Powering phosphoproteomics with large scale data analysis and machine learning

Anthony S. Barente

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Judit Villén, Chair

William S. Noble

Shao-En Ong

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Powering phosphoproteomics with large scale data analysis and machine learning

Anthony S. Barente

Chair of the Supervisory Committee:
Judit Villén, Associate Professor
Genome Sciences

Cells are the fundamental biological units of organisms and are constantly changing their internal state in response to external stimuli and stresses. A common way in which they do this is through the addition and subtraction of chemical tags from proteins, which allows the cells to exert fine grained control over protein activity. One of these tags, phosphorylation, is unique for its essential role in signaling cascades. By linking together chains of proteins turning on and off each other through phosphorylation, cells can build sophisticated networks capable of transforming stimuli into the appropriate biological response. High throughput tools such as mass spectrometry are ideal for studying phosphorylation, as they provide the capability to track the dynamics of thousands of modified sites across treatments. In recent years, this technique has only become more popular, with the number of submissions to public repositories for mass spectrometry data growing every month. By bringing together

multiple phosphorylation studies into one dataset, we have the potential to learn fundamental properties about how phosphopeptides behave across instruments, and improve our assays. In addition to the amount of data, phosphoproteomics datasets have continued to grow in size with the improvement of sample preparation and data acquisition technologies. While this growth allows for more conditions and subjects to be included in a single study, it comes along with fundamental computational and statistical challenges. Within this thesis, I will present two stories which explore these avenues of research. First, I will present the analysis of a large scale yeast phosphoproteomics perturbation screen. With this I will show how the comparison of phosphosite dynamics across multiple treatments can lead to prioritized targets for further research and provide valuable information about the regulatory relationship between phosphosites. After this analysis, I will present my efforts to build a centralized resource for building targeted phosphoproteomics assays. Here I will first present pyAscore, a versatile and fast python package for performing an essential step in phosphopeptide identification. Then, I will detail an automated and reproducible pipeline for integrating publicly available phosphoproteomics data into a centralized knowledgebase, Phosphopedia 2.0. Finally, I will present work to predict phosphopeptide retention time and charge state from amino acid sequence, which has allowed Phosphopedia 2.0 to move beyond detections and provide information about any phosphopeptide.

ACKNOWLEDGEMENTS

I have had many mentors over the course of my career in science, both at the University of Washington and before. My sincerest gratitude goes to Dr. Lynn Pillitteri, who decided to give me not only the opportunity to try out science as a career during my undergraduate education at Western Washington University, but was trusting enough to allow me to lead a major project. The skills that I learned during my time in the Pillitteri Lab were essential to my success in graduate school. Similarly, I would like to thank Dr. Judit Villén. Early in graduate school, when I was struggling to understand where I wanted to apply my efforts, Dr. Villén was kind enough to give me a place where I could be free to perform the research that interested me and learn skills which continue to bear fruit today. Certainly, many things went awry during my time at UW, but Dr. Villén was always there to provide kind words, sound advice, and prudent direction. Finally, I would like to thank my fellow graduate student Dr. Ian Smith for being one of the best mentors throughout my time at UW. I started with almost no knowledge on the topics and technologies detailed in this thesis. Through patience and kindness, Dr. Smith built my confidence in this subject and is one of the main reasons I was able to complete my goals.

Other individuals have had a profound impact on me throughout my time at the University of Washington, and I would be remiss to neglect their mention. The whole Villén lab has been one of the most collaborative environments that I have had the pleasure of working in. My thanks in particular goes out to Kyle Hess, Mario Leutert, Alex Hogrebe, Ricard Rodriguez, and Miguel Martin, who have been wonderful resources of proteomics knowledge. In addition, I would like to thank members of the

William S. Noble Lab, who were instrumental in helping me to understand the computational side of proteomics.

Finally, I would like to thank my family. My parents have always been unconditionally supportive of my education goals, and I could not have done this without them cheering me on. In addition, I would like to thank my wife, Breanne Barente. Graduate school is full of uncertainties and stresses, which continue to take their toll day in and day out. I benefited so much from having Breanne by my side through the whole process, and her support means the world to me. Whatever lies ahead, I get to tackle it hand in hand with my best friend.

Table of Contents

Chapter 1. Introduction

Cells must constantly synthesize information about their internal and external environment, and respond to diverse signals with the appropriate change in cellular state. Much of this signal transduction is facilitated by proteins, whose behavior can be precisely tuned via the addition and removal of chemical post-translational modifications (PTMs). The cleavage of an ATP molecule to add a phosphate group to a protein, termed protein phosphorylation, is particularly important do to its far reaching roles in biological processes.[1,2] Precise control of this type of PTM is mediated by dense networks of interconnected kinases and phosphatases, which exhibit complex and context dependent logic in response to cellular stimuli.[3–5] Dysregulation of the phosphorylation network has been identified as a hallmark of many diseases, and it is likely that rewiring happens frequently in cancer due to mutation and drug resistance.[4,6] This makes the study of protein phosphorylation ideal for building new therapies and elucidating the dynamics of cellular signaling in general.

Despite catalogs of hundreds of thousands of phosphorylated residues,[7] our understanding of the regulatory mechanisms of the phosphoproteome is only just developing.[5] Recent years have seen a boom in the amount of phosphoproteomics data uploaded to public repositories such as PRIDE and MassIVE.[8] Computational integration of these data can yield novel insights about the phosphoproteome, or even aid in the production of future assays.[9,10] A continued push to increase sample throughput has also facilitated the production of large scale datasets, able to characterize the dynamics of thousands of sites across panels of perturbations and tissues.[4,11–13] These avenues have brought an era of big data to the field of

phosphoproteomics, with unique computational challenges. The goal of this thesis is to describe the computational and analytical challenges unique to integrating information about the phosphoproteome across diverse treatments and datasets, and detail the benefits of these avenues of research.

1.1 Building knowledge from single large datasets

Continued improvements in instrumentation and sample preparation techniques have led to deep characterization of cellular phosphoproteomes, with discovery of new phosphosites outpacing our ability to describe their regulation.[5,14] Through panels of tissue types and complementary network perturbations, the underlying signaling cascades have been re-imagined as highly interconnected networks of kinase-substrate relationships capable of complex logic.[3,4,13,15] This can lead to large portions of the phosphoproteome experiencing differential regulation across many conditions. But with the expansion of treatment panels, phosphosites may eventually begin to show differences in activation or deactivation, allowing description via statistical methods which discover regulatory modules, such as weighted gene correlation network analysis.[16] If a high quality protein-protein interaction network is available, and treatments are selected to target highly interconnected sets of nodes, it may even be possible to produce a descriptive model of the underlying signaling logic.[17] By producing a large-scale panel of perturbations, a picture of the phosphoproteome can be developed in unprecedented detail. With recent work into decreasing sample preparation time through robotic automation and decreasing instrument time through continued expansion of sample multiplexing, more researchers may be tempted to take

this route in building this kind of study.[11,18] Here, we will briefly describe the types of insights which can come from large panels of treatments and tissues before delving into the statistical considerations which come along with their analysis.

1.1.1 The promise of large-scale descriptions of signaling networks

Efforts to describe the diversity of signaling networks both within and between organisms have been especially fruitful. Early studies compiled atlases of phosphorylation across tissues of mouse and rat, defining shared and tissue specific network logic.[13,19] These works were able to show that proteins which experience phosphorylation are often expressed in multiple tissues, yet the set of occupied phosphosites may vary greatly. This implies a highly dynamic nature to the underlying signaling network, with linkages experiencing substantial rearrangement to suit a tissue's needs. While a similar systematic effort exists to describe the tissue diversity in the global proteome in humans,[20] the field is still open to thorough descriptions of the phosphoproteome. Across species, principles of network evolution have started to emerge as well. One study of phosphorylation across yeast species showed rapid divergence in occupied sites and reliance on particular classes of kinases, implying high flexibility of signaling networks to adapt to species environment.[21] These type of studies often must rely on qualitative descriptions of phosphoproteome dynamics, due to the relatively low overlap in detections. Despite this, their importance remains undiminished to contextualize our study of phosphosite dynamics across perturbations.

In contrast to studying the global phosphoproteome across diverse biological systems, focusing on a single model system provides a snapshot of network organization which can be probed in depth. Yeast provide a straightforward route for this

analysis, as they allow the systematic production of kinase and phosphatase knockouts covering a signaling sub-network of interest.[3,15] Phosphosites which are differentially abundant after the knockout may be counted as putative downstream targets of the kinase or phosphatase, although further work is needed to ensure that differential abundance is not due to an overall change in cellular state. Perhaps the most striking results that have emerged from these efforts come from Bodenmiller *et al.* (2010), who observed that indirect effects of kinase knockouts often outnumbered the direct effects.[3] The authors suggest that this might reflect functional redundancy, which keeps signaling stable in a changing environment, but also emphasize that neutral evolution may lead to many off-target hits with low functionality.

The availability of commercial kinase inhibitors can greatly aid in producing short time point perturbation experiments.[22] This approach led to successful characterization of the PI3K/Akt/mTOR and MEK/ERK signaling axes in a cancer cell line by Wilkes *et al.* (2015).[4] Its notable that the authors were able to show that chronic application of kinase inhibitors potentially induces dramatic changes in network wiring, which may suggest that these experiments are more likely to produce accurate network models than kinase knockouts. By systematically applying inhibitors to all the nodes in a particular signaling axes, and combining data with a known protein-protein interaction network, the authors were able to produce a final boolean network representation of the underlying kinase-substrate relationships, which modeled the network as an interconnected series of on/off nodes.[17,23] This shows the particular benefit that integrating data from previous experiments can have in understanding the dynamics of phosphoproteomics data.

1.1.2 Addressing statistical challenges in large datasets

As the scope of phosphoproteomics experiments continues to increase, questions about the quantitative relationship between phosphosites will become more intricate. Inevitably, fundamental issues which plague all proteomics datasets, such as missing data and batch effects, will increase alongside dataset size.[24,25] Thus, care must be taken to statistically address these challenges when modeling cellular systems, so that biological signals can be differentiated from noise. Given the potential large upfront investment in sample acquisition time for assay panels, it is important to make sure that sufficient time is given to detailing which experimental steps will introduce confounding variables and that the chosen experimental design allows these confounders to be appropriately modeled during statistical analysis.

A major limitation of discovery mass spectrometry methods such as DDA and DIA is the variance in detected phosphopeptides from run to run.[24] Thus, researchers must make a choice on how to handle missingness within their dataset, keeping in mind the tradeoff between number of sites which can be analyzed and the real information content of the included sites. The most straightforward approach is to first set a defined threshold for allowing a phosphorylation event to be included in an analysis. This could be a threshold on the percentage of samples containing a given analyte, the percentage of treatments the analyte is detected in, or a combination of the two. At this stage, one can choose between a number of imputation strategies for creating a full matrix.[26] Depending on the total amount of missing data and the proposed reason for its absence, some imputation strategies may be more appropriate than others. Often, individuals will choose an imputation strategy which explicitly encodes the hypothesis

that analytes below the limit of detection tend to dominate missing data, which is a form of missing data termed missing-not-at-random.[27] Once downstream analysis is complete, it may be a worthwhile investment to evaluate how dependent conclusions are on choice in imputation strategy.

Aggregation of quantitative data into higher level representations, such as peptides into proteins or phosphopeptides into phosphosites, can often lead to statistics with a more complete representation across samples. Phosphoproteomics allows for a particularly interesting version of this technique for phosphosites that can be mapped to known kinases. This technique, termed kinase set enrichment analysis, looks at the log fold changes among all target phosphosites of a kinase in a given treatment to define an activity for the individual kinase against a control treatment.[28,29] The list of underlying kinase targets detected in each analyzed treatment can vary dramatically, allowing the comparison of widely different cellular systems. The main requirement is a confident list of kinase-substrate relationships, which for humans can be downloaded from the Phosphositeplus online resource,[30] but for other organisms may be difficult to curate.

An important aspect of large scale datasets which must be planned and carefully controlled at every step in the sample preparation, acquisition, and statistical analysis pipeline is the effect of confounders.[25,31] Any change between a block of samples, such as plate for phosphopeptide enrichment or liquid chromatography column, has the potential to introduce batch specific biases into analyte intensities. Thus, it is ideal to maintain as close to a balanced design as possible, with every treatment or tissue represented in each batch. This setup has the best potential to mitigate any confusion between true treatment effects and underlying confounders.

The choice in how to mitigate batch effects within the data will largely depend on the analysis a researcher wants to perform. Techniques which estimate batch effects and produce a final dataset with the effects corrected out, such as ComBat,[32] are ideal when subsequent analysis steps don't allow explicit modeling of batch variables. Evaluating whether batch effects have been effectively removed can be done visually or statistically.[25] The former can be performed through principal component analysis colored with the individual batch labels before and after batch correction. Statistical analysis of batch effect removal is possible using principal variance component analysis, which can model the relative proportion of signal coming from treatment variables and batch variables before and after correction.[33] Other statistics, such as the median coefficient of variation between samples from the same treatment, should also substantially improve with batch correction.

When the goal of analysis is to use linear modeling to analyze differential expression either between individual treatments or between each treatment and a control, it is often more desirable to include batch variables directly during modeling. In the case of a partially unbalanced design, when not all treatments are distributed equally across batches, this is also strictly necessary due to the inflation of the test statistic during modeling.[34] If confounding variables are known, the design matrix is directly appended to the design matrix of the treatments, and linear modeling can then proceed as normal. However, some authors also choose to learn confounding variables directly from the data, using techniques such as surrogate variable analysis to directly model the remaining variation in the data after treatment variables have been corrected out.[35] These techniques can be powerful, as they have the potential to recognize

underlying confounding variables even if they were not originally recorded. However, their benefit will be severely diminished if confounders correlate with treatment variables.

1.2 Compiling knowledge across datasets

Even with the increase in straight-forward sample preparation and analysis techniques, building large scale phosphoproteomics datasets is often outside the reach of many laboratories. Alternatively, qualitative and quantitative information about phosphoproteomes is abundant in publications and public repositories.[8] Careful curation of this information can provide deep insights, potentially equivalent to months of mass spectrometry time.

