acids in certain coupling. The initial order contained four highly related libraries, differing only by the serial addition of alanine to the N-terminus of the sequence (Figure 21). The libraries contain either 60 expected unique sequences with one phosphorylation site on each. Libraries were received as lyophilized powders. Peptides were
solubilized with 5% formic acid, 95% water solution, agitated for approximately 2 minutes, then diluted to 0.1% formic acid prior to LC/MS/MS analysis.

Figure 21. The sequences of the first four combinatorial libraries used in this paper. Peptides are written N-terminal to C-terminal. Multiple amino acids in a column indicates that a mixture of the indicated amino acids was introduced at that site. Phosphorylated residues are indicated in red text and lower case. Each combinatorial library is expected to contain 60 distinct phosphopeptides.

LC/MS analysis over 1-hour reverse phase gradients proved highly successful at identifying the expected peptides in these relatively simple mixtures. All expected peptides were detectable in at least one charge state (Figure 22), with most observable as +1 and +2 ions and some even identifiable in the +3 charge state. However, there were certain warning signs that should have, perhaps, steered me away from using libraries in any larger scales. In some cases, it was impossible to achieve chromatographic separation of the peptide peaks, leading to overlapping isotope distributions and making MS$^2$ spectra of pure peptide impossible to obtain. Additionally, there were a multitude of peaks that did not correspond to any expected peptide
match. I further increased the variety of spectra obtainable by dephosphorylating the peptides using calf intestinal alkaline phosphatase (CIAP). This allowed access to the spectra of unphosphorylated cognates of the library spectra. The most prevalent of the contaminant peaks did not respond to the phosphatase (Figure 23).

Figure 22. Contour plot showing identifications from the library phos1. This contour plot shows the separation of phos1 with reverse phase peptide retention time on the x-axis and the m/z of peptides on the y-axis. This area is expected to contain the doubly-charged peptides present in the library. A isotope distribution corresponding to each expected sequence is observed; peaks are annotated with the variable amino acids. This demonstrates the ability of the libraries to provide information about the peptides; however, two important caveats are demonstrated. A set of peptides that are not fully resolved in mass and time are indicated in blue. Contaminant peaks that do not correspond to any expected peptide sequence are indicated in red.
Figure 23. The effects of phosphatase treatment on phosphopeptide libraries. Treating phosl1 with 1 unit/microgram peptide calf intestinal alkaline phosphatase for 60 minutes effectively dephosphorylates the peptide portion of the library. This is indicated by the separation of the phosphatase treated sample (green) from the untreated (red overlay) sample in the areas containing peptides. However, there are intense peaks that do not respond to phosphatase treatment and smear significantly more than the peptides in reverse phase chromatography.

Phosphopeptide libraries on this scale, therefore, proved successful at providing a variety of fragmentation spectra. However, these libraries only provided 60 sequences each, with relatively little diversity among those sequences. While, this is a more varied sample than is included in many anecdotal studies on fragmentation, it is did not prove enough for rigorous statistical evaluation of the factor affecting neutral loss. Consequently, I decided to produce more diverse libraries.
Scaling up to more diverse phosphopeptide libraries

In scaling up to larger phosphopeptide libraries, I decided to maintain the rule that there should be no isobaric peptides within a library. This proved to be less trivial endeavor than it would initially appear. For a library with 23 possible amino acids (the twenty naturally-occurring amino acids, plus three phosphorylated amino acids), there are $2^{23} = 8,388,608$ possible degenerate combinations. Consequently, for as few as four degenerate sites there are almost $5 \times 10^{27}$ possible combinations. Upon realizing this and the resulting fact that the calculations would take approximately a trillion years, I gave up the attempt to determine whether libraries were isobaric by brute force calculations.

To simplify the calculation of potential combinatorial sequences, I made some assumptions about the desired nature of the libraries. Since I am primarily interested in the sequence around phospho-residues, I decided that one site should contain the residues pSer, pThr, pTyr, and Val. There should be four total degenerate sites; the others should contain one each of pSer, pThr, and pTyr. Given the need for isobaric sequences and the fact that the fixed residues ensure that each degenerate site includes one residue that also shows up in another site, it is impossible for any other residue to show up more than once. This leaves $4^{20}$ or approximately 3.5 billion combinations. After three days of number crunching, 32 distinct degenerate combinations proved to yield 720 isobaric phosphopeptides, the maximum obtainable variety.

I developed another script that made the most diverse possible overall library of peptide sequences by varying the constant sites around the acceptable degenerate combinations. I
ordered an additional six peptide libraries (Figure 24). These libraries were more diverse than the simpler libraries; however, they were not as well behaved. The contaminant peaks that had been relatively minor in the simpler libraries became dominant when the peptide signal was diluted over a much broader sequence space (Figure 25). The possibility of an unexpected contaminant at the mass of an expected peptide made identification by mass alone impossible. Furthermore, the majority of contaminants contain part of the expected peptide sequence. Consequently, the contaminants resemble expected peptides even when analyzed at the MS² level. This ensured that the identity of the peptide had to be verifiable by MS², even then only high scoring examples could be accepted. Despite these limitations, the more complex libraries yielded more identifications than the simpler ones and brought the number of identified spectra from the hundreds into the thousands.
Figure 24. The sequences of the complex combinatorial libraries used in this paper. Peptides are written N-terminal to C-terminal. Multiple amino acids in a column indicates that a mixture of the indicated amino acids was introduced at that site. Phosphorylated residues are indicated in red text and lower case. Each library is expected to contain 720 distinct phosphopeptides.
Figure 25. Contour plot of a combinatorial peptide library with 720 expected phosphopeptides. The contour plot shows significantly more complexity than was observed in libraries that only contained 60 expected peptides. Note especially the prevalence of chromatographically unresolved contaminant peaks that extend throughout the mass range.

Section 2: Analyzing the fragmentation patterns in phosphopeptide libraries

Creating a data structure for fragmentation analysis

Having data on a scale large enough to make manual curation impractical revealed a paucity of tools for the analysis of fragmentation. Most proteomics-based data structures focus on the identification of peptides and make the underlying data about the fragmentation difficult to access; many even discard the underlying data entirely. Consequently, the development of tools for the analysis of fragmentation was essential to further experimentation.
To facilitate analysis I developed an ion- and ion-annotation-centric relational database structure that would easily allow access to fragmentation information within or among spectra (Figure 26). Logically, the database consists of two tables representing experimental inputs and a circular relational arrangement of five tables representing spectra and their annotations. The input tables include `experiment`, which contains information about the sample that was run, and `search`, which contains the parameters used in data processing and annotation. The raw data is represented by the `spectrum` table which holds information about the MS scans. This joins by a one-to-many relationship with the `fragment` table which contains information about each ion observed in the spectrum. These tables are joined to a corresponding pair of related tables, `spectrumsid` (in a one-to-many relationship) and `fragids` (in a many-to-many relationship), which contain the annotations of the spectrum and fragments, respectively.
Figure 26. Schema of the relational database used for the storage of spectral information and annotations. Tables with white backgrounds are experimental inputs, tables with red backgrounds hold raw mass spectroscopic information, tables with green backgrounds hold annotations of spectra and fragments. The experiment table holds metadata about the sample used. Spectrum and fragment contain the information from MS1 and MS2 scans, respectively. Search contains the parameters of the Mascot search. Spectrumid holds the information about a particular identification, while fragid holds the annotations for individual fragments in the spectrum. Fragment_fragid exists to allow a many-to-many join of fragments and their annotations along with metadata for that annotation.
Peaks were annotated using a priority system similar to that which Katheryn Resing and I had developed in Sun, et al. Peaks were initially identified with a stringent tolerance of ±0.125 Thompson (Th) plus 200 parts per million (ppm). Ions matching within this window were then assigned to the peak with the following priority: neutral loss from the parent, singly charged b- or y-ions, singly charged a-ions provided the corresponding b-ion was present, singly charged b- or y-ions with a single neutral loss provided the intact b- or y-ion was present, singly charged a-ions with a single neutral loss provided the corresponding a-ion was present, singly-charged b- or y-ions with multiple neutral losses provided the intermediate neutral loss ions are present. This list was then repeated for doubly- and triply-charged ions. If no matches were found, the mass tolerance was increased to 0.250 Th plus 400 ppm and the process was repeated. If the identification was ambiguous according to the priority system, multiple identifications were assigned. The total neutral loss in a spectrum is the total intensity annotated as having lost phosphoric acid divided by the total identified intensity. Peaks that have ambiguous neutral loss state had their intensities divided between the identifications.

By the use of this data structure, it is possible to probe effects at the ion-by-ion, spectrum-by-spectrum, or even dataset-by-dataset level, with theoretically equal ease. Unfortunately the size of the database ensured that certain joins of all ions were inconveniently, if not prohibitively, computationally expensive. The final database structure is consequently intentionally denormalized with some information about the fragment identifications contained within the spectrumid table. This denormalization allowed much more convenient pre-filtering of spectrum-level analyses and greatly sped up analysis.
Section 3: Variability in neutral loss patterns based on instrumentation used

We primarily use two instruments an LTQ-Orbitrap (Thermo Electron, Bremen, Germany)\textsuperscript{55} and a Qtrap 4000 (Applied Biosystems, Foster City, CA). Both instruments fragment peptides through collisional activation; however, the mechanisms differ in subtle but important manners. The LTQ isolates the ion of interest in a 3D-quadrupole ion trap, then excites it through RF at the harmonic frequency of the isolated mass in the trap. Consequently, only ions at the parent mass are heated. Conversely, the Qtrap fragments ions in a transmission quadrupole by accelerating all ions across a fixed electrical potential. Consequently, daughter ions continue to be heated after fragmentation. Additionally, the shorter isolation time of the transmission quadrupole requires that higher energies are used in fragmentation.

Fragmentations in both instruments were performed under standard proteomics conditions. Both instruments used 1 mTorr of nitrogen as a collision gas. In the LTQ, ions were fragmented for a 30ms activation at 35\% of normalized energy. In the Qtrap, ions were quadrupole selected and fragmented with default rolling collision energy.