1.2.1 Efforts to compile public data into central repositories

As stated above, there is a high diversity in detected sites between datasets in discovery phosphoproteomics. While this can be a source of frustration for quantitative phosphoproteomics, it can provide a deep catalog of potential phosphosites if reported detections are combined between studies.[36] Perhaps the most famous example of this is Phosphositeplus, whose repository represents a core database of phosphosites reported in individual phosphoproteomic publications.[7] This expert curation has facilitated the combined analysis of a large number of confident sites, especially when the data is augmented with known kinase substrate relationships.[30,37] Other groups have taken a similar approach, but focusing on curating the quantitative data reported in the literature.[38]

While datasets curated from publications can provide high coverage catalogs of signaling events, these often suffer from the accumulation of false positive hits.[10] This occurs due to the compounding of small but additive errors in phosphopeptide detections and phosphosite localizations in individual publications. Several attempts have been made to deal with this, however. In recent years, the concept of dataset reanalysis has become increasingly popular, i.e. the search and scoring of data from individual mass spectrometry datasets through a common pipeline.[39] This allows an analysis to control phosphopeptide and phosphosite discovery globally across the database, and control FDR at definable levels. Recognizing this same limitation in its own data, Phosphositeplus began including a core set of re-analyzed data in 2014.[30]

There have been notable attempts to build large scale reanalyses of phosphoproteomics data, either to build information to inform new phosphoproteomics assays or characterize the phosphoproteome in general. Our lab was one of the first to address this challenge with Phosphopedia, which reanalyzed nearly 1000 phosphoproteomic mass spectrometry runs in order aid the construction of targeted phosphopeptide assays.[10] Users of the accompanying resource were able to look up phosphopeptides of interest and map out the likely elution times and most abundant charge state for each, severely reducing the up-front cost to build PRM runs. Other work of interest was performed by Ochoa *et al*. (2020) who used the Maxquant proteomics software to reanalyze thousands of individual phosphoproteomics runs and produce a final list of high confidence phosphosites.[9,40] These sites were then able to be fed into a machine learning algorithm along with evolutionary properties about sites and features

describing protein sequence and secondary structure in order to produce a final predictor for the functionality of a site.

1.2.2 Computational challenges in producing large scale re-analyses

Although it is true that individuals save themselves a great deal of instrument time and resources by performing re-analysis of public datasets, much of the investment is shifted to production scale computational resources. Furthermore, there are unique challenges that arise as more data is consumed, especially if individuals want to avoid the accumulation of errors in the final product. In fact, the same principle holds true both when building large scale experiments, and when building large analyses – time invested up-front to set up an analysis correctly will be paid back as the project progresses.

The first major hurdle which individuals must overcome is choosing a pipeline capable of performing the desired re-analysis on available hardware. For example, the use of MaxQuant in Ochoa *et al.* (2020) required a custom server with hundreds of gigabytes of RAM, and the total computational time was in the hundreds of hours.[9] While cloud servers, such as those provided by Amazon Web Services, Google Cloud, of Microsoft Azure, make it easy to build a virtual machine that can handle this kind of analysis, choosing a monolithic analysis software can hinder flexibility if parameters need to be changed or if there is a desire to integrate more data. An alternative approach is to employ modern workflow management systems such as Snakemake or Nextflow to build reproducible data pipelines.[41,42] These technologies handle file and job management for the user, allowing them to define pipelines which perform individual analysis tasks in parallel when possible and only rerun if absolutely necessary. At

completion, if a user wants to integrate more data into the analysis, many of the search and scoring steps may only need to be run on that new set of data, with a global integration step performed at the end.

This final integration step is where analysts often have to put the most amount of planning, as datasets which are made up of runs spanning many instruments and labs can be extremely heterogeneous. The first decision that must be made is how phosphopeptide and phosphosite false discovery rates will be controlled across the database. One solution is to allow scoring software such as Percolator or Mokapot to calculate scores for all PSMs using a common model across the database.[43,44] However, this may not necessarily address the intrinsic heterogeneity of the underlying data, especially if data comes from multiple instruments. It is possible that a grouped analysis, which takes into account variables such as instrument type and datasets may be fruitful, but more work needs to be done on the theoretical aspects of this analysis to guide researchers.[45] One particular aspect of false discovery rate estimation which has received only a small amount of attention is the accumulation of errors in modification sites. For many algorithms, individuals can receive a score per PTM per spectrum, and set a global threshold for localization across the dataset.[46,47] However, it is unclear how this should be combined with underlying PSM scores to define a global false discovery rate at the modification level. Past heuristics have worked well, such as that used in Lawrence *et al.* (2016).[10] Here the authors set a global localization score cutoff and collapsed unlocalized sites into localized sites if they could be explained by their more localized counterparts. Then, FDR was calculated by target–decoy competition at this level similarly to phosphopeptides, by taking the maximum peptide score corresponding

to each phosphosite. Even if heuristic techniques such as these control the true false discovery rate at a level higher than specified by the user, they can still produce a useful repository of information that can be refined with future experimentation by users.

1.3 Organization of Dissertation

Following the introduction chapter (Chapter 1), the thesis is composed of two major case studies for dealing with large scale phosphoproteomics datasets. First I detail the analysis of a large scale phosphoproteomics perturbation assay in *S. cerevisiae* (Chapter 2), providing an overview of the unique challenges of mass spectrometry datasets with hundreds of samples. Then, I provide an in depth analysis of my work to build Phosphopedia 2.0, a centralized online resource for mass spectrometrists to build targeted phosphoproteomics assays for any phosphopeptide. This section begins with an overview of pyAscore, a python package which provides fast and versatile post-translational modification scoring (Chapter 3). The next two chapters showcase the production of Phosphopedia 2.0 and its automated pipeline (Chapter 4), and my work to extend the web resource beyond phosphopeptide detections with deep learning (Chapter 5). Finally, Chapter 6 will provide concluding remarks and implications for this thesis on the study of phosphoproteomics.

Chapter 2. A large-scale phosphoproteomics perturbation screen in budding yeast.

Author contributions: This chapter was a joint collaborative effort between Anthony Barente and Mario Leutert. Mario Leutert designed and performed all experiments and data collection and performed the initial processing of DDA and DIA phosphoproteomic data. Mario Leutert also curated kinase targets from the literature. Anthony Barente performed data quality control, filtering, and batch correction. Methods sections on experimental procedures used to generate yeast data were written by Mario Leutert. All other analyses and discussions of results presented in this chapter are the work of Anthony Barente.

2.1 Abstract

S. cerevisiae is a powerful system for the study of cellular signaling under diverse biological, chemical, and physical stresses. Often studies of this system analyze small sets of perturbations in isolation. Comparing the effects of the treatments across studies is often difficult, due to study specific confounders. In contrast, scaling up these studies to include more stresses allows analysis of the similarities and differences among treatments, as well as dissection of the underlying phospho-regulatory modules. Scale also brings challenges in the form of stochastic detection and confounding variables, which must be carefully handled. In this study we present the Yeast Phosphoproteome Atlas, one of the largest phosphoproteomics perturbation screens in yeast to date. We show how this data can be used to answer important questions about the underlying

phosphoproteome, such as the reproducibility of regulation and how phosphosites can be broken into modules of regulation.

2.2 Introduction

The budding yeast *S. cerevisiae* occupies an important position in modern cell signaling research. Many of the main components of the human regulatory architecture have direct analogues in yeast, the extent of which is reviewed in Mohammadi *et al.* (2015),[48] and although yeast lack much of the diversity in tissue specific expression that can be found in multicellular organisms,[49] their simplicity has made them attractive models. In fact, the yeast genome only encodes 161 kinases and phosphatases compared to the 665 in humans, which has led to systematic attempts to dissect yeast's underlying network architecture.[3,15] A main theme of phosphoproteomics research has been to show just how widespread and interconnected phosphorylation of proteins is, but effort is still needed to describe the regulatory relationships between these sites in the diverse biological contexts that the organisms may experience.[5,50]

The simple laboratory manipulation of yeast makes measuring the consequences of biological stresses on the phosphoproteome relatively easy. Much work has gone into uncovering the unique phosphoproteome responses to conditions such as high osmotic conditions,[51] heat and cold shock,[52] and endoplasmic reticulum stress.[53] Some of this work has even provided insight into the promiscuity of kinase-target relationships and intricate temporal profiles of phosphosignalling.[54,55] Research in human cell lines has indicated that remodeling may be common in cell lines which are subjected to chronic kinase inhibitors,[4] which suggests that short stresses may be ideal for determining the

true regulatory architecture of wild type cells. With the improvement of automated sample preparation workflows, we are now at a point where the signaling consequences of a diverse panel of stresses can be studied all at once.

Here we present the Yeast Phosphoproteome Atlas, an effort to study yeast phosphoproteomics under a diverse range of biological, chemical and physical conditions. By analyzing these treatments together and removing between-study heterogeneity, we have unprecedented power to understand phosphorylation dynamics across treatments. This allows us to address questions about both the relationship between phosphosites, as well as between the treatments themselves. The data and analyses presented here are a valuable resource for the generation of future questions in the field of yeast phosphoproteomics.

2.3 Methods

2.3.1 Treatment of cells and preparation

Single colonies of the S288C-derivative BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) budding yeast strain were picked from a fresh plate and grown overnight, shaking at 30° C in Synthetic Complete media with 2% Glucose (SC+C). Next day 50ml shake flask cultures were inoculated in SC+C at an OD600 of 0.1 and grown for ~5h until they reached an OD600 of 0.6. Treatments were performed on these cultures and for treatments that required complete media exchange yeast were filter purified and filters were incubated in prewarmed new media. Each culture was treated as an independent replicate and treatments were performed shaking at 30° C for exactly 5 min (or as described for specific treatments). A description of exact treatment conditions can

be found in Table 2.1. To stop treatments, cells were metabolically arrested and harvested by adding 100% (w/v) Trichloroacetic acid (TCA) directly to the liquid culture to a final concentration of 10% TCA, incubated on ice for 10 min centrifuged, decanted, washed once with 100% acetone at -20° C, centrifuged, decanted, followed by a wash with ice-cold water, centrifuged and decanted. Cell pellets were snap-frozen in liquid nitrogen, and stored at −80°C until lysis.

Frozen cell pellets were resuspended in a lysis buffer composed of 8 M urea, 75 mM NaCl, and 50 mM HEPES pH 8. Cells were lysed by 4 cycles of bead beating (30-s beating, 1-min rest on ice) with zirconia/silica beads followed by clarification with centrifugation. Protein concentration of every lysate was measured by BCA assay and lysates adjusted to 1 mg protein per ml lysis buffer. Proteins were reduced with 5 mM dithiothreitol (DTT) for 30 min at 55°C and alkylated with 15 mM iodoacetamide in the dark for 15 min at room temperature. The alkylation reaction was quenched by incubating with additional 10 mM DTT for 15 min at room temperature. Lysates were stored at -80°C until further processing.

Lysates were scrambled across 96-deep well plates and biological replicates were blocked from being on the same 96-well plate. Each plate contained 4 samples containing the same pooled lysate to assess sample preparation reproducibility between 96-well plates. To purify proteins and perform digestions, 96-well plates were processed using the R2-P1 (Rapid-Robotic proteomics) protocol implemented on a KingFisher™ Flex (Thermo) magnetic particle processing robot as established by our group before in Leutert *et al.* (2021).[56]

2.3.2 Deep phosphoproteome sample preparation

For the deep phosphoproteome a pooled sample (only 5 min treatments) was created by combining equal amounts reduced and alkylated lysates. In-solution digestion were performed on lysates containing 5 mg of protein using either Trypsin (Promega), Chymotrypsin (Promega), Glu-C (Promega) or Lys-C (Wako chemicals) according to manufacturer protocols and peptides were desalted on C18 SepPak cartridges. Phosphopeptides were enriched using R2-P2 as described above. Offline pentafluorophenyl reverse-phase (PFP) chromatography was performed on phosphopeptides derived from the different digests individually using a XSelect HSS PFP 200 × 3.0 mm; 3.5 µm column (Waters) as described in Grassetti *et al.* (2017),[57] fractions were combined into 12 pooled fractions and lyophilized.

Additionally, strong cation exchange (SCX) chromatography was performed on tryptic digest using a polysulfoethyl A, 200 × 4.6mm; 5µm, 300A column (PolyLC) and two buffers: (A)10mM Ammonium formate, 0.05% formic acid, 25% ACN and (B) 500mM Ammonium formate, 0.05% formic acid in 25% ACN. Peptides were fractionated with a gradient ranging from 5% Buffer B to 100% Buffer B, 12 fractions were collected, lyophilized and phosphopeptides enriched with R2-P2.[58]

2.3.3 Mass spectrometry data analysis

The S. cerevisiae S288C reference protein fasta database containing the translations of all 6713 systematically named ORFs, except "Dubious" ORFs and pseudogenes created on 05/11/2015 by SGD (https://www.yeastgenome.org/) was used for all searches.

DDA data was searched with Comet (2019.01.2).[59] The precursor mass tolerance was set to 20 ppm. Constant modification of cysteine carbamidomethylation (57.021463 Da) and variable modification of methionine oxidation (15.994914 Da) were used for all searches, and additional variable modification of serine, threonine, and tyrosine phosphorylation (79.966331 Da) was used for phosphopeptide samples. Search results were filtered to a 1% FDR at PSM level using Percolator.[60] Phosphorylation sites were localized using pyAscore, an in-house implementation of the Ascore algorithm.[61] Phosphorylation sites with an Ascore > 13 (p-value < 0.05) were considered confidently localized.

For spectral library generation and spectral library searches of DIA data Spectronaut v.15 (Biognosys) was used. A phosphopeptide spectral library was generated by searching all DDA data encompassing the deep yeast phosphoproteome coming from tryptic digests together with DIA data encompassing the quantitative phosphoproteomic measurements. Standard search parameters were used, including fixed modification of cysteine carbamidomethylation and variable modification of methionine oxidation and serine, threonine, and tyrosine phosphorylation. A PSM and peptide FDR cutoff of < 0.01 and a PTM localization site confidence score cutoff of > 0.75 were chosen.