As an initial sanity check, I assessed whether the amount of neutral loss observed in one instrument correlated with the amount of neutral loss in the other. Surprisingly, the observed correlation on 343 pairs of matched spectra was remarkably low (Figure 27). While spectra that showed the very lowest amounts of neutral loss tended to do so in both instruments, those that show some neutral loss correlate very poorly. This implies that the rate of neutral loss is highly sensitive to the mechanism of fragmentation. To discover the reasons for this effect, I looked more in depth at what neutral loss ions appear in each instrument.
Figure 27. Correlation if neutral loss in the Qtrap and LTQ mass spectrometers. A scatterplot showing the observed fraction of neutral loss a given peptide sequence exhibits in the LTQ (x-axis) and Qtrap (y-axis) mass spectrometers. The level of correlation is remarkably small indicating that the differences between the fragmentation energies significantly alter the observed neutral loss.

The canonical neutral loss peak in phosphopeptides is the M-98 peak, representing the loss of H$_3$PO$_4$ from the parent. This ion alone is dramatically more prevalent in the LTQ than it is the Qtrap (Figure 28). On average peptides in the LTQ show 10-fold more signal at M-98 than those in the Qtrap. The variability in neutral loss is also much higher in the LTQ. This indicates that the problem of the neutral loss peak, which is not informative to peptide sequence dominating spectra, is significantly less prevalent in Qtrap spectra; however, even in the LTQ only 4% of spectra will show a 2-fold or higher decrease in sequence ion signal-to-noise.
Figure 28. Cumulative distribution of neutral loss from the parent in the LTQ and Qtrap. The cumulative distributions of the intensity of the M-98 peak from 6263 distinct doubly protonated pSer or pThr containing spectra in the LTQ (green) and the same plot for 3124 nonmatched spectra in the Qtrap. There is a large diversity of intensities in both instruments, though the neutral loss peak is far more prevalent in the LTQ. Spectra in the LTQ display an M-98 peak approximately ten-fold larger than those in the Qtrap. Despite this neutral loss 95% of LTQ spectra have less than half of the intensity in this peak; consequently it is rare that the M-98 peak will contribute more than a two-fold decrease in signal-to-noise on sequence ions.

Since the observed neutral loss of H$_3$PO$_4$ in the Qtrap is not accounted for by losses from the parent ion, it must be contained in neutral losses from sequence ions. These ions are especially detrimental to peptide identification, since, without the ability to accurately predict the balance between the ion series, identification algorithms have to effectively guess which ion series to consider. I assessed the ratio of intensity between intact sequence ions and their cognate that had undergone neutral loss (Figure 29). The sequence ions in the Qtrap are, on average, twice as intense as those in the LTQ. The lack of correlation in neutral loss in the two instruments can, therefore, be attributed to the fact that fundamentally different ions are
counted as neutral loss peaks in each instrument. Interestingly, while the Qtrap shows more neutral loss than LTQ, this increase makes the expected neutral loss state of the sequence ions more, not less, predictable, since the ions predictably undergo neutral loss.

![Graph](image.png)

Figure 29. Comparison of the amount of intact and neutral loss sequence ions in the Qtrap and LTQ. The cumulative distribution of ratios of intensities of b- and y-ions that display neutral loss to the amount that contains intact phosphate. The LTQ sample (green) consists of 22,479 matched pairs of ions, while the Qtrap consists of 13,288. In the LTQ intact and neutral loss sequence ions are approximately equally likely; while in the Qtrap ions displaying neutral loss are approximately twice as common.

The observations of reduced variability in neutral loss both from the parent and from sequence ions indicate that phosphopeptides should be easier to identify in beam-type CAD fragmentation. This may explain a recent movement in phosphoproteomics. More recent versions of the LTQ contain an additional collision cell called the HCD. This cell has slower scan speeds than the ion trap, but yields fragmentation patterns that more closely resemble those
from beam-type CAD. While the underlying mechanism has never been explained, several groups have found that fragmentation in the HCD dramatically increases the identification rate of phosphopeptides.$^{57-59}$

**Section 4: Assessing determinants of neutral loss**

Since most of our high-throughput results are acquired in the LTQ, I focus on that instrument for the remainder of this work. In an attempt to predict the amount of neutral loss from phosphopeptides, I first assigned each spectrum values for 56 attributes. These attributes included measures of spectral quality, the identity of the phosphorylated residue, and the position of the phosphate relative to the peptide termini. Additionally, the sequence of the peptide was encoded, both in a phosphate-centric manner, noting the identities of the 13 residues N- and C-terminal to the phosphate, as well as, in the overall amino acid composition of the peptide. Finally, the amount of neutral loss from the peptide was noted.

Using the WEKA data mining package,$^{60}$ I attempted to classify the amount of neutral loss, stratified by decile, that peptides undergo, using the phosl libraries as a training set and biological samples for testing (Figure 30). I was concerned that the high sequence homogeneity in the library spectra would make overfitting to sequence motifs a significant problem when testing against a biological sample. However, when only considering the adjacent amino acid residues, analogous to those important in backbone cleavage, the fitting was remarkably good with as many as 63% of spectra classified within one decile of the true value. There was cause for concern, though, as the training sets performed much better than the test sets. Furthermore, when more extensive sequence motifs were considered, the amount of over
fitting became severe, with the test performing only marginally better than guessing. This suggests that I am making progress in understanding the patterns of neutral loss; however, I may need better data to uncover more.

Figure 30. The quality of classification of neutral loss by naive Bayesian classifier. Using spectral dataset from 2219 synthetic phosphopeptides in the LTQ for training, the performance of a naive Bayesian classifier of neutral loss, stratified by decile in the training set, is assessed using a variety of attribute sets. The test set is derived from 541 naturally occurring phosphopeptides from biological sample. Both exact matches and those within one decile are shown in the respective traces. The adjacent attributes include the identity of the two amino acids to the left of the phosphate and on to the right (analogous to those found to be important in backbone cleavage). NumBasic represents the number of basic residues, position indicates the position of the phosphate relative to the N-terminus, phosphoresidue is the identity of the phosphorylated residue itself. All includes all the previous attributes, plus the identity of each residue within 13 amino acids of the phosphate. The fitting shows limited success in classifying neutral loss, especially when considering adjacent residues, the position, and the number of basic residues. There is, however, evidence of severe overfilling when all attributes are considered.

The significant over fitting when more distal sequence effects were considered could be attributable to the unimportance of distal residues or it could indicate that the sequence
diversity of peptide libraries is not sufficient to effectively represent these residues. The spectra of the sequences \[1\]pTPAVALTSMEV[2] where \[1\]=D,T, or R and \[2\]=A,K are telling (Figure 31). The variable site at the N-terminus is immediately adjacent to the site of phosphorylation and it does affect the amount of observed neutral loss of H$_3$PO$_4$. However, the variable site at the C-terminus, 12 residues away from the phosphate, has a much more severe impact on the rate of neutral loss. When this residue is Lys, neutral loss of H$_3$PO$_4$ is prevalent, but when it is Ala, neutral loss of H$_3$PO$_4$ essentially stops and instead neutral loss of water is observed. This implies that distal residues are able to dictate the rate of neutral loss. However, in order to investigate these phenomena, more diverse data is necessary. Given the theoretical and technical difficulties associated with the use of peptide libraries on the scale needed, the use of data derived from biological sources is suggested.
Figure 31. Distal amino acids can affect neutral loss more dramatically than proximal amino acids. The spectra of the sequences [1]pTPAVALTSMEV[2] where [1]=D, T, or R and [2]=A, K are shown. Altering the amino acid at the N-terminus, immediately adjacent to the phosphate, only slightly alters the amount of neutral loss. However, changing the residue at the C-terminus, 12 residues from the phosphate, from Lys to Ala universally inhibits the neutral loss of $\text{H}_3\text{PO}_4$, instead water is lost. This implies that distal residues can have dramatic effects on neutral loss.
Chapter 4: Understanding the variations in neutral loss from phosphopeptides

Previous studies of phosphopeptide fragmentation mechanisms are largely based on small sets of phosphopeptide MS², limiting their applicability for developing a general model of phosphopeptide neutral loss. In this study, I comprehensively examine neighboring residue effects that determine rates of neutral loss in phosphopeptide CAD MS², using 34,057 spectra from public spectral databases supplemented by spectra obtained in our laboratory from biological and synthetic sources. Using a robust non-parametric statistic based on changes in the quantile distribution of neutral loss levels, conditioned on local sequence features, I show that immediately adjacent residues and those up to seven residues distal to the phosphosite influence the total amount of observed H₃PO₄ neutral loss. Distal basic sites, most notably the N-terminus, show strong effects on neutral loss and suggest mechanisms contrary to the mobile proton model, in which immobile protons participate in charge directed mechanisms by forming secondary structures.
Section 1: Evaluation of the types and amounts of neutral loss