For the spectral library searches of the phosphoproteomic DIA data standard spectral library search setting with following adjustment were chosen: decoy limit strategy was set to dynamic with a library size fraction of 0.1, but not less than 5000, a precursor FDR cutoff of < 0.01 was enforced by choosing the data filtering setting “Qvalue”, no imputing or cross run normalization was performed, a PTM localization site

confidence score cutoff of > 0.75 was chosen, multiplicity was set to false, and PTM consolidation was done by summing. For computer memory reasons raw files were searched in batches of 100 files and combined all together using the “SNE combine workflow” in Spectronaut to merge the identification results of individual batches in a FDR controlled manner.

Spectral library searches of the total proteomic DIA data were done in similar manner as the phosphoproteomic searches, with a total proteome spectral library.

2.3.4 Analysis of differential phosphorylation and site co-regulation

Quantitative phosphoproteomics analysis was performed at the phosphosite level across the data by summing peptide quantification values and median normalization across individual samples. As described in the text, a stringent filter for sites which were present in at least 50% of samples was applied, and remaining missing values were imputed using the QRILC algorithm implementation within the imputeLCMD R package.

For differential phosphorylation analysis, surrogate variable analysis was performed on all samples using the SVA R package, with the number of significant surrogate variables automatically determined.[62] Differential expression was determined using LIMMA on all samples at once, with a model matrix including both the treatment for each sample and the individual surrogate variables to correct batch information.[63] Significant differential expression was then calculated for each treatment against the untreated cells, and p-values were corrected globally using Benjamini-Hochberg correction.[64] Finally, a site was called as significantly regulated in a treatment if its absolute log2 fold change against untreated cells was at least 1 and its adjusted p-value was less than 0.01.

For site by site co-regulation analysis, a second dataset was created from the filtered data with R2-P2 batch effects directly corrected out using the ComBat method from the SVA package.[32] This data was intentionally not used for differential expression analysis due to the undesirable effect that an unbalanced design can have on false discovery rate.[34] Weighted gene correlation network analysis (WGCNA) was performed using the WGCNA package, with a soft threshold chosen as the first power which resulted in a scale free model R^2 passing 0.8, as recommended by the method's authors.[16] The WGCNA method produces a final topological overlap matrix which is used to call modules through dynamic tree clustering.[65]

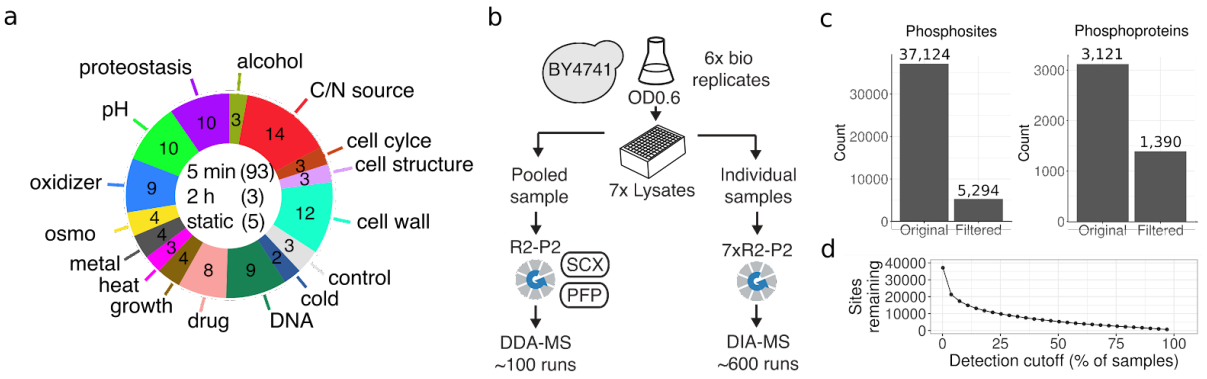


Figure 2.1. Description of the yeast phosphoproteome atlas dataset. a) Number of individual treatments included in the original dataset and their distribution of treatment types. For the current chapter, we focused on treatments which were applied for 5 minutes, to minimize changes to the underlying proteome. **b)** A graphical description of the experimental design used to generate the yeast phosphoproteome atlas with generation of a deep phosphopeptide spectral library via DDA mass spectrometry shown on the left and the generation of quantitative samples via DIA mass mass spectrometry on the right. **c)** Final number of phosphosites and phosphoproteins represented in the filtered dataset and analyzed in this chapter. **d)** Number of sites remaining in the dataset after application of increasingly stringent cutoffs on the percent of all samples that a site was detected in.

2.4 Results

2.4.1 Overview of the Yeast Phosphoproteome Atlas

In order to study the dynamics of the yeast phosphoproteome in diverse biological contexts, a large panel of 101 perturbations was generated across several environmental, chemical, and biological classes (Fig. 2.1a; Table 2.1). Most of these perturbations were applied for 5 minutes, and several were included which either analyzed cells at a long time point (2 hours) or applied a static stress. The scale of this effort necessitated uniform handling of samples to limit technical noise, and thus R2-P2 robotic sample preparation was used to perform phosphoproteome fractions (Fig. 2.1b).[11] Since the R2-P2 method allows up to 96 samples per batch, it was not possible to include every sample in a single enrichment. Instead, 4-6 biological replicates per treatment were distributed across 7 R2-P2 enrichment batches, with at most one biological replicate included in each (Fig. S2.1a). In addition, each R2-P2 batch included several untreated control biological replicates, guaranteeing that each batch could contribute to differential expression analysis. After enrichment, quantitative samples were measured using high resolution DIA mass spectrometry with the goal of producing high phosphosite coverage across the dataset. At 558 samples, the generated Yeast Phosphoproteome Atlas constitutes one the largest DIA phosphoproteomic perturbation panels to date.

A deep spectral library is a powerful tool for analyzing DIA experiments,[66] and thus the included 5 minute treatments were pooled and subjected to digestion by multiple proteases and deep fractionated for analysis by DDA mass spectrometry (Fig. 2.1b). This produced a high quality library of 36,405 high-confidence phosphosites at a

PSM FDR of 5% and localization probabilities of >95%, which enabled the quantification of 37,124 phosphosites on 5,294 proteins across the DIA samples (Fig. 2.1c).

One of the main challenges with a dataset of this size is missing values, which may occur when there is insufficient evidence for a peptide in a sample due to low abundance, mislocalization, or other processes affecting signal quality. Across the dataset, the number of samples in which a phosphosite was detected dropped off exponentially, with the fraction of samples missing a quantification for a site being dependent on the median intensity of that site across the dataset (Fig 2.1d; Fig S2.1b). Since we were interested in analyzing the global behavior of phosphites across treatments, we wanted to find a compromise between the number of included sites in the analysis and the amount of global quantitative information they contained. Thus, we chose to filter for phosphosites which were quantified in at least 50% of all DIA samples, which gave a final dataset of 5,294 phosphosites distributed across 1,390 phosphoproteins (Fig. 2.1c).

At this point, we had an incomplete dataset, with a global 33.5% of values missing. Given that individual batches varied dramatically in their median number of detected phosphosites per sample, the percent of values missing also varied from 39.3% in batch 1 to 10.4% in batch 6 (Fig S2.1cd). Informed by the dependence of phosphosite missingness on intensity, we chose to impute the remaining values using the QRILC algorithm.[67] This algorithm explicitly assumes missing data was generated from the lower tail of a sample's intensity distribution and the effect of the algorithm is apparent in the downward shift of each batch's intensity distribution (Fig. S2.1e).

2.4.2 Modeling differential phosphorylation across batches

We wanted to understand how individual treatments affected the relative phosphorylation of sites across the proteome, so we turned to linear modeling via the LIMMA R package.[63] As discussed above, the individual biological replicates for each treatment are distributed across 7 R2-P2 batches. A principle component analysis of this data revealed that the dataset experiences a great deal of batch specific expression, with the group of batch variables able to explain a majority of the signal within the dataset according to principal variance component analysis (Fig. S2.2ab). Given this issue, it was important that we incorporated variables which capture confounders directly into the linear modeling.

A popular technique for capturing profiles of confounding effects is surrogate variable analysis.[62] Applying this analysis to our dataset revealed 4 distinct surrogate variables (SVs), which showed remarkably smooth batch and sub-batch trends even though the analysis ignores all sample ordering (Fig. S2.2c). After fitting an SV-only model to the data with LIMMA and subtracting out the SV signal, median within-treatment sample-sample correlation improved by a median of 25.02% across treatments, which suggests the SVs are capturing between batch heterogeneity (Fig. S2.2d). Furthermore, inclusion of the SVs alongside treatment variables in the LIMMA analysis increased the median explained variance from 24.28% to 61.44% (Fig. S2.2e).

With the full model, we used LIMMA to test all treatment-phosphosite pairs (529,400 total tests) for differential abundance against the untreated samples, and called significance at a 1% FDR and log2 fold change of 1 (Fig. S2.3a). Globally, this resulted in 3,749 phosphosites with at least one significant treatment-phosphosite pair

on 1,155 proteins (Fig. S2.3b). As discussed above, our dataset contains a number of treatments which apply a constant or long stress. There is a potential that these treatments may lead to a change in the underlying proteome, which can confound our analysis of phosphosite dynamics. Thus, the focus of this chapter will be only the 5 minute treatments, which resulted in 3,128 sites with at least 1 significant treatment-phosphosite pair on 997 proteins (Fig. S2.3b).

A striking result is that downregulation of phosphorylation outweighed upregulation for nearly every treatment with at least 5 regulated sites (Fig. S2.3c). In order to investigate the effect of data preprocessing on this phenomenon, we broke sites into four groups based on percent missing data, and we counted the proportion of sites in each group which showed more down regulation than up regulation. This revealed that the dominance of down regulation over upregulation was anti-correlated with missingness (Fig. S2.3d). One interpretation of this is that our stringent missing data filter enriches for ubiquitously phosphorylated sites with relatively high stoichiometry who rely on dephosphorylation to start or stop activity.

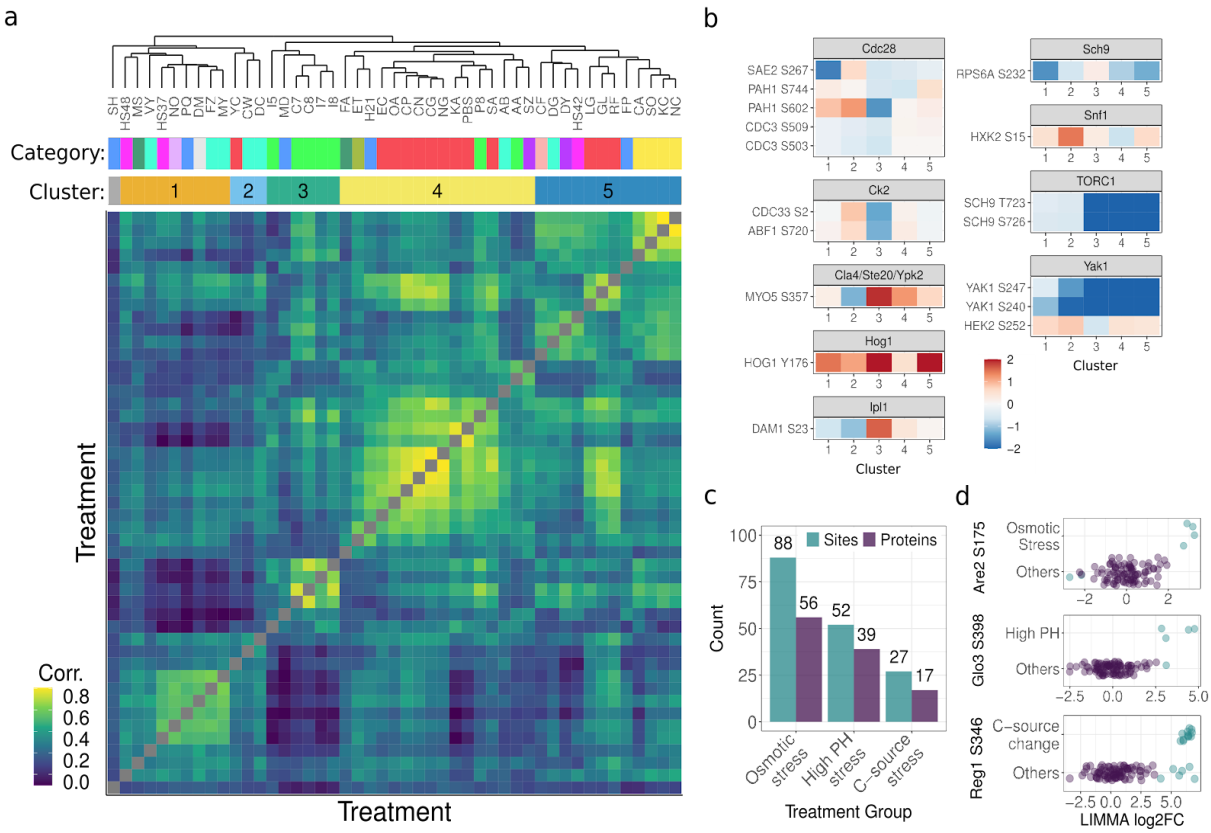


Figure 2.2. Analysis of treatment clustering and robust differential expression. a) Pearson correlation heatmap of log2 fold changes as determined by LIMMA between all 47 of the 5 minute treatments with at least 100 regulated sites using the 2996 sites which were regulated in at least 1 of the selected treatments. Rows and columns were arranged with hierarchical clustering and the resulting dendrogram is displayed with treatment category (colors defined in Fig. 2.1a) and clusters determined by cutting the dendrogram at a uniform height. **b)** Mean log2 fold change for sentinel sites within a cluster filtered to show sentinel groups with at least one site reaching a 1.5 mean log2 fold change. **c)** Number of phosphosites regulated in every treatment of sub-groupings and the number of proteins represented in those phosphosites. The 3 subgroupings that were formed based on the hierarchical clustering were osmotic stress (CA, SO, KC, NC), high pH stress (C7, C8, I7, I8), C-source change (EC, OA, CP, CN, CG, NG, KA, PBS, SA, LG, GL, RF). **d)** Representative examples of robustly expressed phosphosites in osmotic stress, high pH stress, and C-source change.