Characterization of types of neutral loss

Heteroatom-containing side chains frequently exhibit neutral losses upon CAD, providing competitive fragmentation pathways to $\text{H}_3\text{PO}_4$ loss in phosphopeptides. The resulting complexity in fragment ion patterns complicates de novo sequencing and identification by database search algorithms. Thus, improved understanding of the relative rates of major neutral loss events is essential for improving phosphopeptide MS$^2$ interpretation. $\text{H}_2\text{O}$ and $\text{NH}_3$ are the most commonly observed neutral losses during CAD of protonated peptides. I first examined the prevalence of $\text{H}_2\text{O}$ and $\text{NH}_3$ neutral loss events in combination with $\text{H}_3\text{PO}_4$ loss in a large set of 5749 validated phosphopeptide MS$^2$ collected on an LTQ-Orbitrap instrument (WM239A dataset), comparing the ion intensity attributable to each neutral loss event normalized to the total ion signal (Figure 32a). Only 32% of ion intensity represented peaks showing no neutral loss. As expected, neutral loss of $\text{H}_3\text{PO}_4$ was the most common neutral loss event, followed by water and ammonia neutral losses, which differ by one Da and are often indistinguishable in low resolution LTQ MS$^2$, and thus are considered together in this analysis. Furthermore, all combinations of two distinct neutral losses accounted for significant signal, while ions displaying three or more neutral losses were uncommon.
Figure 32. The distribution of neutral loss. Neutral loss characteristics of the fragmentation spectra of 5749 unique peptide ions containing a single phosphothreonine or phosphoserine derived from samples of the human cell line WM239A. (a-d) Distribution of identifiable ion intensity by type of neutral loss observed. (a) represents the total signal within the spectra, including b- and y-ions that do not contain the site of phosphorylation. (b) Parent ions include only those ions that do not have a cleavage at the backbone. (c) b-ions and (d) y-ions include only those ions of the given series that contain the site of phosphorylation. (e-h) The distribution of the fraction of identifiable signal annotated to result from the neutral loss of phosphoric acid, binned by increments of 10% of all signal. Separated by identity of the phosphorylated residue (e) or the charge mobility of the peptide ion (f). (g) Signal annotated as parent ion neutral loss. (h) Same as (f) except that only signal annotated to have at least one cleavage of the peptide backbone counted.
When neutral loss from parent ions and b- and y-series were treated separately, trends strongly deviated from the overall distribution of neutral loss (Figure 32b-d). Remaining intact peptide was not observed (Figure 32b), as would be expected due to resonant activation of the parent in ion trap CAD. Interestingly, neutral loss of H$_3$PO$_4$ was dramatically different for b- and y-ion series. Only 22% of y-ion intensity shows neutral loss, compared to 72% for b-ions, confirming trends suggested by a previous statistical learning study that examined over 3,000 spectra 61. This indicates either that ions that have undergone neutral loss form a product that is more susceptible to b-ion formation or that b-ions are more likely to undergo neutral loss. However, neither mechanism shown in Figure 10 suggests a reason for this bias. One potentially explanatory difference in the composition of b- and y-ions in tryptic peptides is absence of a C-terminal basic residue in b-ions, which would affect mobility of the remaining proton, or result in gas-phase interaction with the resulting H$_3$PO$_4$. However, the dramatic difference in b- and y-ion neutral loss persisted when ions were partitioned by the presence of basic residues.

Importantly, this effect is not a result of the bias to the presence of basic residues on the C-terminus as it persists in peptides with non-tryptic C-termini and those constrained to have N-terminal arginine or lysine (Figure 33). One implication of the greater observed stability of phosphate on y-ions in CAD spectra is that precise localization of phosphorylation sites relies more heavily on y-ion series, since observation of intact phospho residues is required for unambiguous localization.
Figure 33. Assessment of neutral loss types in non-tryptic peptides. The proportion of signal intensity representing each neutral loss state, as was shown in Figure 1 a-d, but in peptides that differ from typical tryptic peptides. The left column contains peptides with either R or K as the N-terminal amino acid, these spectra almost all contain an additional R or K at the C-terminal amino acid. The right column represents spectra with non-tryptic C-termini.
Variation in neutral loss between peptides

Models that estimate neutral loss based on global averages have been successful at increasing the number of phosphopeptides that can be identified\textsuperscript{61}. However, if neutral loss rates vary widely between peptide MS\textsuperscript{2} in a sequence dependent manner, such methods are likely to introduce biases against outliers to the main trend. To assess global variability of H\textsubscript{3}PO\textsubscript{4} loss in phosphopeptide CAD, I examined the distribution of loss over all spectra in the WM239A dataset (Figure 32e). For peptides containing either pSer or pThr, the observed neutral loss of H\textsubscript{3}PO\textsubscript{4} varies from undetectable to accounting for the entire signal. pSer shows slightly more neutral loss than pThr, as noted previously\textsuperscript{62}. Interestingly, this trend is opposite of the analogous neutral loss of water from unmodified serine and threonine\textsuperscript{62}, suggesting that the loss of water and loss of H\textsubscript{3}PO\textsubscript{4} have different rate limiting factors. That pSer, while less basic than pThr, shows higher neutral loss, suggests that proton availability is not rate limiting in the neutral loss of H\textsubscript{3}PO\textsubscript{4} as would be predicted for charge directed mechanisms such as the oxazoline mechanism (Figure 10b). While the difference between neutral losses from pSer and pThr is useful for inferring mechanism, the magnitude of the difference is very small compared to the overall variation in neutral loss. Thus, other factors must contribute to the variability of neutral loss.

Previous studies have noted an inverse correlation between proton mobility and the amount of H\textsubscript{3}PO\textsubscript{4} neutral loss of in small sets of phosphopeptide MS\textsuperscript{2}, with lower charge mobility correlating with higher neutral loss\textsuperscript{43}. To quantify this effect over a significantly larger data set, I examined the distribution of neutral loss in spectra stratified by proton mobility
(Figure 32f). The differences between mobility groups accounted for 24% of the observed variance in neutral loss, a much greater effect than was observed between pSer and pThr. However, the variance within each group is still large, suggesting other factors are important in loss of H₃PO₄.

Section 2: Effects of local sequence on neutral loss

Having exhausted the factors that were known to affect the neutral loss of H₃PO₄, I examined local sequence effects. Although the WM239A dataset was nearly twice as large as the largest previous study on phosphopeptide neutral loss⁶¹, I found that statistically meaningful estimates of sequence effects required much larger sets of phosphopeptide MS². The PhosphoPep project at the Institute of Systems Biology (ISB)⁶³ provides a public database currently containing more than 30,000 phosphopeptide spectra obtained from several experiments. The database currently contains data derived from yeast, C. elegans, fly and human samples. The exact experimental details vary; however, in general, tryptic phosphopeptides were enriched from cellular extract and then analyzed by LC/MS/MS on either an LTQ-Orbitrap or LTQ-FTICR. Spectra are denoised and replicates matched to the same peptide sequence and charge state are combined to create consensus spectra. Fragment ions are annotated using the SpectraST toolset⁶⁴. In this study all of the included spectra are assumed to be correct identifications. I obtained a database of 34,057 unique peptide ions containing a single pSer or pThr. These ions were all either doubly or triply protonated. Similar results are obtained when using only the most confident 50% of spectra (xCorr score greater
than 3.0, data not shown). However, the consequent reduction in the number of spectra reduces the statistical significance of the results.

Two important caveats for the use of consensus spectra became apparent in the analysis of the PhosphoPep data. Consensus spectra in this case were obtained through the averaging of multiple observed spectra of the same peptide. Firstly, the spectra varied widely in the amount of noise and the quality of annotation. When analyzing spectra that deviated greatly from observed trends in neutral loss, many proved to be the result of questionable annotation or especially noisy spectra. It is possible to filter out the poorest signal-to-noise spectra from the dataset; however, doing so significantly lowered the number of spectra available. Second, the consensus spectra in the PhosphoPep data showed 12% lower neutral loss than our unprocessed spectra collected in our lab, an effect that I speculate is the result of peak voting procedures that emphasize low-abundance ions and dampen high-abundance ions. This effect appears to be general to consensus spectra and is observed in other libraries that do not contain phosphopeptides (Figure 34). Creating data mining metrics that were robust to these artifacts is essential to the effective use of these rich repositories of data.
Figure 34. Spectral databases compress the observed range of intensities. Comparison of the observed intensity of all matched ions from the NIST 2011 library (y-axis) and locally obtained unprocessed spectra, ph3 (x-axis) is shown. For this analysis the intensity of the highest peak in the spectrum is defined as 100. This demonstrates that the ratio of intensities between the highest and lowest ions is significantly suppressed in the library spectra compared to spectra obtained locally.

**Visualizing neutral loss patterns**

To assess the effect of localized residues in the presence of intensity biases, I developed a metric that is robust to wide variability in the underlying data, poor spectral quality, misannotation, and distorted intensity distributions. I found that problems associated with variability and noise were solved by considering only large changes in neutral loss. This suggested using a non-parametric quantile-based statistic, in which the central portion of the
data was ignored. In this method, the overall frequency distribution of neutral loss intensity among all spectra is divided into three equal classes: the third of the spectra with the highest neutral loss is classified as high neutral loss, the middle third of the data is excluded, and the bottom third of the spectra is classified as low neutral loss. The metric is an odds ratio of high neutral loss to low neutral loss conditioned on putatively explanatory factors, expressed as

\[
\frac{p(NL=\text{high}|AA_i=X)/p(NL=\text{low}|AA_i=X)}{p(NL=\text{high})/p(NL=\text{low})},
\]

where \( p(NL = \text{high}|AA_i = X) \) is the probability that a spectrum is among the tercile of spectra with the highest neutral loss, given that it was chosen from the subset of spectra in which the amino acid at position \( i \) is \( X \); \( p(NL = \text{low}|AA_i = X) \) is the conditional probability for a spectrum being in the lowest tercile, and

\( p(NL = \text{high})/p(NL = \text{low}) \) represents the prior odds of high versus low neutral loss. Since the data were divided into equal terciles, the prior odds are by definition 1, simplifying the equation to

\[
\frac{p(NL=\text{high}|AA_i=X)}{p(NL=\text{low}|AA_i=X)}.
\]

Peptide spectra were stored in my ion-centric relational database. The sequence of identified peptides was encoded in a phosphosite-centric manner in which the phospho-residue was position 0. The distances of a phosphosite to each of the termini, the identity of the phosphorylated residue and the identity of each residue from one to 13 residues N-terminal to the phosphosite (positions -1 to -13) and one to thirteen residues C-terminal to the phosphosite (positions 1 to 13) were recorded. The termini were annotated one residue position distal to the terminal residues. Since peptides are of variable length, many peptides do not extend to all 27 residue positions in this encoding, corresponding to the largest phosphopeptide length. For instance, the peptide XXpSXX, has the N-terminus at the -3 position and amino acids at the -2, -
1, 1 and 2 positions, while all other positions are outside of the sequence. Any position that is not within the peptide sequence are annotated as missing (‘‐’), thus placing all peptide sequences regardless of length on the same scale. In addition to the sequence information, the charge of the peptide ion was recorded.