A tradeoff that is apparent with an exploratory dataset of this size is that the total amount of regulated sites per treatment can vary highly (Fig. S2.4a). Within our treatment panel, a total of 76 treatments had at least 10 differentially phosphorylated sites and 47 had at least 100 (Fig. S2.4b). Within treatments where the number of differentially regulated sites is low, it is difficult to determine whether the response to a

treatment is highly specific or if the treatment was not sufficient to activate underlying signaling pathways. Thus, we focus on treatments with at least 100 regulated sites to analyze the similarity between treatment effects. Using the 2,996 phosphosites which were differentially regulated in at least one of the remaining conditions, we calculated the Pearson correlation of log2 fold changes between the individual treatments and arranged treatments by hierarchical clustering (Fig 2.2a). It is noteworthy that the off-diagonal elements are nearly all positively correlated or uncorrelated, which suggests that, on average, when pathways are regulated they are regulated in the same way. As expected, a series of prominent blocks of related treatments are visible in the data, such as the osmotic stresses, pH stresses, and carbon source change stresses. It is particularly interesting that two distinct clusters of carbon source change stresses appear, with raffinose (RF), glycerol (GL), and low glucose (LG) clustering together away from the rest of the nutrient stresses.

Certain phosphosites, termed sentinels, can give insight into the dynamics of signaling pathways, and we wanted to determine whether these sites could provide insight into our sample clustering. We thus cut the dendrogram at a constant height to produce five main clusters, allowing us to average the log2 fold change for individual sites between treatments within a cluster (Fig. 2.2a). This revealed a stark down-regulation of TORC1 complex signaling, as represented by T723 and S726 on Sch9, in clusters 3, 4, and 5 which covered the majority of perturbations, including the pH, carbon source change, and high osmolarity stresses (Fig. 2.2b). A similar response was seen for all but cluster 3 for the Sch9p target RPS6A S232, which provides insight into cellular modulation of translation. Cluster 3 groups the high pH stresses and shows

interesting dynamics in several kinases. This cluster shows the highest phosphorylation increase on the Ipl1p target Dam1 S23, whose phosphorylation directly inhibits attachment of the kinetochore to the microtubules, as well as the Cla4p/Ste20p/Ypk2p target Myo5 S357, whose phosphorylation is required for polarization of the actin cytoskeleton. Cluster 3 also shows downregulation in Cdc2 S2 and Abf1 S720, targets of the Casein kinase Ck2p. Dephosphorylation of these sites signals a down regulation of translation and transcription respectively, hinting a general shutdown of RNA and protein production in cluster 3.

2.4.3 Robust differential phosphorylation in treatment groups

The clustering of similar treatments makes apparent a particular benefit provided by scaling up treatment categories. In many studies, probing cellular responses with a single treatment can provide a prioritized list of sites for further analysis. However, here we can also provide analysis of robustly perturbed sites showing regulation in all treatments of a given type. In order to show how we can apply these types of analyses to our dataset, we defined a series of broadly interesting treatment groups. The first two groups which include the osmotic stresses–NaCl, KCl, $CaCl_2$, and Sorbitol–and the high pH stresses–C7, C8, I7, I8–fully cluster together within the treatment heatmap (Fig. 2.2a). The next group of interest, which constituted stresses involving the exchange of cellular carbon source–YC, EC, OA, CP, CN, CG, NG, KA, PBS, SA, LG, GL, and RF–was split across three different clusters. For the current analysis, we decided to remove YC and combine the two largest subsets into the carbon source stress group, since their overlap was relatively high (Fig. S2.4c).

We first counted the total number of robustly regulated sites within each treatment group. This revealed that robust regulation was generally low across the defined treatment groups, with a total of 88 sites for the osmotic stresses and 52 sites for the high pH stresses out-measuring the 27 sites for the carbon source stresses (Fig. 2.2b). The low count and correlation with group size is expected with the stringency of requiring that a site be reproducibly regulated across multiple treatments. Interestingly, the total number of treatments in which the robustly regulated sites experienced differential phosphorylation was often relatively high, with a median of 18, 16, and 21 total treatments per phosphosite for the osmotic, high pH, and carbon source stresses respectively (Fig. S2.5d). One interpretation of this result is that analyses which seek to define sites which are regulated in multiple treatments enrich for sites which are core points in the underlying phosphorylation network. For further characterization of the treatment groups, we ordered phosphosites by total number of significant treatments, in order to focus on sites which were more unique to a stress type (Fig. S2.5abc).

Several sites of interest can be found in these protein lists. The sterol O-acyltransferase, Are2, plays a fundamental role in membrane maintenance in S. cerevisiae by catalyzing the esterification of sterols to fatty acids and sequestering them away from the plasma membrane.[68] A canonical HOG1 target on this protein, S175,[69] shows increased phosphorylation in all osmotic stressors, and only shows decreased phosphorylation in 2 other treatments, formamide (FA) and k-acetate media (KA) (Fig. 2.2d). Within the high pH group, S398 of Glo3, experienced robust regulation, with the only other significant treatment being tris-buffered alkali stress (P8), which was not included in the high pH group since it clustered with the carbon source stresses in Fig.

2.2a. Glo3 has an overlapping function with Gcs1, which together regulate retrograde transport from the golgi.[70,71] In contrast to the osmotic and high pH stresses, 23 out of 27 of the robustly regulated sites in the carbon stress treatment group are down-regulated, mirroring the overall down-regulation in the global data (Fig. S2.5c). Reg1 is the regulatory subunit of Glc7p, and is involved in the regulation of glucose repressible genes.[72] One of its sites, S346, shows robust upregulation in all of the carbon source change treatments, as well as several other stresses including two high pH treatment (C8 and I8) (Fig. 2.2d).

In sum, these data represent a flexible resource for individuals seeking to explore the consequences of stress application. Treatments can either be viewed individually or combined together to make refined lists of phosphosites for further study. In the next sections, we will look at another way to combine these data to learn about the underlying regulatory architecture.

2.4.4 Dynamics of HOG1 targets across treatments

Within yeast, kinase substrate relationships can be complex,[3] and we wanted to understand to what extent targets of a single kinase could be expected to be regulated together under a wide range of treatments. In order to pursue this question, we decided to build a final dataset which corrected out batch specific effects so that we could measure phosphosite co-regulation directly. Thus, we employed the ComBat batch correction method from the SVA R package to remove confounders and produce a final dataset which was free of most known batch effects (Fig. S2.6a). A PVCA analysis of the corrected data revealed that residual noise within the dataset still outweighed treatment variables in average signal, but the median within-treatment sample-sample

correlation improved by a median of 30.8% across treatments (Fig. S2.6bc). Finally, we averaged the expression level of each phosphosite within treatments, with the goal of minimizing residual variance as much as possible.

Recent work by Romanov *et al.* (2018) defined high quality HOG1 targets based on their dynamics under HOG1 induction and inhibition, as well as the phosphosite's correspondence to HOG1's preferred motif.[69] Within our filtered and batch corrected dataset, we were able to detect 10 of these sites which partitioned into 3 distinct groups after hierarchical clustering of the phosphosite by phosphosite Pearson correlations (Fig. 2.3a). Two of these groups show high within-cluster correlation and thus represent interesting sets for further investigation–cluster 1, which includes S602 of Rod1, T808 of Ask10, T750 of Gal11, and T264 of Gip3, and cluster 2, which includes Ser175 of ARE2, Ser476 of AIM21, Ser158 of ART5, and Ser898 of Reg1.

In order to investigate why these 2 clusters may separate, we went back to our LIMMA analysis to determine whether we could see any differences in differential phosphorylation. Since HOG1 is the main signaling component of the yeast osmotic stress response, we first looked to see whether the clusters showed different dynamics in the treatments, NC, KC, CA, and SO. Plotting the log2 fold change against untreated cells showed that cluster 1 experienced a lower induction under all osmotic stresses than cluster 2 (Fig. 2.3b). S602 of Rod1 in this analysis shows abnormally low log2 fold changes compared to the rest of the sites in cluster 1, and we thus removed this site before determining that cluster 1 experiences a 3.1 times lower induction under our osmotic stresses than cluster 2 ($p < 1e\text{-}4$, Paired Student's T-test). This explains why

cluster 2 shows differential phosphorylation for nearly all treatment-site pairs, whereas cluster 1 shows none (Fig. 2.3b).

We also searched previously published results to determine whether other groups have seen differences in the dynamics of sites within our clusters. Interestingly, a 1 minute time course analysis under similar osmotic conditions from Kanshin *et al.* (2015) includes data for phosphopeptides representing T808 of Ask10 and T750 of Gal11 from cluster 1 and S476 of Aim21 and S158 of Art5 from cluster 2.[54] While all of these sites were activated under osmotic stress, the authors determined that the phosphopeptides representing sites from cluster 1 experienced earlier induction than the phosphopeptides representing sites in cluster 2. While this does provide specific insight into the dynamics of these sites between the 1 and 5 minute timepoints, it does suggest that the phosphosites in cluster 1 and 2 may be positioned differently in the underlying phosphorylation network.

Within this study, we have the ability to describe the differences in signaling dynamics across distinct treatments, so we wanted to investigate what other treatments may be driving the underlying correlations. We thus calculated the z-score of the phosphosite log2 fold changes within each cluster per treatment and plotted the difference between clusters (Fig. 2.3c). This revealed two quadrants of interest, each of which showing lower z-scores for cluster 1 than cluster 2. The first quadrant, where cluster 2 z-scores are greater than 2, includes the set of osmotic stresses as well as 1 pH stress and 1 heat stress, which are known to have some overlap with the osmotic stress pathways. The other quadrant, with cluster 1 z-scores less than -2, includes several nutrient stresses, including no glucose (NG), PBS media exchange (PBS),

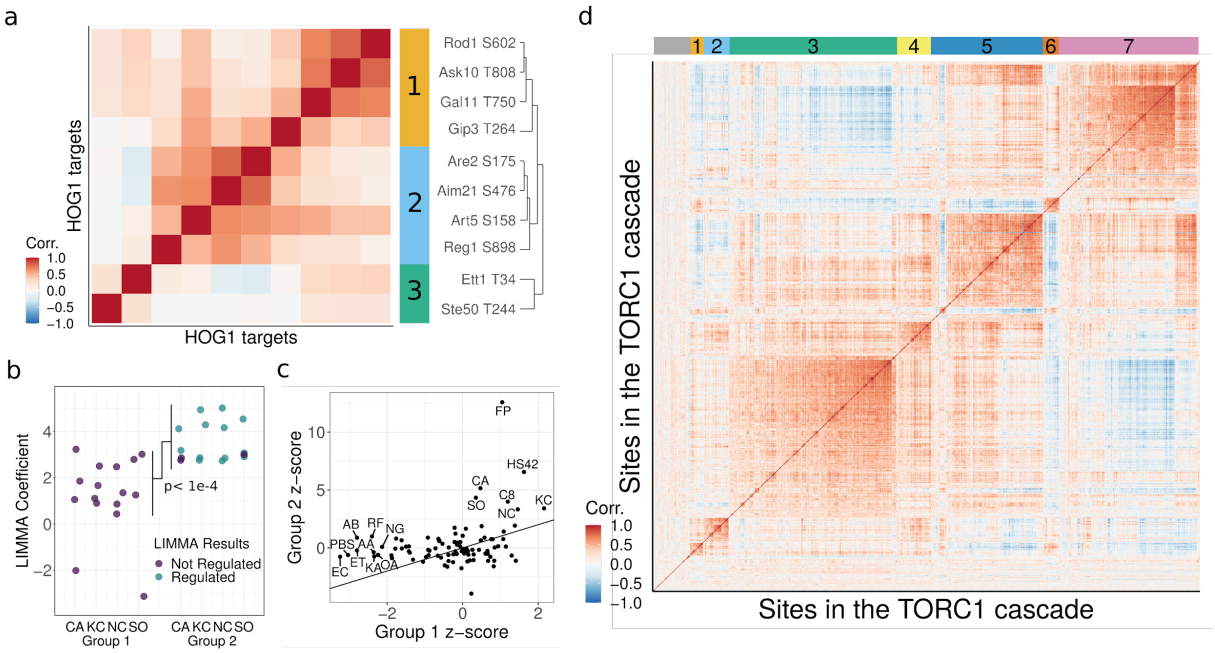


Figure 2.3. Co-regulation of downstream kinase targets. a) Hierarchical clustering of the Pearson correlation between sites defined as direct targets of HOG1. Expression was averaged within a given treatment before calculating Pearson correlation. **b)** Log2 fold change for each osmotic stress against untreated samples (LIMMA Coefficient) for the sites in groups 1 and 2 from (a). Points represent the LIMMA Coefficient for one of the sites in one treatment. Significance of the difference between 1 and 2 is the p-value for the group variable in a blocked linear model calculated without Rod1. **c)** Aggregate relative activity of group 1 sites vs group 2 sites for all 5 min treatments with treatments crossing an absolute value of 2 highlighted. Aggregate relative activity is defined as the z-score of the LIMMA coefficients for all sites in a group. **d)** Clustering of the 934 TORC1 cascade targets clustered and broken into 7 modules using Weighted Gene Correlation Network Analysis (WGCNA).

glucose to raffinose exchange (RF), and glucose to oleic acid carbon source exchange (OA). Interestingly, this lower quadrant includes both ethanol stress (EC and ET) as well as both stresses which introduce acetate (AA and KA).

2.4.5 Regulatory modules in the TORC1 cascade

Our previous analysis suggested to us that while defining kinase-target relationships is important, measuring how sites are regulated together across treatments may be more fruitful for understanding network organization under stress.

With this in mind we decided to investigate the modular organization of a larger pathway. We selected a series of kinases from the TORC1 cascade, as defined in Dokládal *et al.* (2021)[15]–TORC1, SCH9, SLT2, YAK1, RIM15, TAP42, NPR1, GCN2, and ATG1–and matched these with literature curated sets of downstream targets to produce a list of 934 total phosphosites with which to investigate regulatory dynamics. It is noteworthy that a majority of these sites are labeled as downstream of the TORC1 complex and ATG1, which are situated early in the cascade, and that several additional kinases beyond the original selected are labeled as upstream of final set of phosphosites (Fig S2.7a).