Since I am using phosphosite-centric positional indexing, there is no guarantee that any given position is actually within the extent of the peptide. For example, if the phosphate is at the second position, there is no amino acid at the -2 to -7 positions. Rather than allowing sparse data, I explicitly encode these positions as ‘‐’. Because the position of the phosphate within the peptide is a confounding factor to discovering sequence-specific effects, I corrected the prior odds at each position to exclude peptides that did not have an amino acid at the position. The resulting odds-ratio, which I use for the remainder of this work to measure the importance of an effect, is

\[ \frac{p(NL=high|AA_i=X)/p(NL=low|AA_i=X)}{p(NL=high|AA_i=-)/p(NL=low|AA_i=-)} . \]

A p-value for the significance of the observations can be obtained from a two-tailed binomial distribution. This p-value corresponds to the likelihood that a result at least as extreme as the observed would be obtained by chance, if the criterion had no effect on neutral loss. Since I am simultaneously testing 22 amino acids at each of 14 positions around the phosphate there are 308 simultaneously-tested hypotheses. The p-value was adjusted for multiple hypotheses using the Sidak method \(^66\).

To assess the importance of proton distribution on fragmentation, the spectra were further divided by proton mobility. Mobile proton spectra are defined as having a charge
greater than the sum of the number of arginines, lysines and histidines in the sequence. Immobile proton spectra were defined as having a charge less than or equal to the number of arginines in the sequence. All other spectra were classified as partially mobile proton. Once these subsets of spectra are generated, they are binned by tercile (likely with different cutoffs from before) and tested for odds-ratio, as above.

Residues immediately adjacent to the phosphosite show the strongest effect on neutral loss of $\text{H}_3\text{PO}_4$. When proximal to $\text{pSer}$ or $\text{pThr}$, glycine, basic amino acids and acidic amino acids increase the rate of neutral loss (Figure 35, left panel). Proline reduces neutral loss, but other aliphatic side chains increase neutral loss when immediately adjacent. Threonine, serine, carbamidomethyl-cysteine, glutamine, and asparagine all reduce neutral loss when N-terminal to the phosphosite. Twelve of the twenty amino acids show significant effects even when five or more residues from the phosphosite.
Figure 35. The effects of local sequence on neutral loss. The effects of peptide sequence adjacent to the site of phosphorylation on 34,057 unique spectra from the ISB PhosphoPep database. The horizontal axis represents the position of residues from 7 sites N-terminal to 7 sites C-terminal relative to the phosphorylated residue. The vertical axis denotes the presence of a particular amino acid at the position. ‘‐’ indicates that the peptide is not long enough to contain the site, thus ‘‐’ serves as a cumulative marker for the peptide terminus. A blue dot indicates that the peptides with the specified amino acid at the specified position are more likely to have high neutral loss than low neutral loss. Red dots indicate that low neutral loss is more likely. Low neutral loss is defined as having neutral loss less than at least two-thirds of the peptides in the dataset (bottom tercile). High neutral loss is defined as greater than at least two-thirds dataset (top tercile). The width of the iris represents the multiple-hypothesis-corrected probability that this ratio could occur by chance if the indicated residue had no effect on neutral loss.

**Simpson’s paradox: controlling for charge mobility, proximal basic residues**

**suppress H₃PO₄ neutral loss**

Because the presence of basic residues correlates strongly with charge mobility, the positive effect of nearby basic residues could simply reflect a correlation with previously shown effects of charge mobility. To assess this possibility, I repeated the flanking residue analysis with the data stratified by charge mobility (Figure 35, center and right panels, Figure 36).
Surprisingly, in contrast to the positive effect of proximal basic residues on neutral loss in the global analysis, mobile and partially mobile MS<sup>2</sup> show decreased neutral loss when basic residues are adjacent to the phospho-residue. Separating the effects by charge mobility reveals that the basic residue effect observed in the global analysis is likely a proxy for the correlation between the presence of basic residues and the charge mobility of the peptides. The lack of available data from immobile proton spectra reduces the number of results that attain significance after multiple testing corrections. The only significant results for immobile proton cases show that arginine at the -1 position strongly inhibits neutral loss and that proximity to the C-terminus enhances neutral loss. Partially mobile and mobile cases are observed at high enough frequency that many significant effects can be observed. When the data is separated by mobility, most amino acids show statistically similar effects in all charge states. However, the effects of the basic residues and the peptide termini differ dramatically with proton mobility.

![Figure 36](image_url)

Figure 36. The effects of flanking sequence on neutral loss from immobile proton spectra. The data is presented exactly as in for the other mobilities. The low number of observed immobile cases leads to very few significant results.
Whereas basic residues are positively correlated with neutral loss when examined over all spectra, when spectra are stratified by charge mobility, basic residues proximal to the phosphate are negatively correlated with neutral loss in all mobility groups. This example of Simpson’s paradox occurs because the presence of basic residues shows strong positive correlation with charge mobility. Simpson’s paradox occurs when a correlation is present in an overall population, but when that population is divided, each subgroup shows correlation in the opposite direction. I propose that basic residues have the direct effect of slowing neutral loss rates when proximal to the phosphoresidue. In non-mobile proton MS², the lack of available protons slows the aggregate of other fragmentation pathways to a greater degree, increasing apparent neutral loss due to reduced fragmentation pathway competition.

Once I control for the shifts in overall reaction rate caused by changes in proton mobility, I can more closely examine the direct variation in neutral loss caused by basic residues. The only observed trend with basic residues that was universal across all charge mobility classes is that basic residues at the -1 position strongly inhibit neutral loss. I hypothesize that this effect occurs because the protonated basic residue will form a very stable hydrogen-bonded structure with the phosphate, producing steric hindrance that inhibits formation of an oxazoline and subsequent neutral loss. In mobile proton spectra, basic residues from the -2 to the +5 position inhibit neutral loss, though some of these are not significant. I suspect that when the basic residues are close to the phosphate, they will hydrogen bond, preventing the interaction between the phosphate and mobile protons or other basic sites that could better catalyze the neutral loss reaction, such as the N-terminus. As proton mobility
decreases, the presence of basic residues is less detrimental to neutral loss, suggesting that absent other reactive acidic sites, the hydrogen bonding of protonated-bases to the phosphate is less inhibitory. At the lowest levels of proton availability, the base-catalyzed mechanisms may be the prevalent method of neutral loss.

**The N-Terminus is the primary driver of neutral loss in mobile proton peptides**

The effect of the N-terminus on neutral loss of H$_3$PO$_4$ varies depending on proton mobility states. In spectra with immobile protons, proximity to the N-terminus inhibits neutral loss, with partially mobile protons the terminus has minimal effect, and with mobile protons the N-terminus strongly enhances neutral loss when 1 to 6 residues away. To test the extent of this effect I stratified the spectra by level of neutral loss and examined the average distance of the phosphosite to the termini (Figure 37a). For peptides with mobile protons, those with low neutral loss tended to be more than seven residues from the N-terminus, while being within four residues was required for the highest levels of neutral loss. While proximity of the phospho-residue to the N-terminus strongly enhances neutral loss, it is not sufficient; many such peptides still exhibit low neutral loss, indicating that the N-terminus alone is not sufficient for the elevation of neutral loss. Despite this lack of sufficiency, proximity to the N-terminus explains 4% of the variance in neutral loss exhibited by phosphopeptides with mobile protons, a larger effect than was observed for any other explanatory factor (Supplementary Table 1).
Figure 37. The effects of the N-terminus on neutral loss. a) The observed changes in proximity to the N-terminus (solid lines) and the C-terminus (dashed lines) in mobile (blue) and partially mobile (red) based on the observed amount of neutral loss displayed by a peptide. b) Distribution of neutral loss in b-ions derived from 37 peptides with acetylated N-termini. c) Proposed mechanism for neutral loss of H$_3$PO$_4$ catalyzed by the N-terminus to form an oxazoline. d) Spectrum of doubly protonated AQISpSPNLR. Red markers with the ion series denoted (x+2) indicate ions annotated as the x-type ions that have been reported as markers of oxazoline or macrocycle formation. e) Proposed mechanism for neutral loss of H$_3$PO$_4$ catalyzed by the N-terminus to form a macrocycle.

Enhancement of neutral loss by N-terminus proximity suggests involvement of the N-terminal amine group in the mechanism of H$_3$PO$_4$ loss. However, peptide length, proximity to the C-terminus, or other correlated factors could also explain the observed effects. To assess the importance of the N-terminal amine, I examined 37 mobile proton phosphopeptides from the WM239A dataset with acetylated N-termini. Acetylation of the N-terminus reduces the
basicity of the N-terminus. Consistent with the hypothesis of direct action, the acetylated peptides displayed on average 42% less neutral loss than the rest of the data. Interestingly, the distribution of neutral losses from the acetylated b-ions is more similar to that of amino-N-terminated y-ions than amino-N-terminated b-ions (Figure 37b). Thus, the bias of high neutral loss from b-ions is partially explained by the effects of proximity to the N-terminus. The effects of N-terminal acetylation imply that direct action by the N-terminal amine is responsible for the activation of neutral loss.