Given the number of phosphosites available to analyze, we wanted to pursue a powerful technique for determining co-regulation modules, and thus turned to Weighted Gene Correlation Network Analysis (WGCNA).[73] In order to build modules, the WGCNA R package requires a soft threshold to which the correlation based adjacency matrix will be raised. The underlying network is hypothesized to exhibit scale free topology, so we followed recommended practice to use the first power which created a scale free topology model fit with an $R^2 > .8$, which allowed for an optimal trade off between the hypothesized topology and connectivity (Fig. S2.7b).[16] The final result of the analysis was a partitioning of the TORC1 cascade targets into 7 modules, which varied in size from 23 phosphosites to 287 (Fig. 2.3d; Fig. S2.7c).

WGCNA allows each module to be transformed into a representative eigensite, defined as the first principle component of the sites belonging to the module.[16] A hierarchical clustering of the Pearson correlation of these eigensites showed correspondence between modules 1 and 6, between modules 5 and 7, and between

modules 2, 3, and 4 (Fig. S2.7d). Earlier, we showed that many treatments within our panel have low differential signaling, so we decided to focus again on the treatments which had at least 100 regulated phosphosites according to our LIMMA analysis. Measuring the relative induction of the eigensites against the untreated cells showed distinct dynamics according to clusters determined in section 2.4.2 (Fig. S2.8a). Module 1 from the WGCNA analysis showed most of its upregulation in the high pH treatments from treatment cluster 3, and behaves opposite to module 7, which is mostly downregulated in the same treatment cluster. Module 2 showed most of its upregulation within treatment cluster 5, which contains the osmotic stresses as well as one of the heat shock responses, and is visually similar to module 6 which also experiences upregulation in treatment cluster 5 in addition to treatment cluster 4. Modules 3-5 all show mostly down regulation, with module 3's partitioned mostly to treatment cluster 1 and modules 4 and 5 experiencing downregulation amongst the carbon source exchange treatments of treatment cluster 4.

We also wanted to get a first look into the underlying kinase regulatory landscape which may be governing our modules. After assigning the motif of each phosphosite within each module to either the acidic, basic, proline directed, or other classes, we looked to see if any enrichment for motif type were occurring (Fig. S2.8b). This hinted that the two smallest modules, modules 1 and 6, enrich for the basic or other phosphosite motif classes respectively. Module 6 is also most de-enriched for the proline directed class, while cluster 1 is most de-enriched for the acidic class. One of the largest modules, module 5, is de-enriched for the acidic class as well. Looking at the distribution of kinase targets across modules revealed that all kinases had targets split

between WGCNA modules (Fig. 2.8c). The kinases with the most representation, TORC1 and ATG1, had phosphosites partitioned across all the modules. Others, with downstream targets in either 2 or 3 modules, also spanned modules with distinctly different dynamics in the treatments with at least 100 regulated phosphosites, such as GCN2 and SLT2. It is likely that the kinases targets still represent signaling that is downstream of the individual kinases, but it is noteworthy that individual phosphosites show much more coregulation than can be directly explained by major kinases in the TORC1 pathway.

2.5 Discussion

In this study we presented a quantitative analysis of one of the largest yeast phosphoproteomics panels to date. The resolution provided by DIA mass spectrometry allowed us to assess the impact of nearly 100 treatments on 5,294 phosphosites on 1,390 phosphoproteins. Previous undertakings to study the yeast phosphoproteome's response to stress have often analyzed single classes of stresses,[51–53] which can make the comparison of treatments difficult if they were not conducted in the same experiments. Within our dataset, perturbation of cells from the same culture and careful parallel processing of treatments together in each enrichment batch allowed us to globally control confounders in modeling and compare distinct treatments. This revealed interesting similarities within treatment groups such as the generally high correlation across carbon source switch treatments, which likely represents a broad glucose depletion signaling.

We capitalized on inclusion of multiple treatments of similar type to define robust regulation, or phosphosites which experienced reproducible differential phosphorylation. By focusing only on the reproducible sites, we were able to create a core list of targets for future research. Several of these robustly regulated phosphosites occurred on proteins which are known to play key roles in high osmolarity, high pH, and carbon source stresses. Furthermore, many of the phosphosites experienced regulation in treatments outside the defined treatment group, which may mean they occupy a core position in the underlying network, or are targeted by multiple kinases.

Broad regulation is a common feature of the yeast phospho-regulatory network, which stems from the interconnectedness of kinase cascades.[3] While this likely allows for complex cellular logic,[74] it also means that directly studying the activity of kinases is difficult. In the past, individuals have been able to define a proxy for the activation of kinases through the dynamics of known substrates.[75] While this might be effective with highly specific kinase substrate relationships, we showed that even with quality target annotations for HOG1, sites could be partitioned into distinct subsets across treatments. When looking at a large set of phosphosites, such as that built off of the TORC1 cascade, phosphosites could be broken into broad modules which experience coherent dynamics upon stress. This does not imply that we must completely abandon the notion of a kinase-substrate relationship, but that we must build an understanding that fluidly moves between direct kinase action and modules of phosphosite regulation.

This study provides both a valuable resource to the phosphoproteomics community as well as case study in how to handle the challenges of a dataset of this size. Further studies with carefully selected perturbation panels have the potential to

greatly enhance our understanding of yeast stress signaling, as long as care is taken to address the limitations inherent to large-scale discovery proteomics datasets. For individuals across the cellular signaling field, this resource also provides a gold mine of new hypotheses which can be tested with targeted experiments.

Table 2.1. Description of 5 minute treatments analyzed in this chapter.						
Code	Description	Category		Code	Description	Category
AA	Acetic Acid Stress	pH		CS23	Cold Shock 23°C	Cold
AB	Amphotericin B	Cell Wall		CU	CuSO4	Metal
AF	Alpha Factor	Drug		CV	Canavanine	Proteostasis
AN	Anisomycin	DNA		CW	Calcofluor white	Cell wall
AY	Actinomycin D	DNA		CX	Cycloheximide	Proteostasis
AZ	Azetidine	Proteostasis		CY	Calyculin A	Drug
BM	Beta-mercaptoethanol	Proteostasis		CZ	Clotrimazole	Cell wall
BU	Butanol	Alcohol		DC	Deoxycholate	Cell wall
BY	Benomyl	Cell structure		DG	Digitonin	Cell wall
BZ	Bortezomib	Proteostasis		DI	Diamide	Oxidizer
C3	pH 3, 0.1M McIlvaine	pH		DM	DMSO	Control
C5	pH 5, 0.1M McIlvaine	pH		DT	Dithiothreitol	Proteostasis
C7	pH 7, 0.1M McIlvaine	pH		DY	Doxycycline hyclate	Proteostasis
C8	pH 8, 0.1M McIlvaine	pH		EA	EDTA	Cell Wall
CA	CaCl2	Osmo		EB	Ethidium Bromide	DNA
CF	Caffeine	Drug		EC	Low ethanol	C/N Source
CG	Galactose	C/N Source		ET	High ethanol	Alcohol
CI	A23187 Ca^{2+} ionophore	Cell wall		FA	Formamide	DNA
CM	Camptothecin	DNA		FE	FeSO4	Metal
CN	C/N depleted media	C/N Source		FP	FCCP	Oxidizer
CO	Cobalt	Metal		FY	Formaldehyde	DNA
CP	Pyruvate	C/N source		FZ	Fluconazole	Cell wall
CS10	Cold Shock 10°C	Cold		GL	Glycerol	C/N source
CS18	Cold Shock 18°C	Cold		H21	1mM H2O2	Oxidizer

Table 2.1. Description of 5 minute treatments analyzed in this chapter.						
Code	Description	Category		Code	Description	Category
H2O1	0.1mM H2O2	Oxidizer		NY	Nystatin	Cell wall
HS37	Heat Shock 37°C	Heat		OA	Oleic acid	C/N source
HS42	Heat Shock 42°C	Heat		OY	Oligomycin A	Oxidizer
HS48	Heat Shock 48°C	Heat		P8	Tris buffered - alkali	pH
HU	Hydroxyurea	DNA		PBS	PBS media	C/N source
I3	C3 + 2mM DNP	pH		PQ	Paraquat	Oxidizer
I5	C5 + 2mM DNP	pH		RF	Raffinose	C/N source
I7	C7 + 2mM DNP	pH		RN	Rotenone	Oxidizer
I8	C8 + 2mM DNP	Cell pH		RP	Rapamycin	Drug
KA	Potassium Acetate	C/N Source		SA	Sodium Acetate	C/N Source
KC	KCl	Osmo		SD	SDS	Cell wall
LA	Latrunculin A	Cell structure		SH	Sodium hypochlorite	Oxidizer
LG	Low glucose	C/N Source		SO	Sorbitol	Osmo
MC	Mitomycin C	DNA		ST	Staurosporine	Drug
MD	Menadione	Oxidizer		SZ	Sodium azide	Proteostasis
MN	MnCl2	Metal		TA	Trichostatin A	Drug
MS	MMS	DNA		TU	Tunicamycin	Proteostasis
MT	Methanol	Alcohol		UTF	Media exchange	Control
MY	Myriocin	Cell wall		VC	Verrucarin	Proteostasis
NC	NaCL	Osmo		VY	Valinomycin	Cell wall
NG	No glucose	C/N source		WT	Wortmannin	Drug
NN	No nitrogen	C/N source		YC	Change to YEPD	C/N source
NO	Nocodazole	Cell structure		ZC	Zenocin	DNA
NR	Nicotinamide riboside	Drug		ZN	ZnSo4	Metal

Chapter 3. A Python Package for the Localization of Protein Modifications in Mass Spectrometry Data

This chapter is adapted from the following work:

Anthony S. Barente and Judit Villén. A Python Package for the Localization of Protein Modifications in Mass Spectrometry Data. doi:10.1101/2022.04.04.487044[61]

3.1 Abstract

Determining the correct localization of post-translational modifications (PTMs) on peptides aids in interpreting their effect on protein function. While most algorithms for this task are available as standalone applications or incorporated into software suites, improving their versatility through access from popular scripting languages facilitates experimentation and incorporation into novel workflows. Here we describe pyAscore, an efficient and versatile implementation of the Ascore algorithm in Python for scoring the localization of user defined PTMs in data dependent mass spectrometry. pyAscore can be used from the command line or imported into Python scripts and accepts standard file formats from popular software tools used in bottom-up proteomics. Access to internal objects for scoring and working with modified peptides adds to the toolbox for working with PTMs in Python. pyAscore and is available as an open source package for Python 3.6+ on all major operating systems and can be found at pyascore.readthedocs.io.

3.2 Introduction

Post-translational modifications are fundamental for fine tuning protein function and can be studied at the proteome scale with mass spectrometry (MS). In a typical bottom-up proteomics experiment, whole protein extracts are proteolyzed and the peptides are measured by MS to obtain their mass and fragmentation pattern.[76] Software tools for protein sequence database search, such as Comet, perform well at matching MS/MS fragmentation spectra to peptide sequences, and identifying when a modification is present in the sequence.[59] However, they have poor sensitivity at identifying the precise site of modification when the peptide contains multiple acceptor residues.[77]

The Ascore algorithm was one of the first tools to explicitly score the confidence of PTM localization for a peptide-spectra pair.[46] It is able to provide a probabilistic score for localization confidence by scoring the presence of site-determining ions, i.e. fragment ions that report for the presence of a modification on the specified site. One advantage of this score is that it is calculated and can be consistently interpreted at the level of individual PSMs, and thus can be used in intelligent MS data acquisition workflows to inform subsequent MS events.[78]

The original implementation of Ascore focused solely on phosphorylation and was only available to the community via a web server which restricted submissions to 500 PSMs, a service which itself is now unavailable.[46] Similarly, a current open source implementation exists within the pyOpenMS software suite, but is limited to phosphopeptides analyzed with CID or HCD type fragmentation and requires the use of suite specific data structures.[79] There has been growing interest in the proteomics

community in analyzing other PTMs and current datasets report 10,000-100,000s spectra assigned to modified peptides. Thus, we reasoned it would benefit the community to expand the capabilities of the Ascore algorithm and make it more broadly accessible, so that it can be used as an alternative to or in conjunction with other open source PTM localization algorithms.[47] Here we present pyAscore, a fast and extensible open source implementation of the Ascore algorithm, which can handle a wide variety of modifications and MS/MS fragmentation modes.

3.3 Methods

Mass spectrometry data for individual experiments was downloaded directly from MassIVE or PRIDE, converted to mzML format with ThermoRawFileParser (v. 1.3.4) and then searched with the Comet database search software (v. 2021010).[59] Precursor tolerance was set to 20ppm for all searches and fragment tolerance was set to 0.02 Da for high resolution data and 1.0005 Da with 0.4 Da offset for low resolution data, as recommended by the Comet documentation. Human samples were searched with the UniProt *Homo sapiens* reference proteome with isoforms included (downloaded Feb 8, 2022). All synthetic peptide data was searched with a FASTA file that combined the human proteome with synthetic peptide sequences, which was defined in Marx *et al.* (2013) and located online in PXD000138.[80] All files were searched with carbamidomethylation on cysteines as static modification and up to 3 oxidized methionines as variable modifications. Since dataset PXD007145 reports TMT-labeled peptides, a TMT 10-plex modification was added as a static modification on lysines and peptide N-termini. For the synthetic peptide datasets, PXD000138 and PXD000759, and

the other phosphoproteomic datasets, PXD007740 and PXD007145, the modification of interest was phosphorylation on serine, threonine, and tyrosine, and this was included in the variable modification list. For dataset MSV000079068, acetylation of the protein N-terminus and internal lysines was included as a variable modification, but peptide C-terminal lysines were not allowed to be acetylated. Finally, HCD and CID data were searched with the b and y ion series, whereas ETD data was searched with the c and z+H ion series. All searches were subsequently grouped by dataset and fragmentation method and analyzed with Mokapot using default parameters.[44] Unless otherwise noted, for pyAscore localization, the fragment error was set to 0.05 Da and the same ion series as the searches was used. All searches and tests were performed on Intel Xeon Gold 6312U 2.4Ghz processors.

3.4 Results and Discussion

3.4.1 Overview of package and implementation

pyAscore provides an accessible interface to perform probabilistic scoring of PTM localization directly from the command line or incorporated into Python scripts (Fig. 3.1a). In both cases, our package relies on the pyteomics Python package to provide the ability to read from popular file formats containing mass spectra and peptide spectrum matches (PSMs),[81] so that scoring can be incorporated into a wide variety of workflows. For each input of MS/MS spectrum, peptide sequence, and a modification of interest, pyAscore will provide the overall best localization and an ambiguity score, which describes how much better the best localization is compared to the next best alternative. This procedure generalizes to any modification of interest just by specifying

the mass of the modification and the potential acceptor amino acids, e.g. 79.9663 Da and Ser, Thr, Tyr in the case of phosphorylation. Users can also customize analytical parameters to their specific MS acquisition method, with parameters for fragment ion mass tolerance, ion types to score, and presence of neutral losses.

The core algorithmic components of pyAscore are implemented in C++ and exposed to Python using the Cython package. This facilitates the production of fast and efficient code for scoring while allowing the flexibility of analyses provided by the Python programming language. The full scoring procedure and score definitions correspond to the original descriptions by Beausoleil et al.,[46] and we briefly describe the main scoring steps here. First, the MS/MS spectrum is binned into 100 m/z windows and the fragment ions within each bin are ranked. pyAscore then iterates over every possible localization of the PTM of interest on the peptide backbone and produces a PepScore, which is based on the number of matching theoretical peaks between a modified peptide sequence and the ranked set of fragment ions. The best scoring localization is then reported to the user. Peptides can potentially have a large number of permutations of unmodified and modified sites, so care was taken at this step to increase speed by reducing repeat fragment mass calculations and corresponding peak search.

Finally, for each modified site on the best localized peptide sequence, pyAscore calculates a score based on the number of site-determining ions for the best localization according to the Pepscore *vs.* the next best localization with that site unmodified. Like in the original implementation,[46] this score is termed the ‘ambiguity score’, or Ascore, and gives a probabilistic measure of how much better the best localization is than the next best possible modification site.

If more intricate analyses are desired, the components of pyAscore can be imported into Python scripts or any software that can link C++ libraries. Users then have direct access to the internal scoring class, which can be combined with spectrum preprocessing or further localization scoring (Fig. 3.1b). Since multiple scoring objects can be created with their own specific parameters, users also have the option to tailor score calculation to individual scan types. For users in need of tools for fast prototyping of new localization algorithms, pyAscore also provides access to its internal class for iterating over permutations of modified residues on peptides and calculating their individual theoretical fragment masses (Fig. 3.1c).

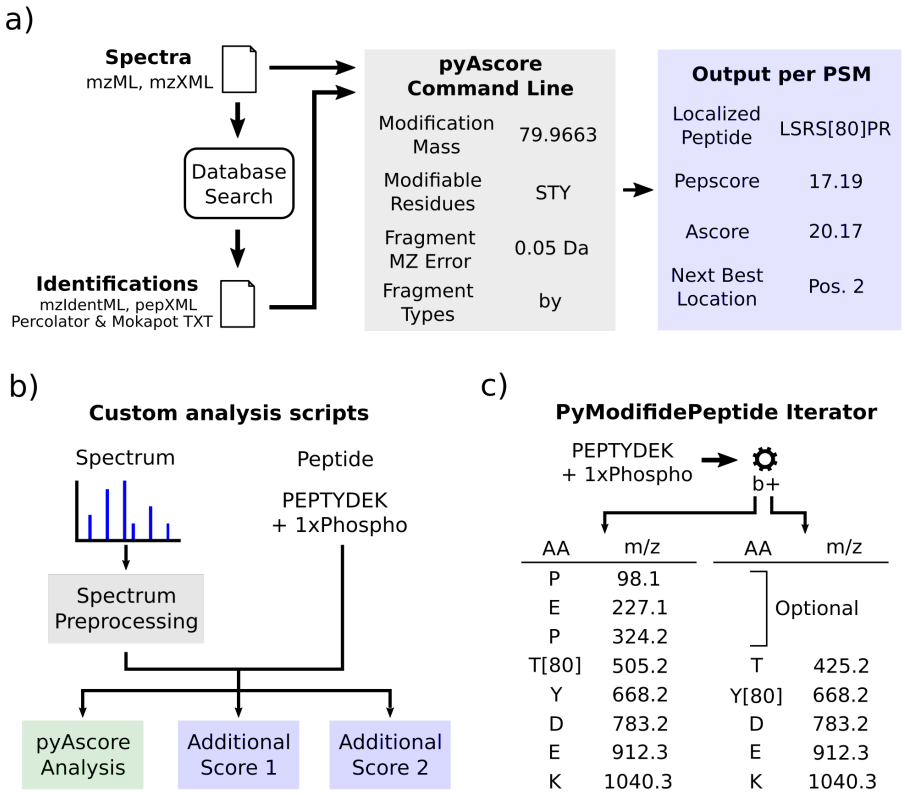


Figure 3.1. Outline of features included in the pyAscore python package. a) pyAscore can be used directly from the command line as a standalone application with a user specified modification and instrument parameters. The required input is a file containing spectra and a file containing peptide spectrum matches, both of which can be supplied via multiple popular mass spectrometry data formats. pyAscore outputs localization information per modified PSM in a tsv file. **b)** Individual scans and PSMs can be passed to pyAscore in Python scripts, allowing on the fly logic and the combination of multiple analyses per PSM. **c)** pyAscore provides Python use of its internal iterator pyModifiedPeptide, which allows users to efficiently step through theoretical fragment masses for all permutations of a peptide and a modification, with the b series ions shown here.

3.4.2 Validation of pyAscore on a dataset of synthetic phosphopeptides

We wanted to evaluate the performance of pyAscore's PTM localizations on data where the modification site of peptides was known. Thus, we turned to a large synthetic phosphopeptide library produced by Marx *et al.* (2013).[80] This library was analyzed across 2 studies with 3 MS acquisition approaches: high resolution HCD, high resolution

ETD, and low resolution CID.[80,82] The analysis of the same data with different fragmentation modes and resolution settings allowed us to compare how parameter choices in pyAscore affected performance, so we downloaded files from both repositories corresponding to the analysis of the original library.

After database search, we filtered for singly phosphorylated PSMs which were part of the synthetic library and obtained a total of 103,561, 153,088, and 13,854 PSMs in the high resolution HCD, low resolution CID, and high resolution ETD datasets respectively (Fig. 3.2a). While the number of PSMs was much lower in the ETD dataset than the others, there was still an ample number of PSMs to evaluate false localization rate (FLR). For PTM localization, we decided to evaluate the effect of tailoring pyAscore's parameters to instrument parameters during scores. Thus, for each dataset we tested two parameter settings for pyAscore's fragment ion mass tolerance. The wide tolerance, ±0.5 Da, corresponds to the tolerance used in the original Ascore paper, but is presumably too tolerant for high resolution data.[46] Therefore, we also tested a narrow tolerance setting, ±0.05 Da.

After scoring, we ordered PSMs by decreasing localization score, and evaluated the relationship between the score and the true FLR. For the two high resolution datasets, using the narrow mass tolerance drastically increases the number of PSMs at a given FLR, while the low resolution dataset performed better with the wide mass tolerance (Fig. 3.2b). This directly shows the benefit of tailoring the scoring parameters to acquisition parameters. It is notable that an Ascore cut off of 20 varied in FLR depending on dataset and parameter settings. When localization parameters did not match acquisition parameters, the FLR at an Ascore of 20 was 2.1%, 4.7%, and 0.7%

for the HCD, CID, and the ETD data respectively, while the FLR for tailored parameters was 3.6%, 2.2%, and 1.7% for the same datasets. However, taken together with the acetylation results presented above, these data suggest that an Ascore cutoff of at least 20 can still achieve a consistently low FLR if parameters are tailored correctly. In total, with tailored parameters, pyAscore scored 52.96%, 40.53%, and 38.28% of PSMs at a 1% FLR for the HCD, CID, and ETD respectively, and 69.38%, 53.45%, and 44.15% of PSMs at an Ascore cutoff of 20 for the same data (Fig. 3.2a).

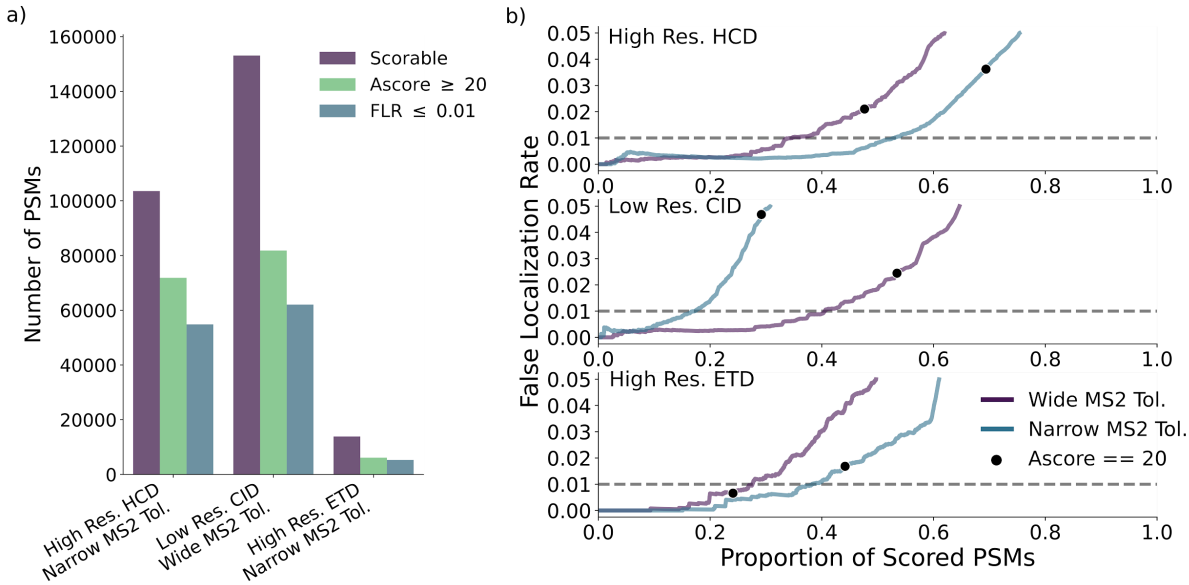


Figure 3.2. Validation of pyAscore on synthetic phosphopeptides. a) The number of PSMs which can be scored by pyAscore, the number of PSMs passing an Ascore cut off of 20, and the number of PSMs at a true FLR cutoff of 1% for the Marx *et al.* synthetic peptide dataset measured with 3 fragmentation methods. PSMs were considered scorable if the peptide came from the synthetic library, had a single phosphorylation event, and had at least 2 common phosphorylatable amino acids (i.e. Ser, Thr, Tyr). For each dataset, only the pyAscore results with parameter settings best matched to the instrument acquisition settings are shown, i.e. narrow tolerance settings for high resolution data and wide tolerance settings for low resolution data. **b)** False localization rate vs proportion of scored PSMs at decreasing Ascore cutoffs. Results are shown for pyAscore run with wide fragment mass tolerance (i.e. ±0.5 Da) and pyAscore run with narrow fragment mass tolerance (±0.05 Da). False localization rate (FLR) is calculated as the number of PSMs with Ascore passing at a given threshold where pyAscore called the wrong localization, divided by the total number of passing PSMs.

3.4.3 Speed of localization scoring

To evaluate pyAscore run times and usability in production settings, we reanalyzed three replicate high-resolution data-dependent acquisition (DDA) label-free phosphoproteomic runs from PXD007740.[83] After database search, pyAscore phosphosite localization was timed for each PSM.[44] Median localization times for individual replicates ranged from 0.061 ms to 0.073 ms per PSM with 99.7% of localizations taking 1 ms or less (Fig. 3.3a) and all PSMs in the run being scored in a matter of seconds. Plotting the localization time by the number of localization permutations in logarithmic scale revealed a linear correspondence. This is likely driven by the PepScore calculation, which must perform a calculation for each permutation of phosphorylated and not phosphorylated sites. Despite this, we found that the rate limiting step in the pyAscore calculations for a single file tended to be reading spectra and PSMs into memory.

3.4.4 Versatility of the pyAscore package

We wanted to showcase the versatility of pyAscore to extend to other modifications. Thus in addition to the data from PXD007740 we analyzed three replicate high-resolution DDA runs from both PXD007145,[84] which contained TMT labeled phosphopeptides, and MSV000079068,[85] which contain label-free acetylated peptides. For both PXD007740 and PXD007145, pyAscore was set to localize phosphorylations to serine, theonine, or tyrosine residues, and for MSV000079068 pyAscore was set to localize acetylations to any lysine in the peptide. An Ascore cutoff of 20 (theoretical confidence of 99%) has often been used in previous studies for determining localization,

and thus we considered any site which achieved this score as localized. For the label free phosphoproteomics data, pyAscore determined that an average of 58.8% of modified PSMs contained at least one localized site (Fig. 3.3b). The TMT-labeled phosphoproteomic data performed similarly, with an average of 57.1% of modified PSMs containing at least one localized site. On the acetylation dataset, pyAscore determined that an average of 87.7% of modified PSMs contain at least one localized site. This is likely due to the overall lower number of acceptor sites per peptide in the acetylation data than the phosphorylation data.

Our pyAscore analysis of the acetylation data (MSV000079068) was slightly less stringent than our database search, the latter of which did not allow acetylation on C-terminal lysines. This provided another look into the underlying FLR of pyAscore, as we could count any acetylations that pyAscore placed on the C-terminal lysines as incorrect. We thus filtered the PSMs of MSV000079068 for those that contained a single acetylation event, a C-terminal lysine, and at least one other internal lysine (N=14,608). We then sorted the PSMs by decreasing Ascore and counted the cumulative number of C-terminal acetylations as we moved down the list. This showed a direct correspondence between the Ascore and the rate of C-terminal acetylation, with the Ascore cutoff of 20 achieving a rate of 1.44% C-terminal acetylation (Fig. 3.3c). While this provides a great deal of confidence that pyAscore is performing well, the fact that these spectra were provided by Comet as ones that already had low potential for C-terminal acetylation led us to look for another independent dataset to directly evaluate pyAscore's false localization rate.