The N-terminus may participate in neutral loss of H$_3$PO$_4$ by nucleophilic attack or charge stabilization. Mechanisms involving nucleophilic attack by the N-terminus, such as the formation of diketopiperazine b-ions, also show maximal effect at the 2$^{nd}$ peptide bond$^{67}$. However, ab initio modeling studies indicate that nucleophilic attack by the N-terminus on the $\beta$-carbon of pSer and pThr to be higher energy than oxazoline pathways for neutral loss, even under optimal conditions$^{68}$. Furthermore, nucleophilic attack requires the N-terminus to be uncharged. Since the N-terminus is predominantly protonated in mobile but not in partially mobile cases, such mechanisms would be expected to be at least as prevalent in partially mobile cases as they are in the mobile cases. Consequently, the observation that proximity to the N-terminus enhances neutral loss in mobile proton spectra but not in partially mobile implies that the enhancement is dependent upon action by a protonated N-terminus. I hypothesize that under the oxazoline-forming mechanism (Figure 10b), the neutral loss of phosphoric acid by charge directed mechanisms is more efficient when the charge is stabilized on a more basic site and the phosphate interacts with the protonated N-terminus (Figure 37c).
Further evidence for charge directed, ring-forming mechanisms was found by several groups that independently identified neutral-loss-dependent cleavage of the backbone\textsuperscript{52, 69, 70}. This diagnostic cleavage occurs between the $\alpha$-carbon and amide carbon, one bond N-terminal to the ring that was formed by neutral loss. The resulting ion has been variously described as an “x-type ion”\textsuperscript{69}, a “y+10 ion”\textsuperscript{70} and a neutral loss\textsuperscript{52}. The diagnostic ion is seen both in neutral loss from pSer and pThr\textsuperscript{69} and during neutral loss of water from the non-phosphorylated cognates\textsuperscript{70}. Harrison also observed an additional ion attributed to a neutral loss mechanism analogous to oxazoline formation, but leading to a larger cyclized product incorporating one or more residues\textsuperscript{52}. Although the influence of surrounding amino acid side chains on dissociation events was not systematically explored, these studies point to previously unappreciated mechanisms underlying neutral loss in phosphopeptides that are currently not accounted for by current peptide identification algorithms.

Interaction of a charged N-terminus and the phosphoester group might bring the N-terminal amide carbonyl in proximity to the $\beta$-carbon of the phosphorylated residue, facilitating nucleophilic attack and formation of large cyclic structures after elimination, similar to those that Harrison\textsuperscript{52} proposed in neutral loss of water from unmodified serine and threonine. To assess this possibility, I looked for the x-type ion, here denoted (x+2), derived from cleavage at the first alpha carbon-carbonyl carbon bond, reported previously as a marker for this macrocyclic-cleavage\textsuperscript{52}. While less common than the x-type ion immediately N-terminal to the site of phosphorylation\textsuperscript{69}, peaks consistent with macrocycle x-type ions were found in MS/MS of several peptides. The spectrum of doubly protonated AQJSpSPNLR shows an especially avid
example. The \((x+2)_8\) ions suggest the formation of a macrocycle, while the \((x+2)_5\) ions indicate the formation of an oxazoline. Unexpectedly, in cases where an x-ion was present, I also observed a charge-reduced form, presumably due to loss of \(\text{H}_3\text{O}^+\). The presence of x-ions distal to the site of phosphorylation suggests that charge stabilization is leading to the formation of macrocycles in the neutral loss of phosphoric acid (Figure 37e).

**Proline provides a competitive pathway and reduces backbone flexibility required for distal basic interactions**

The analysis of the effects of flanking residues (Figure 35) reveals that proline strongly inhibits neutral loss regardless of position. I initially ascribed this effect to competition from the well-known enhancement of backbone cleavage N-terminal to proline. I examined cases from the synthetic phosphopeptide library dataset where the effect of single proline substitutions could be evaluated, and found evidence for competition in some spectra. An extreme example of this effect can be seen in Figure 38a, which compares the spectra of doubly protonated TPHVITEANpSGPR and TYHVITEANpSGPR, which differ only by the substitution of tyrosine for proline at the 2\(^{nd}\) position. Proline at the second position strongly activates the creation of the \(y_{11}^{++}\) ion relative to a tyrosine at that position. The ratio of the M-98 peak to other sequence ions is similar between the two spectra. Spectra such as this indicate that competition is an important factor in the inhibition by neutral loss; however, it does not fully explain the effect. A second group of proline-containing spectra shows a decrease in neutral loss peaks relative to all other ions. For instance, Figure 38b shows a spectrum in which the change of an alanine to a proline one residue N-terminal to the phosphosite significantly reduces the amount of neutral
loss relative to the b- and y-ion series and neutral losses of water and ammonia. Competition does not explain the change in ion ratios, indicating that these prolines may interfere with the mechanisms for neutral loss.

Figure 38. The effects of proline on neutral loss. a) Comparison of the spectra of doubly-protonated TXHVITEANpSGPR where X=P in the top panel and X=Y in the bottom panel. For lettering: blue denotes y-ions, green denotes b-ions, and red denotes losses from the parent. For ion markers: black denotes no neutral loss, red denotes loss of H3PO4, blue denotes loss of water or ammonia, purple denotes loss of H3PO4 and water or ammonia. b) Comparison of the spectra of doubly protonated DEIxpSFALQ where X=P in the top panel and X=A in the bottom panel. c) Representation of the increased chances of peptides from the ISB dataset being in the lowest tercile of neutral loss, given that there is a proline at the indicated position relative to the phosphosite. The y-axis is in units of log2 of the enrichment. Error bars represent the 95% confidence interval for the mean.
Proline’s cyclic side chain reduces the flexibility of peptide backbones, interfering with the formation of turns and loops. Given the evidence that neutral loss is enhanced by structures that bring the phosphate into proximity with immobilized protons, I hypothesized that proline’s direct reduction of neutral loss was caused by inhibiting loop formation. To examine this possibility, I plotted the magnitude of inhibition as a function of position relative to the phosphosite (Figure 38c). In peptides with mobile protons, proline shows strong inhibition when N-terminal to the phosphosite, consistent with interference with the loop required for interaction between the phosphate and the N-terminus. Proline slightly inhibits neutral loss when C-terminal to the phosphosite or when it is more than five residues N-terminal. I attribute the reduction in neutral loss from distal prolines to competition from increased backbone fragmentation. Peptides with partially mobile protons show high levels of inhibition when the proline is either C-terminal or N-terminal to the phosphate. This suggests that, while enhancement by the N-terminus that was observed with mobile protons remains, loops that allow the phosphate to interact with the basic residue at the C-terminus additionally encourage neutral loss. In immobile cases, proline shows most significant inhibition of neutral loss when C-terminal to the phosphosite. This can be rationalized under the loop-forming hypothesis since the N-terminus is expected to be uncharged in these peptides, leaving tryptic C-terminus as the only site that is generally protonated in these peptides.

Aspartic and glutamic acid

Aspartic and glutamic acid both increase the amount of neutral loss from position -6 to +5. It is possible for a carboxylic acid to initiate neutral loss through nucleophilic attack, yielding
a phosphoether, which would rapidly eliminate phosphoric acid under CAD conditions. This has been proposed as the mechanism for neutral loss of HPO$_3$ in pTyr$^{72,73}$. Alternatively, a carboxylate group could directly donate a proton needed to for the reaction or simply compete for hydrogen bonding with the phosphate, helping the peptide to escape from secondary structures that stabilize the phosphate rather than catalyzing its neutral loss.

To test the plausibility of direct action by acidic residues, I examined the phosphopeptide libraries for cases where a single acidic residue replaced another residue. Contrary to the strong enhancement suggested by the analysis of the PhosphoPep database, substitution incorporating single acidic residues has minimal effect on the observed amount of neutral loss. Indeed, single replacements tend to slightly lower the observed neutral loss rather than increase it. It should be noted the presence of one acidic residue greatly increases the chances of finding more acidic residues due to the presence of acidophilic kinase motifs. Thus, the increased neutral loss observed for nearby acidic residues in the global analysis shown in Figure 35 may be a synthetic effect of multiple acidic residues, rather than an effect particular to that position. I attribute the rise in the rate of neutral loss to the creation of an extensive hydrogen bond network within the peptide that is able to modulate or stabilize the transfer of protons to the phosphate.

**Section 3: Neutral loss of water and ammonia from threonine and other amino acids interferes with the loss of H$_3$PO$_4$**

While aspartic and glutamic acid showed a tendency to increase neutral loss, other amino acids with propensity for loss of water or ammonia from their side chains reduced the
observed neutral loss of H₃PO₄, especially when these amino acids are N-terminal to the phosphate and in peptides with mobile protons (Figure 35). Because this N-terminal bias is similar to the N-terminal enhancement of neutral loss of H₃PO₄, I hypothesized that a similar mechanism leads to analogous loss of water from modified and unmodified Ser and Thr, along with the loss of water and ammonia from asparagine, glutamine, and carbamidomethyl-cysteine.

If neutral loss of water and ammonia competes with the loss of H₃PO₄, there should be negative correlation between the two pathways. To assess the relationship between the pathways I examined the distribution of the neutral losses of water and H₃PO₄ among spectra (Figure 39a), which shows a negative correlation between the intensity of the two neutral loss events. Immobile proton cases favor the neutral loss of H₃PO₄, while mobile cases show higher loss of water. The reduction in neutral loss of water in peptide ions with less mobile protons implies that proton availability is a rate-limiting factor. While the difference in neutral loss between proton mobility groups is striking, within groups there remains a strong negative correlation between water and H₃PO₄ loss pathways. The immobile and partially mobile cases exist over a broad range of neutral losses of both water and phosphoric acid. In spectra with mobile protons, those peptides with high neutral loss of H₃PO₄ show negative correlations between the neutral loss of H₂O and H₃PO₄ similar to the correlations seen in immobile and partially mobile peptides. However, for low neutral loss of H₃PO₄, neutral loss of water becomes uncorrelated with loss of H₃PO₄. This suggests two regimes of neutral loss in mobile
proton cases: one in which there is high water loss that has effectively shut off loss of $\text{H}_3\text{PO}_4$ and another where the two neutral loss pathways compete.