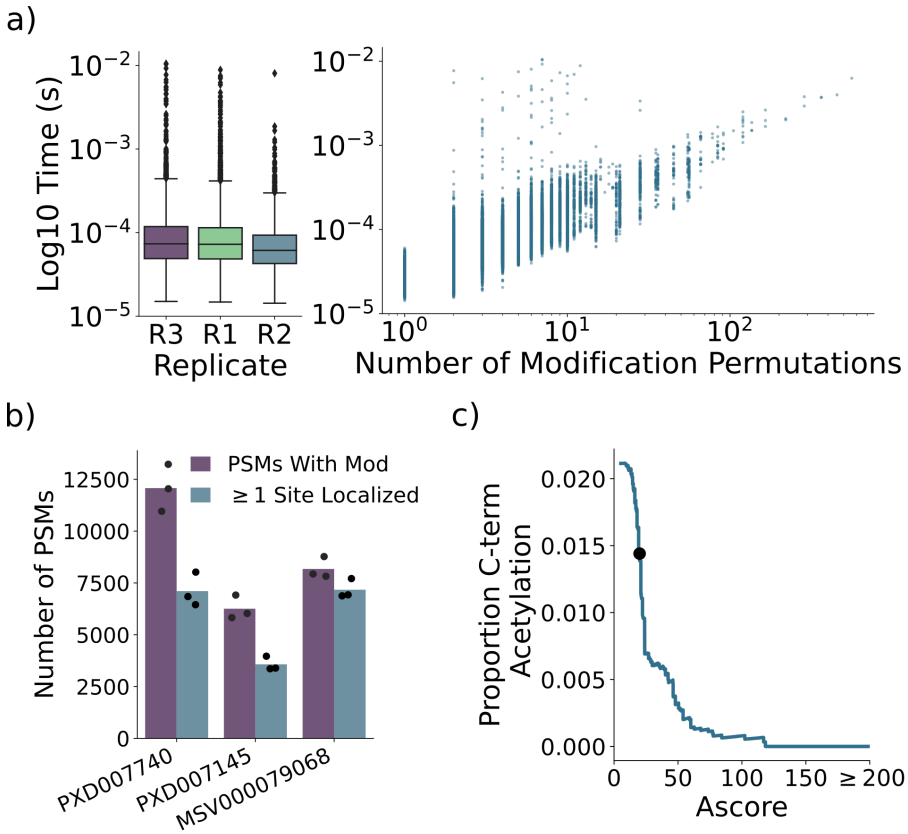


Figure 3.3. Experiments demonstrating pyAscore's speed and versatility. a) Analysis time per PSM for 3 replicate phosphoproteomic runs from PXD07740 (left), and the analysis time split by the number of unmodified/modified site permutations (right). **b)** Number of PSMs with a modification and the number of PSMs with at least one site localized for 3 replicate runs from each of PXD007740 (label-free phosphoproteome), PXD007145 (TMT-labeled phosphoproteome), and MSV000079068 (label-free acetylome). **c)** Cumulative number of falsely localized C-terminal acetylations by Ascore for all PSMs containing both an internal lysine and C-terminal lysine from MSV000079068.

3.5 Conclusions

Here we describe pyAscore, a flexible open source Python package for probabilistic scoring of PTM localization. Our package can be incorporated into a variety of analytical workflows and provides full access to internal parameters to tailor analyses to the user's instrument and PTM of interest. We showed that pyAscore is fast enough to be used in high performance settings such as in intelligent mass spectrometry data

acquisition alongside real time database search. In our online documentation, we supply examples of potential workflows in order to provide users with a detailed look at how they can apply pyAscore to their own data.

3.6 Data availability

Analyses within this study were performed on publicly available data from the ProteomeXchange[86] partner repositories, PXD007740, PXD007145, PXD000138, and PXD000759 from PRIDE (https://www.ebi.ac.uk/pride/archive) and MSV000079068 from MassIVE (https://massive.ucsd.edu). Result files from search and scoring steps can be found at the PRIDE repository PXD032908, and all code for reproducing validation experiments and generating figures can be found on GitHub at https://github.com/AnthonyOfSeattle/pyAscoreValidation.

Chapter 4. Building a phosphoproteomics knowledgebase through automated public data synthesis

4.1 Abstract

Global mass spectrometry methods are the workhorse of phosphopeptide identification and quantification but can be prone to low reproducibility due to stochastic acquisition. Alternatively, targeted phosphoproteomic assays, such as parallel reaction monitoring, can achieve high reproducibility and sensitivity, but generating the data necessary to design these assays can require significant monetary and time investment. One solution is to mine public data for information about peptides, synthesizing diverse experiments into a single database. Here, we build an automated and scalable data synthesis pipeline, where datasets are automatically downloaded from PRIDE or MassIVE before analysis by a scalable and reproducible pipeline powered by the workflow management system, Snakemake. The resulting database provides a powerful resource for building targeted phosphoproteomic assays in humans and allows the production of a large-scale spectral library which can be used to power peptide-centric analyses such as data independent analysis.

4.2 Introduction

Discovery mass spectrometry is the standard for characterizing the diversity of phosphorylation across the proteome, allowing the cataloging of tens of thousands of phosphorylation events in a single study.[14,49,87] Advances in sample preparation and instrumentation will likely continue to improve these techniques and provide previously

infeasible phosphoproteome coverage.[11,12,88] Even so, the most widely used discovery method, data dependent acquisition (DDA), suffers from a major limitation in its bias towards characterizing the most intense precursors in a sample, which can lead to high run-to-run detection variability.[24]

Accordingly, targeted methods have been heralded for their ability to enhance sensitivity and reproducibility in phosphoproteomics.[89] These methods have been successfully applied to validate discovery proteomic results without the need for site specific antibodies,[90] comprehensively characterize the dynamics of core pathways,[91,92] and probe global cellular signaling states through reduced sets of targets.[93,94] One technique which is of particular interest is parallel reaction monitoring (PRM), which repeatedly isolates and fragments a target mass-to-charge ratio during a specified retention time.[95] This technique provides a measurement for all resulting peptide fragments during this time, allowing individual ions to be filtered for intensity and interference.[96] However, generating a map of when desired peptides will elute and what charge state to target can be daunting, especially for difficult to detect peptides.

One solution is to mine the publicly available mass spectrometry data in ProteomeXchange member repositories like PRIDE and MassIVE for information about detectable phosphorylation events and their representative peptides.[8] Several databases have integrated this information by cataloging reported phosphosites and other post-translational modifications from publications.[38,97–101] An alternative approach is large scale re-analysis of datasets through a common computational pipeline, which offers the ability to control errors across the entire dataset and limit false discoveries.[9,102] This route is especially relevant for targeted proteomics, as it facilitates building a

detailed catalog of the detectable peptides covering a phosphosite along with their retention times and preferred charge states. With this goal in mind, in 2016 we built Phosphopedia, a public web portal for phosphopeptide information which simplifies the production of PRM assays.[10]

Unfortunately, Phosphopedia was built with a fundamental limitation–it was designed to be static, without the ability to continuously grow with the field. In this work, we address this limitation by redesigning Phosphopedia's input pipeline from the ground up, turning to the modern workflow management systems, Snakemake,[41] which provides reproducibility and scaling to thousands of mass spectrometry runs. This allows us to take full advantage of the growing wealth of phosphoproteomics data to build the core knowledgebase of our new web resource, Phosphopedia 2.0.

4.3 Methods

4.3.1 Phosphopedia's computational pipeline

Pipeline interface and data retrieval: Phosphopedia's re-engineered pipeline is built as a two-part system. The most important component is an integrated Snakemake workflow which manages the job graph and cluster deployment. The workflow is determined entirely at runtime allowing the pipeline to add new data with ease and only run the minimum number of rules necessary. Jobs are also matched to an individual conda environment specification, which provides reproducibility by handling all open source software download and installation. The pipeline's second component is a SQL database backend which augments the Snakemake functionality by efficiently storing information about completed tasks and acts as the repository for final pipeline output.

Phosphopedia's backend also makes use of SQLAlchemy as a programming framework, which allows portability between database management systems.

Currently, the Phosphopedia pipeline is available through a command line interface. The pipeline takes a single TOML configuration file where users can specify target datasets and individual parameters of the pipeline, allowing customization of individual pipeline steps. Target files are automatically identified and gathered from user specified ProteomeExchange accessions and local repositories of data. Remote targets are then downloaded using the ppx Python package, which currently allows download from Massive as well as the PRIDE repository. Once downloaded, raw files are converted to mzML format by ThermoRawFileParser. If any files fail to download or convert, they are marked and all other files are allowed to continue through the pipeline.

Database search and PSM scoring: After the pipeline has finished its preprocessing steps on all files, each file is automatically scanned for mass analyzer information so that data can be matched with the proper analytic parameters during subsequent search steps. Database search is then performed using the Comet distribution supplied in the Crux software suite (v. 3.2). The most important user inputs for this step are a FASTA file and a variable modification for the pipeline to focus on, but since these are arbitrary, the pipeline is theoretically generalizable to any feasible organism with known proteome and most PTMs. Furthermore, the user can control most of Comet's internal parameters for individual mass analyzer types, the only hard requirements being concatenated target-decoy search and PIN file output.

After database search, PSM scoring is performed by running Percolator (also from Crux v. 3.2) with default parameters independently on each run. This helps cut

down on memory consumption for the scoring step, as producing a single Percolator model for the entire database can be costly. It also adds flexibility to the models, allowing models to respond to the most important parameters for individual datasets. However, all scoring is kept at the PSM level, since individual matches to peptides and sites from all files need to be integrated together to score FDR at higher levels.

The final step in the search portion of the pipeline is scoring localization quality for the PTM of interest. Here, the top scoring match for each spectrum is outputted, and PTM localization is performed using pyAscore, our fast, in-house implementation of the Ascore algorithm.[61] This step provides several key pieces of information for the PTMs of interest -- their best location for the PTM in a PSM, the score of this localization, and the next best localizations for the PTM – which are taken into account in the integration step. Again, the user has the ability to control software parameters per mass analyzer type.

Peptide and phosphosite scoring: After database search, PSM scoring, and PTM localization, we collapse PSMs into peptides with a protocol that focuses on conservative discovery of PTM sites. First, PSMs from all searched files representing the same underlying peptide sequence are grouped together to determine a core set of potential modified peptides. Some of these PSMs will have well localized PTMs, defined as having an Ascore > 13, and others will not have confident localizations. Within a group of PSMs defined by having the same underlying unmodified sequence, we enumerate all PTM positions which are well localized as the first set of potential true sites. Part of the pyAscore output is a list of alternative sites that are the next best choices for PTM position. For all potential PTMs which are poorly localized, we look to

see if one of the alternative sites is already present in the well localized set. If it is, the PTM is relabeled as being in the position of the more confident identification; otherwise, it is counted as a new potential PTM site. Only the user-specified PTM of interest is potentially relocalized, while the others are allowed to stay in place.

Once a final set of modified peptides is enumerated, we score peptides with the maximum Percolator score of all corresponding PSMs in the database. We then use the maximum parsimony approach to protein mapping, and only maintain the maximum coverage protein for all peptide-protein correspondences. This then allows us to group together overlapping modified peptides, and score PTM sites with the maximum Percolator score for all peptides holding evidence for that site. This gives us Percolator scores for the PSM, peptide, and site level for both targets and decoys. Finally, we enforce a 1% FDR at each level using target-decoy competition.

For large database builds, it is often necessary to perform the above scoring steps on tens of millions of PSMs, which can be quite memory intensive. Thus, Phosphopedia's pipeline employs a pair of memory saving steps to reduce consumption. First, each file's FDR is cut off at a user specified value, which is set at a 5% default. This has the effect of ignoring PSM which will not contribute meaningfully to scoring. Second, the pipeline attempts to partition proteins into smaller sets, based on shared peptides. This allows the pipeline to only analyze a handful of PSMs at a time, which substantially reduces memory consumption while having no impact on the final scores. While the final scoring steps remain the most memory intensive portion of the pipeline, it should be noted that the human phosphopedia database did not require more than 64 Gb of RAM at this step.

Common retention time estimation: While a database of detections is useful for providing a catalog of peptides which can be detected for an organism, users who are interested in building targeted assays also want a map of when to expect peptides to elute from the liquid chromatography step. Within the database, each PSM for every detected peptide is associated with the retention time of the scan, but this cannot directly be supplied to users. If a user wants to build an assay with peptides detected in widely different runs, the user will want retention time values which are directly comparable to each other. Thus, we need to go through a process of converting retention times from individual files to a common scale. Classically, this has been done with indexed retention times, which require a common set of peptides to have been identified across the database to linearly align files. In our case, we needed a technique which acknowledges the fact that retention time differences between files are often non-linear and that, in large databases, there often does not exist any peptides which have been detected in all files.

We thus turned to an iterative approach to estimate retention times on a common scale. As a first step, we split all peptides with at least 4 hits in the database into a stratified training and test set, where 25% of the detections for each of the peptides are placed in the test set. The true measured retention times for each run in the training set are scaled and centered to make them as comparable as possible. The algorithm is then seeded by recording the average scaled retention time for all detections of a peptide across the training set, which provides a crude but close enough starting point. Then, according to user preference, either a linear or isotonic regression model supplied by the scikit-learn Python package is fit for each run, mapping the common retention

times to the scaled retention times from each run. Finally, these models are used to calculate the squared error of retention time prediction across the training set, allowing gradient calculation and common retention time update by gradient descent. This process is repeated until convergence. The held out test detections are used as an independent indicator of model fit and can serve to identify files which exhibit abnormally poor retention time alignments. Finally, in order to make the common retention time values feel natural to those familiar with iRTs, we define the .1% and the 99.9% percentiles of the common retention times as 0 and 100 respectively and scale the rest of the values to match.