Figure 39. The interaction between the neutral loss of phosphoric acid and the neutral loss of water. a) Contour plots showing the density of spectra when plotted by neutral loss of water or ammonia ($x$-axis) and neutral loss of $\text{H}_3\text{PO}_4$. Ions that showed loss of both were ignored. Shaded areas duplicate the data in the other graphs for ease of comparison, as with the contours blue shading indicates cases with immobile protons, green: partially mobile, and red: mobile. The contour lines increase with the square of density of spectra, so the innermost contour is 25-fold denser than the outer in mobile proton and partially mobile proton, and 9 times denser in immobile. b) Comparison of the spectra of doubly-protonated AGGPXpTPLSPTR where $X=T$ in the top panel and $X=A$ in the bottom panel. c) Proposed mechanism for the neutral loss of water from threonines N-terminal to the phosphate.
To determine what characteristics might be unique to these ions showing very low neutral loss of H₃PO₄, I examined the sequences of these peptides. When there is a mobile proton, peptides that exhibit high water neutral loss (greater than 0.4) and low phosphoric acid neutral loss (less than 0.2) contain 17% more serines and 85% more threonines than the average peptide in the dataset. Threonines that are N-terminal to the site of the phosphate are especially enriched, being more than twice as prevalent as in other peptides. The number of threonines N-terminal to the phosphate explains 3.6% of the variance in the neutral loss of H₃PO₄ from mobile proton phosphopeptides. This competition explains the general decrease of neutral loss of phosphoric acid when there is a threonine N-terminal to the phosphate (Figure 35). As an example of this effect the spectra of AGGPTpTPLSPTpTR and AGGPApTPLSPTpTR, which differ only by the substitution of threonine for alanine at position 5, are shown in Figure 39b. The addition of the threonine adds a highly active pathway for the neutral loss of water, as evidenced by the appearance of very strong M-18 and M-116 ions. This shows that the mechanism for the reduction of neutral loss of H₃PO₄ is competition with the loss of water.

While it is possible for the N-terminus to perform nucleophilic substitution directly, it must be unprotonated to do so. Consequently, I would expect that this mechanism would predominate in partially mobile and immobile proton cases, not mobile proton cases. Therefore, it is likely that the N-terminus is stabilizing the charge required for neutral loss of water from unphosphorylated threonine, in the same way proposed for phosphorylated residues (Figure 39c). It is possible that coordination of the N-terminus and the phosphate brings the terminus into closer contact with the unmodified alcohol side chain, enhancing the
loss of water. The mobility-dependent reduction of neutral loss of water implies that the availability of protons is a rate-limiting factor for neutral loss of water. Thus, I hypothesize that threonine is more effective than serine because the gas-phase basicity of its side chain is about 14kJ/mol higher than serine\textsuperscript{75}, allowing the proton to be more readily transferred to threonine. Competition for a charge-stabilizing partner could explain the strong negative correlations that are observed between the neutral loss pathways.
Chapter 5: Future work suggested by these studies

Section 1: Statistically robust scoring for intensity-based spectral matching

Many of the problems that emerge from the identification of peptides that exhibit large neutral losses could be addressed by the introduction of scoring metrics that better address the variability in the intensity of neutral losses. Current scoring metric tend to overestimate the importance of neutral loss peaks, because, while these peaks tend to be large, they are wildly variable in intensity and difficult to accurately predict. Consequently, statistical methods should be more robust than the similarity-based methods that are generally used. To address this, I propose to use a Bayesian scoring metric based on empirically observed variation in intensity from predicted values.

Given a spectrum $S$ and defining hypothesis $P$ as $S$ corresponding to a particular peptide, Bayes theorem indicates that the probability of $P$ being correct is:

$$p(P|S) = \frac{p(P) p(S|P)}{p(S)}$$

If we assume that spectrum $S$ is comprised of the parent mass $M_0$ and $n$ spectral features $s_1, s_2...s_n$, this can be rewritten as:

$$p(P|s_1, s_2...s_n, M_0) = \frac{p(P) p(s_1, s_2...s_n, M_0|P)}{p(s_1, s_2...s_n, M_0)}$$

Further assuming that the probabilities of $s_1, s_2...s_n$ are independent, this can be simplified to:
\[
p(P|s_1, s_2 \ldots s_n, M_O) = \frac{p(P) \prod_{i=1}^{n} p(s_i|P, M_O)}{p(M_O) \prod_{i=1}^{n} p(s_i|M_O)} \\
= p(P) \frac{p(M_O|P)}{p(M_O)} \prod_{i=1}^{n} \frac{p(s_i|P, M_O)}{p(s_i|M_O)}
\]

Consequently, in order to estimate the probability of \( H \), we must estimate \( p(P) \), \( p(M_O|P) \), \( p(s_i|P, M_O) \), and \( p(s_i|M_O) \). The prior probability, \( p(P) \), is difficult to determine exactly, but, given the need for unbiased identifier, a constant should be used. The inverse of the number of peptides in the candidate database is likely an appropriate value.

Using a model of the expected spectrum from a peptide, \( p(s_i|P, M_O) \) can be approximated. \( s_i \) is defined by the mass, \( m_i \), and intensity, \( I_i \), of the corresponding peak, \( i \). Generating a predicted spectrum, will yield \( n \) peaks. The mass and intensity of the \( j \)th peak will be denoted \( m_j \) and \( i_j \).

Assuming that the errors in predicting intensity are independent of parent mass and that errors in mass, \( e_m \), and errors in intensity are independent:

\[
p(s_i|P, M_O) = p(s_i|P) = \sum_{j=1}^{n} p(e_m \geq |m_i - m_j|) \ p(I_i|i_j)
\]

In practice the mass error term is usually assumed to be a probability of 1 when \(|m_i - m_j|\) is less than a critical mass error based on the mass accuracy of the spectrometer, \( e_m' \), and zero otherwise. \( p(I_i|i_j) \) can be approximated empirically by observing the distribution of observed intensities of ions which are predicted by our model to have intensities of approximately \( i_j \). We will denote distribution of observed intensities as \( O_j \). Equation 4 simplifies as:
\[ p(s_i|P) = 2 \sum_j \mathcal{H}(e_{m'} - |m_i - m_j|) \sup\{ \mathcal{H}(x - l_i) - p(l \leq x|O_j, 0, \infty) : x \in \mathbb{R} \} \]

where \( \mathcal{H}(x) \) is the Heavyside step-function, \( \sup\{Y\} \) indicates the supremum of set \( Y \), and \( p(l \leq x|O_j, 0, \infty) \) is the cumulative distribution function of \( O_j \). A single count at zero and one at infinity are added to the observed distribution to correct for sampling.

Referring again to Equation 3, the probability of observing a feature in a random spectrum, \( p(s_i) \) remains. This can be derived empirically from a set of observed spectra. The background intensity distribution is known to depend heavily on the mass of the peaks. We will denote the distribution of background intensities daughter ion, \( j \), resulting from parent ion of mass \( M_O \) as \( B_{j|M_O} \). By analogous methods to Equation 5 the value is:

\[ p(s_i|M_O) = 2 \sum_j \mathcal{H}(e_{m'} - |m_i - m_j|) \sup\{ \mathcal{H}(x - l_i) - p(l \leq x|B_{j|M_O}, 0, \infty) : x \in \mathbb{R} \} \]

Using a similar treatment the posterior and prior probabilities of the parent mass, \( p(M_O|P) \) and \( p(M_O) \), can be determined. Assuming that we accept any mass which is within the critical mass error, \( e_M' \) (note that this is likely different from the MS² mass error), and that the mass expected under \( P \) is \( M_P \):

\[ p(M_O|P) = \mathcal{H}(e_M' - |M_O - M_P|) \]

To find the prior probability, \( p(M_O) \), we first find the distribution of masses of parent ions from a database, \( X \). We can then calculate the probability of a random match using:

\[ p(M_O) = p(M \leq M_O + e_M'|X, 0, \infty) - p(M \leq M_O - e_M'|X, 0, \infty) \]
Synthesizing equations 1, 3 and 5-8, yields an equation that is expressed solely in terms of observable quantities:

$$p(P|S) = p(P) \frac{\mathcal{H}(e_{M'} - |M_0 - M_p|)}{p(M_0 - e_{M'} \leq M \leq M_0 + e_{M'} |X, 0, \infty)}$$

$$\times \prod_{i=1}^{n} \sum_{j} \mathcal{H}(e_{m_i} - |m_i - m_j|) \sup\{\mathcal{H}(x - l_i) - p(l \leq x|O_i, 0, \infty) : x \in \mathbb{R}\}$$
$$\sum_{j} \mathcal{H}(e_{m_i} - |m_i - m_j|) \sup\{\mathcal{H}(x - l_i) - p(l \leq x|B_{j|M_0, 0, \infty}) : x \in \mathbb{R}\}$$

For any given spectrum the portion that depends on the parent mass will be either a single constant value or zero. In most classification tasks, the prior, $p(P)$, will also be treated as a constant. In many applications it may be more useful to only consider the part of the equation inside the product operator. This part of the equation exclusively contains all information from the MS² scan.

It should be noted that the assumption of independence between spectral features is a weak one. There are numerous highly correlated features in MS² spectra of peptides, notably isotope peaks, complementary ions, and ions offset by the mass of compounds that are expected to undergo neutral loss. In practice, classifiers generally perform well despite such deviations from their assumptions, because such deviation occurs for all cases which are classified. However, the assumption of independence will likely lead to overestimation of classification probabilities. The simplest methods to correct this are to either arbitrarily lower the prior probability of a match or to raise the significance threshold used for acceptance.
More complex methods which explicitly reduce the assumptions of independence at the cost of increased computational cost and database size may be more appropriate.

Average one-dependence estimators is a method to reduce the assumption of independence between variables.76 These methods replace the strictly independent application of Bayes theorem with the more complex:

\[
p(P|s_1, s_2...s_n, M_0) = \frac{1}{n} p(P) \frac{p(M_0|P)}{p(M_0)} \sum_{i=1}^{n} \left( \frac{p(s_i|P, M_0)}{p(s_i|M_0)} \prod_{j=1}^{n} \frac{p(s_j|P, M_0, s_i)}{p(s_j|M_0, s_i)} \right)
\]

Assuming that the spectral model already accounts for the expected correlations, \( p(s_j|P, M_0, s_i) \) can be treated as independent of \( s_i \). The resulting expression is

\[
p(P|s_1, s_2...s_n, M_0) = \frac{1}{n} p(P) \frac{p(M_0|P)}{p(M_0)} \sum_{i=1}^{n} \left( \frac{p(s_i|P, M_0)}{p(s_i|M_0)} \prod_{j=1}^{n} \frac{p(s_j|P, M_0)}{p(s_j|M_0, s_i)} \right)
\]

\( p(s_j|M_0, s_i) \) is the only term that does not appear in equation 3. This term is similar to \( p(s_j|M_0) \) but has one additional conditional. This follows a distribution like the background distribution \( B_{j|M_0} \) which was introduced earlier, except that the new distribution will only contain those background spectra which have an ion like \( s_i \). We will denote this distribution as \( B_{j|M_0,s_i} \). Replacing equation 11 for equation 3 an equation analogous to equation 9 emerges:
\[ p(P|s_1, s_2, \ldots s_n, M_O) \]
\[
= \alpha \sum_{i=1}^{n} \left( \frac{\sum_k \mathcal{H}(e_m' - |m_i - m_k|) \sup\{|\mathcal{H}(x - l_i) - p(l \leq x|O_k, 0, \infty)|: x \in \mathbb{R}\}}{\sum_k \mathcal{H}(e_m - |m_i - m_k|) \sup\{|\mathcal{H}(x - l_i) - p(l \leq x|B_{k|M_O}, 0, \infty)|: x \in \mathbb{R}\}} \right) \times \prod_{j=1}^{n} \left( \frac{\sum_k \mathcal{H}(e_m' - |m_j - m_k|) \sup\{|\mathcal{H}(x - l_j) - p(l \leq x|O_k, 0, \infty)|: x \in \mathbb{R}\}}{\sum_k \mathcal{H}(e_m - |m_j - m_k|) \sup\{|\mathcal{H}(x - l_j) - p(l \leq x|B_{k|M_O,S'}, 0, \infty)|: x \in \mathbb{R}\}} \right)
\]

where: \( \alpha = \frac{1}{n} \frac{p(M_O|P)}{p(M_O)} \)

This method can be augmented by elimination of variables that show high dependence on another variable.\(^7\) This would provide an efficient method for the removal of isotope peaks and highly correlated neutral loss peaks.

The Bayesian scoring method provides a rapid and generalizable method for determining the quality of a spectral match to another spectrum provided either by a theoretical model or by prior observation. It can handle models of arbitrary complexity. It is possible to easily add evidence from other sources, such as chromatographic retention, ion mobility, or higher order mass spectra. Additionally, at least in the version encoded by equation 9 all responses to evidence are independent, which is a useful feature both for model development and in differential analysis. The primary drawback is the requirement that large databases of sample spectra be kept.
Section 2: Predicting the intensities of neutral losses through prediction of secondary structures

Finding that long-range charge pairing is essential for neutral loss reactions implies that an improved kinetic model of peptide fragmentation is needed. At a minimum, this model must allow coordination of charge between pairs of side chains mimicking the formation of secondary structures necessary for proton solvation. This more nuanced model of basic sites will allow more accurate prediction of proton distributions, especially in cases where bases are near to each other. It will also allow the addition of new mechanisms of neutral loss requiring the action of solvated protons, improving the prediction of neutral loss ions.

Since protonation depends exclusively on ion-dipole interactions, the energy of protonation is a function of the inverse-square of proton-base distance. Consequently, for two side chains the with gas phase basicities $GB_1$ and $GB_2$, the basicity of the paired site should be the root sum-squared of the basicities, $GB_{1,2} = \sqrt{GB_1^2 + GB_2^2}$. This resulting basicity is likely very large, but must be modified by penalties accounting for the steric strain and reduction in entropy yielding the basicity,

$$GB_{1,2} = \sqrt{GB_1^2 + GB_2^2} - \Delta H_{1,2} + T\Delta S_{1,2}$$

When this site is protonated, it would be assumed that both sites are protonated for the purposes of repulsion between multiple charges. In this prediction, no new fitted parameters for basicity are needed, allowing the extension of the model with minimal additional complexity. However, values for $\Delta H$ and $\Delta S$ must be fitted.
The values of $\Delta H$ and $\Delta S$ are not known, but it is reasonable to assume that they are a function of the distance, in bonds, between the two bases. It seems reasonable to assume that $\Delta S$ increase approximately linearly with the number of bonds constrained by ring formation. Therefore with a single new fitted parameter $S$, and defining $d_{1,2}$ as the distance between the bases:

$$\Delta S_{1,2} = S \times d_{1,2}$$

The value of $\Delta H$ is more complex; however, since these reactions are side chain to side chain there are not going to be distances shorter than 8 bonds. At this length and longer, rings tend to be flexible enough that bonds are not highly torsionally strained, and entropic effects dominate over enthalpic effects. Therefore, this term can likely be ignored. Given the evidence that proline hinders ring formation, it may, however, be advantageous to include an energetic penalty for rings containing proline.

The inclusion of solvated charge in the kinetic model will improve the quality of the calculated proton distributions. While these calculations require few additional fitted parameters, the computational expense of calculating fragmentation will increase significantly. This is particularly important since finding the proton distribution is the most time consuming portion of the kinetic calculations. These calculations are currently approximately $O(N^2)$ in complexity with the number of basic sites in the peptide. If all possible pairings were considered, this would increase to approximately $O(N^4)$, which would be overly expensive for our purposes. Consequently, I would only consider solvation sites that involve the N-terminus,
Arg, Lys, or His. This would make calculations about fourfold slower than current, but this is an acceptable decrease.

Once the solvated proton distribution is found, the rate of neutral loss from side chains solvated with bases must be found. The amino acids that can undergo such losses are Ser, Thr, Asp, Asn, Glu, Gln, Cys, and Arg. Since most charge-directed processes act through the induction of dipoles followed by nucleophilic attack, the reactivity of a site should be approximately proportional to the cube root of the partial charge at the basic site. The partial charge on each site can be found using the Boltzmann distribution. Thus the reactivity of base 1 in the conjugated site is:

$$\rho'_{1} = \sqrt[3]{\frac{e^{-\frac{(GB_2-GB_1)}{k_B T}}}{1+e^{-\frac{(GB_2-GB_1)}{k_B T}}}}$$

Then, the overall rate of neutral loss reactions where $\rho_i$ is the population fraction of charge state $i$, $\rho'_i$ is the reactivity of charge state $i$, as above, $A$ is a fitted preexponential factor for the reaction, and $E$ is the fitted activation energy for this reaction:

$$k = \sum_i (\rho_i \times \rho'_i) Ae^{-E/k_B T}$$

By these mechanisms it is possible to increase the accuracy of the proton distribution model and the reaction rates of neutral loss, while adding only one new fitted parameter to the kinetic model. The inclusion of solvated protons is an important addition that will more closely mimic the real proton dynamics experienced by protonated peptides. Hopefully this model will also allow a more realistic balance between the rates of neutral loss and backbone cleavages.
Realistic predictions of the rate of neutral loss would stop these ions from being hindrances to peptide identification and instead allow information to be drawn from their presence.
Chapter 6: Conclusions

The inclusion of fragment ion intensities in the identification of peptides from their MS$^2$ spectra greatly increases the useful information in a spectrum and allows more accurate identification in proteomics. The prediction of fragment intensities relies upon accurate computational models, capable of performing in silico fragmentation. Using kinetic parameters for expected fragmentation pathways in the hybrid statistical-theoretical approach that was first introduced in Mass Analyzer,\textsuperscript{32} has proved highly-successful rivaling purely statistical methods. The improvements to Mass Analyzer that we have implemented in S3 refine the capability and allow expansion of the model through the addition of novel mechanisms of gas phase chemistry.

Additions to S3, such as N2 cleavage when proline is at the second position incrementally make the prediction of peptide sequence ions more accurate. Interestingly, while the model is informed by knowledge of gas phase chemistry, it returns the favor by providing clues to the details of mechanism. Attempting to fit the N2 cleavage indicated that the model was significantly more accurate if the mechanism was assumed to be charge remote. Since reactions of this class had never been previously described, it is unlikely that I would have uncovered them without the statistical learning inherent in the model. Through this reciprocal refinement, the accuracy of prediction of ion intensities is only likely to increase.

While the prediction of sequence ions is improving, the prediction of ions resulting from neutral losses from amino acid side chains remains problematic. The prediction of these ions is highly variable compared to the prediction of sequence ions. This problem extends into peptide
identification where it is better to simply ignore neutral loss ions than to try to use the information contained in them. Refinement of the prediction of neutral loss ions is, consequently, an important goal for the development of kinetic models, especially because neutral loss is prevalent in phosphopeptides, which are of significant biological interest.

Investigations of the neutral loss of \( \text{H}_3\text{PO}_4 \) from phosphopeptides have struggled to reconcile the expectation that these reactions should be charge-directed and the observation that the predominance of neutral loss pathways decreases when protons become more available. Previous studies have suggested direct catalysis by protonated basic residues, especially via non-covalent interactions between the phosphate and nearby arginines\(^ {47} \). At first glance, this view appears to be supported by our evidence that the presence of basic residues increases the amount of neutral loss. However, in an example of Simpson’s paradox, when we stratify the spectra by charge-mobility, we find evidence for the converse, that nearby bases inhibit neutral loss regardless of mobility classes. Furthermore these effects were not limited to adjacent residues but extended across the entire peptide. Additionally, we found that the N-terminus showed strong positive effects on \( \text{H}_3\text{PO}_4 \) neutral loss when both proximal and distal to the phosphorylation site, making it the primary predictor of neutral loss in mobile proton peptides.

These observations suggest a model in which the phosphate is always in complex with protons that are immobilized at basic sites. Because these complexes are constitutively present, the basicity of the phosphate itself becomes unimportant to its rate of neutral loss. Instead the rate is dominated by the basicity of the hydrogen bonded base, since that basicity determines
the partial charge that resides on the phosphate and consequently the susceptibility of the phosphorylated side chain to nucleophilic attack. The N-terminus, therefore, becomes the primary driver of neutral loss because it is the least basic site that gets protonated with high occupancy, making it the ideal catalyst. This equilibrium may be further modified by the participation of other side chains, such as by proton donation from the carboxylic acids on Glu and Asp. Because these base-mediated complexes continue, even in the presence of mobile protons, the reaction rate of $\text{H}_3\text{PO}_4$ neutral loss increases much more slowly with charge availability than the rates of backbone fragmentation.

The neutral loss of water from Ser and Thr anticorrelates with the loss of $\text{H}_3\text{PO}_4$ from the phosphorylated cognates. We suggest that this is because the formation of stable complexes between protonated bases and the alcohol side chains is much less favorable. Consequently the formation of the complex becomes rate-limiting, while even small partial charges on the Ser or Thr side chain is enough for the neutral loss of water. Interaction with the N-terminus still appears to be an important determinant for these neutral losses, and the competition for charge pairing with the terminus defines the neutral loss pattern. It further appears that other amino acids with side chains capable of neutral loss show effects when near the N-terminus, suggesting that interaction of side chains with immobilized protons may be the general mechanism of neutral loss in peptides.

It is worth noting that neutral loss reactions catalyzed by forming secondary structure to interact with charge donors cannot be usefully described by the mobile proton model. The mobile proton model posits three reaction mechanisms: charge remote mechanisms, charge
directed mechanisms involving a mobile proton, and charge directed mechanisms involving a proton immobilized at the site of the reaction. We suggest that neutral loss reactions are catalyzed by a proton that is immobilized distal to the reaction site; however, by the adoption of secondary structures, it interacts with the reaction site. Under this model, the peptide acts more like an irregular solvation shell around the proton, than like a series of discrete protonation sites. While the conjugation of charge between multiple bases is important in all CAD reactions, the evidence here suggests that it may be especially important for neutral losses. This may explain some of the difficulties in the prediction of neutral loss using kinetic models based on the mobile proton model\textsuperscript{32-34}.

This study demonstrates the importance of using datasets that are much larger than those traditionally used in the field. Systematic analysis of the influence of peptide sequence requires datasets ten-fold larger than used in previous studies to achieve statistical significance of results. Large datasets permit application of the multiple hypothesis testing necessary to guard against inflation of type I errors without excessive effects of lowering sensitivity. While obtaining very large libraries of spectra locally is prohibitively difficult, we exploited the recent availability of large curated MS\textsuperscript{2} phosphopeptide libraries, which enabled large-scale analysis of local sequence effects on H\textsubscript{3}PO\textsubscript{4} neutral loss. While it is necessary to carefully tailor analyses to overcome the reduction in control of both data acquisition and data processing inherent in publically available curated data, it provides an invaluable resource.

The size of datasets required for comprehensive study means that degenerate peptide libraries, while tempting in their experimental simplicity, are generally unsuited to open-ended
explorations of peptide fragmentation. Technical difficulties in synthesis ensure that libraries with more than tens of expected sequences become difficult to work with. Furthermore the problem of contamination ensures that experimenters cannot rely upon a priori knowledge of sequence, the primary advantage of synthetic libraries over biological spectra. While these problems make peptide libraries difficult to use in open-ended studies, the highly correlated nature of the peptides makes then ideal for more targeted studies in which the effects of minor perturbations under controlled conditions are assessed. Therefore, libraries provide a complementary technique to the growing diversity of publically available data.

The presence of phosphate and the resulting neutral loss inhibit accurate prediction of expected fragmentation ions in the spectra of phosphopeptides. This uncertainty reduces the accuracy of peptide algorithms, since the neutral loss provides information that is at best uninformative and usually confounding. This proves especially true to algorithms that attempt to use the intensity of fragment ions to inform fragmentation, since the size of neutral loss ions causes the variations to overpower the remaining information in the spectrum. Developing models that incorporate the observation that many amino acids have effects not only when they are adjacent to the phosphosite, but also when they are distal to it, requires fundamental redesign of the algorithms. However, it is possible to create a new model of peptide basicities that incorporates the possibility of solvation of protons across multiple sites without significantly increasing the complexity of the underlying model. With improved calculations of the proton distribution the prediction of peptides and phosphopeptides, alike, will be improved. This could be supplemented by the development of scoring metrics for peptide
matches which explicitly take into account the variability of the underlying data, a failure that has, to date, made neutral loss peaks especially detrimental to peptide identification.

Furthermore, knowledge of site specific variations in expected neutral loss, allows prediction of the site of phosphorylation even in the absence of sequence ions to support the localization. By these methods, the variability of neutral loss becomes informative in phosphopeptide identification rather than a detriment.

By improving the sensitivity of identification, it is obviously possible to increase the number of peptide and protein identifications. Consequently depth and completeness of coverage of the proteome will be improved. However, more interestingly, increased sensitivity will allow tests that query the presence of a more diverse range of targets. In current experiments, only a minority of peptides are identified. Many of the peptides that we are unable to identify contain unexpected post-translational modifications. With more powerful identification algorithms, it should be possible to check a wider variety of possible modifications. Since modified peptides are often the most biologically interesting, this would pave the way for great increases in tying proteomics to cellular function.
Appendix A: Methods for chromatography and mass spectrometry

WM239A dataset: Samples from cellular extracts

Scott Stuart prepared and ran the biological samples that were used in this work. Phosphopeptide samples were obtained from the human melanoma cell line, WM239A. Cells were lysed in boiling SDT buffer (100 mM Tris pH7.6, 4% SDS, 100 mM DTT) and lysate was sonicated for 15s. Buffer exchange, iodoacetamide cysteine alkylation and trypsin digestion was performed by the FASP method. Peptides were desalted on Oasis HLB columns (Waters) and dried by vacuum centrifugation. Peptides were fractionated prior to phosphopeptide enrichment, using ERLIC chromatography. Briefly, peptides were solubilized in 70% acetonitrile, 20 mM ammonium formate (pH 2.2) and loaded onto a 4.6 mm x 150 mm PolyWAX LP column (PolyLC) at 1 ml/min using an Agilent 1100 HPLC (Agilent Technologies). Peptides were eluted as follows, collecting 1ml fractions: 0-5 min with buffer A (70% acetonitrile (MeCN), 20 mM ammonium formate pH 2.2), 5-15 min with a linear gradient to 100% buffer B (10% MeCN, 20 mM ammonium formate pH 2.2), 15-20 min linear gradient to 100% buffer C (10% MeCN, 1M ammonium formate, pH 2.2), 15-20 min linear gradient to 100% buffer D (10% MeCN, 1% TFA), 20-24 min wash with buffer D, followed by re-equilibration of the column with buffer A. Fractions were concentrated by vacuum centrifugation to 12.5ul.

Phosphopeptides were enriched using a batch method with titanium dioxide beads (Titansphere, GC Sciences). ERLIC fractions were diluted to a final volume of 400ul with loading buffer (65% acetonitrile, 2% TFA, 140 mM glutamic acid). Titanium dioxide beads were washed in 65% acetonitrile, 2% TFA, followed by a wash with loading buffer, and added at a
peptide-to-bead ratio (w/w) of 1:20 and rotated 15 min at room temperature. Beads were washed once with loading buffer, once with 65% acetonitrile, 0.5% TFA, and twice with 65% acetonitrile 0.1% TFA. Beads were resuspended in 0.1 ml of 65% acetonitrile 0.1% TFA and packed onto the top of a 200 ul C8 Stagetip (ThermoFisher). Phosphopeptides were eluted with 100 ul of 20% acetonitrile, 1% NH₄OH into a receiving tube with 20ul of 25% acetonitrile, 1% TFA to neutralize the pH. Remaining phosphopeptides bound to the C8 resin were eluted with two 100ul volumes of 65% acetonitrile, 1% NH₄OH into the same tube. Samples were dried by vacuum centrifugation.

**Mass spectrometry**

Dried phosphopeptides were solubilized in 0.1% formic acid and directly injected onto a BEH C18 column (25 cm x 75 μm i.d., 1.7 μm bead, 100 Å pore size, Waters, part #: 186003545) on a 2D nanoAcquity system (Waters) in direct injection mode. Peptides were eluted with a linear gradient from 95% buffer A (0.1% formic acid) to 30% buffer B (0.1% formic acid in acetonitrile) in 120 min at a flow rate of 300 nL/min. Mass spectrometry analysis was performed on a LTQ-Orbitrap (ThermoFisher). Survey scans were collected in the Orbitrap at 60,000 resolution (at m/z=300), and MS/MS sequencing was performed by CAD in the LTQ in data-dependent mode, using monoisotopic precursor selection and rejecting singly charged and unassigned precursors for sequencing. The 10 most intense ions were targeted. After two observations of a peptide, dynamic exclusion of ±10 ppm mass lasting 180s was applied. The maximum injection time for MS survey scans was 500 ms with 1 microscan and AGC= 1x10⁶. For LTQ MS/MS scans, maximum injection time for survey scans was 250 ms with 1 microscan and
AGC= 1x10^6. Peptides were fragmented by CAD for 30 ms in 1 mTorr of N\textsubscript{2} with a normalized collision energy of 35\% and activation q=0.25.

**Phosphopeptide identification**

MS/MS spectra were extracted with readw (version 4.3.1) and searched with Mascot (v 2.2, Matrix Science) against a human IPI 3.27 protein database, with methionine oxidation, N-terminal pyroglutamic acid (Gln) and phosphorylation on Ser, Thr and Tyr as variable modifications. Precursor m/z error was 20ppm and fragment m/z error was 0.4 Da. A custom software pipeline was used to extract identifications from Mascot DAT files. Phosphopeptide identifications were accepted at 1\% FDR at the peptide level determined by separate search of a reversed database.
Citations