Retention time outlier detection via shallow learning: While leaving out a test set of detections can provide insight into the error of common retention time estimation, we wanted a general procedure to detect outliers which was directly applicable to the RT values and available for all peptides in the database. One option is a sequence-to-retention time model which gives a predicted value for the common retention time for the peptide of interest that can be scored against the estimated common retention time. However, an issue that arises is that while we need scores for every peptide in the database, we cannot train a model on the same data that we score. We solve this problem by using a cross validation procedure. Here, we split the peptides into 10 random sets and leave each set out in turn as the scoring set while we train a predictor on the other 9. This allows the predictor to be used in an unbiased manner on the held out set to score.

We wanted the final procedure to be usable within the Phosphopedia pipeline, so it was advantageous for it to work quickly without specialized computation hardware (i.e.

a GPU). Thus, for this step we turned to multilayer perceptrons (MLP), which provide adequate low-precision predictions. First, the pipeline generates a feature vector for each peptide, which is defined as the count of each type of amino acid in the sequence, Then, using the MLP implementation in scikit-learn with a learning rate of 0.001, the pipeline builds one model per cross validation split which predicts retention times from the count vectors. These models are used to predict retention times on the held out test CV split, and measure the error in prediction.

Once a set of errors is generated for all splits, the pipeline attempts to calculate an outlier score for each peptide. Here, the pipeline takes a conservative approach and matches the distribution of errors against a heavy tailed theoretical distribution. In order to do this, each quantile of the error distribution from the 25th to 75th (in increments of 1) of the error distribution is matched against the same quantiles from t-distributions with increasing degrees of freedom. Once the optimum match is found, an outlier score is calculated as the -log10 of the two tailed p-value for the error given the closest theoretical distribution.

4.3.2 Dataset Curation

In order to find suitable data for the new Phosphopedia database, a search on PRIDE was conducted to find datasets which performed label free DDA phosphoproteomic experiments. Datasets were filtered for only those which had a subset of samples which were IMAC or antibody enriched and ran on Thermo Fisher Scientific instruments. Datasets were excluded if publication was still pending, as to avoid embargos. At time of writing, a total of 24 external datasets were included within the database, as well as lab data from 3 different instruments.

4.3.3 Search Settings

All files were searched with one of two parameter sets depending on the instrument used for MS2 analysis. Orbitrap data was searched with the Comet recommended high resolution settings, while ion trap data was searched with the Comet recommended low resolution settings. All searches were performed with a 50 ppm MS1 tolerance, an MS2 fragment tolerance of 1.0005 Da and 0.4 Da offset for low resolution MS2 data, and MS2 fragment tolerance of 0.2 Da for high resolution MS2 data. Carbidomethylation was set as a constant modification on Cysteines, and Phosphorylation of STY, Oxidation of M, and n-terminal acetylation were set as variable modifications. A standard proteome FASTA file was downloaded from Uniprot, sp_iso_HUMAN_4.9.2015_UP000005640.fasta, and used for all searches. The database search procedure for Phosphopedia itself is detailed above. Database search for data which was not included in Phosphopedia was performed with Comet and all files analyzed for a single dataset were run through Percolator together. Final analyses were performed on the Percolator results filtered at a 1% FDR at the PSM and Peptide level. All DIA data was analyzed using Spectronaut's directDIA with default parameters. The database and modification settings from the DDA data were carried over to the DIA analysis. For DIA data, all detections were controlled at a 1% precursor level FDR.

4.3.4 MS1 feature detection

For quantification of peptide precursors, we use the MS1 feature detection software Dinosaur and include the workflow as part of an experimental branch of the pipeline available at Phosphopedia github repository. Dinosaur is run with default

settings on every file to obtain an unbiased set of MS1 features without detections. Then, for every peptide with at least one PSM in a run, we take the retention time value for the best PSM for that peptide in the run and query the MS1 feature list for any features with bounds that overlap the retention time and have MZ within 10ppm of the peptide in at least one charge state from 1-6. This allows us to accumulate a charge distribution per run for every peptide with at least one PSM in that run.

4.3.5 Spectral library generation

A spectral library was generated from all phosphopeptides with at least 1 PSM in the high resolution (orbitrap MS2) data. The top scoring PSM, according to Percolator score, for each available charge state for each phosphopeptide was chosen for further processing. Fragment ions in each spectra were filtered out if their intensities fell below 5% of the base peak intensity. Spectrum_utils was then used to label b and y fragment ions with a tolerance of 0.2 Da,[103] and any spectrum with less than 3 annotations was filtered out. The median PPM error of the annotated peaks vs their theoretical masses was then used to globally correct the masses of all fragment ions remaining, and the precursor mass was replaced with the theoretical mass of the full ion. Finally, spectra were output in spTXT form, which is broadly compatible with software in the field.

4.4 Results

4.4.1 Pipeline Automation

In anticipation of the demands from ever growing repositories of phosphoproteomic data, we sought to build a MS data analysis pipeline that was fully

automated and scalable. In order to achieve this goal we employed Snakemake, a Python based workflow management system which offers the ability to build large scale reproducible data pipelines.[41] This allowed us to define a system which was built on completely open source components, and only run essential analysis steps upon database update. Files can be automatically downloaded from ProteomeExchange member repositories via ppx, and are then searched independently with Comet before PSMs and PTM localization are scored by Percolator and pyAscore respectively.[59,60,104] Once searched, results are aggregated from all runs in order to build a final global database which incorporates PSM, peptide, and site discoveries as well statistics about the behavior of peptides across instruments (Fig. 4.1a).

In building our pipeline, we recognized the broad applicability of such a tool for modified peptides. Thus, care was taken to allow the pipeline to generalize to most modifications of interest and any organism with an available FASTA. Data can either be housed locally or automatically retrieved from the MassIVE and PRIDE repositories, and each file is automatically scanned for mass analyzer information, so that data can be matched with the proper analytic parameters during database search steps. We also implement a procedure to limit the accumulation of false detections at the modified peptide and modification site levels, which is essential in databases which aggregate large amounts of data. These features make our pipeline a powerful tool for analyzing PTMs in public and private repositories of MS data. A detailed overview of pipeline steps and parameters is provided in the supplementary methods.

4.4.2 Database Update

We wanted to evaluate the performance of our new automated pipeline against the results from our analysis in Lawrence *et al.* (2016).[10] Thus, we re-analyzed the 995 files included in the original dataset and compared the final number of phosphopeptide-spectrum matches (phospho-PSMs), phosphopeptides, and phosphosites between the two searches. While both analyses generated similar numbers at the phospho-PSM and phosphopeptide levels, our new pipeline reported 8.2% less phosphosites (Fig. 4.1c). However, this is most likely due to the fact that the original database did not control phosphosite discoveries at a 1% FDR, suggesting the true phosphosite detections are lower. Taken together, the results of the two pipelines matched well, which suggested we could confidently start adding new data to the database.

With the new update, we continued to integrate both data produced within our lab and publicly available datasets. At time of writing, we added an additional 1,009 internally produced files and 2,517 files from public repositories, bringing the human phosphoproteomic database to a total of 4,521 individual mass spectrometry files (Fig. 4.1b). This data was then analyzed with the new automated analysis pipeline, producing a final database of detections with FDR controlled at the phospho-PSM, phosphopeptide, and phosphosite level globally. The necessity of this is clearly demonstrated by the fact that, for a given global score cutoff, the global peptide and site FDRs are consistently at least an order of magnitude higher than the PSM FDR, unless carefully controlled at each level (Fig. S4.1a). When looking at individual files, the global FDR control allows for some variability in the number of included false matches, but

~95% of files have a filewise PSM FDR < 0.02, implying the vast majority of the database is built on high confidence hits (Fig. S4.1b).

In total, our update brings the number of phospho-PSMs within the Phosphopedia database up to ~25.5 million and increases the total number of unique phosphopeptides by 35.7% to 259,745 (Fig. 4.1c). Notably, however, this only represented a total of 116,937 phosphosites, or an increase of 13.03%, demonstrating the logarithmically decreasing return with more data. To see why this is, we looked at the uniqueness of detections between repositories and discovered that the majority of phosphopeptides and phosphosites were detected in at least 2 datasets, recapitulating the problem that DDA mass spectrometry data tends to repeatedly measure the same peptides (Fig. S4.1d).

However, this phenomena provides an overall increase in the amount of unique data associated with individual phosphopeptides and phosphosites. When comparing the database builds resulting from Phosphopedia's new pipeline using the old and the new data sets, we observe an increase in the number of phospho-PSMs for 77.02% of phosphopeptides (Fig. S4.1c). In the new build 46.7% percent of phosphopeptides are represented in at least 2 charge states and 49.6% of phosphopeptides have both FTMS and ITMS spectra (Fig. 4.1de).

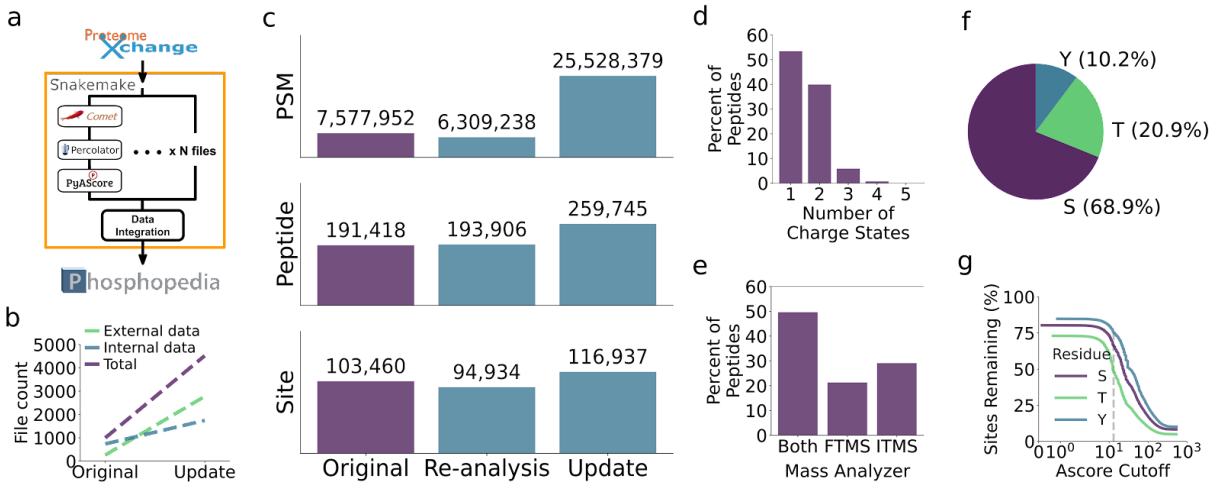


Figure 4.1. a) Overview of Phosphopedia's Snakemake powered data pipeline, which offers automated and distributed analysis of DDA mass spectrometry data with a focus on post translationally modified peptides. The pipeline handles automatic data download from PRIDE and Massive, performs database search and PSM scoring with Crux, and scores PTM localizations with Pyascore. Files are analyzed in parallel and then merged into a final set of identifications with FDR controlled at the PSM, peptide, and modification site level. **b)** Breakdown of data in Phosphopedia's original publication vs the current database update with colors describing whether data was derived from our own lab's instruments (Internal data) or from outside sources (External data). **c)** Matches and detections for 3 database builds. Original numbers are derived from the original Phosphopedia pipeline and data, and re-analysis uses Phosphopedia's new Snakemake pipeline to analyze the original set of data. The updated numbers describe the application of the new pipeline to the updated set of mass spectrometry runs. All FDRs are controlled at their respective levels except the site level count for the original Phosphopedia which is controlled at a 1% peptide level FDR, implying the site level FDR may be significantly greater than 1%. **d)** Histogram describing the number of unique charge states among all spectra matching to a given peptide in our database update. **e)** Histogram describing the percent of peptides matched to low resolution spectra (ITMS), high resolution spectra (FTMS), or both. **f)** Representation of serine (S), threonine (T), and tyrosine (Y) phosphorylation sites among all phosphorylation sites in our database update. **g)** Breakdown of number of phosphosites in our database with localization score at or greater than the given cutoff colored by type of site. The vertical grey bar indicates an Ascore cutoff of 13, the minimum cutoff for a site to be considered localized in Phosphopedia's pipeline.

This increase in information translates to the level of phosphosites as well. Phosphopedia's new database represents 80,559 serine sites, 24,392 threonine sites, and 11,986 tyrosine sites, with at least 40% sites in each category represented by at least 2 covering peptides (Fig. 4.1f; Fig. S4.2a). Regardless of type of phosphorylated residue, more than 60% of sites are represented by at least one PSM with an Ascore for that site of 13 or greater, our cutoff to label a site as localized (Fig. 4.1g). We wanted to understand how well our database represents current knowledge of protein regulation, so we downloaded a list of regulatory sites from Phosphositeplus which had been confirmed in both high and low throughput assays.[100] While we observed that our database covers 3296 of the 6222 listed sites (52.97%), it is likely that many of the remaining sites are only available in select conditions or still out of reach for DDA mass spectrometry (Fig. S4.2b). In addition, we obtained curated lists of sites representing biologically relevant signals from Krug *et al.* (2019) and matched each list against the sites in our database.[105] This gave a median of 7 sites per list, with 30.75% having 10 or more overlapping sites (Fig. S4.2c). In our resource, we have made an effort to provide easy access to peptides representing these curated sites, so that they can be combined into deeper targeted runs.

In all, we see experimental data as the gold standard for information in the database for its ability to build confidence in detections and provide information about the behavior of phosphopeptides in our instruments. Our automated pipeline makes it simple to continue this synthesis of public data, and we plan to add periodic updates as new data becomes public.

4.4.3 Peptide retention times

One of the main challenges in providing users with the optimum retention times and charge states for targeting phosphopeptides is synthesizing information from many individual experiments. For example, retention time (RT) is unlikely to be comparable between a large set of samples unless corrected computationally. A classic approach to this problem is to use a common set of peptides which have been identified in all samples to estimate linear transformations between the retention time spaces of individual runs. This posed a problem in Phosphopedia for two reasons. The first is that no peptides have been consistently identified in every single run across the database, even though most files have thousands of peptides identified (Fig. S4.3a). The second problem lies in the fact that transformations between gradients can be non-linear, necessitating more flexible models.

We tackle these problems by using a non-linear factorization model, which estimates both a comparable retention time value for every peptide in the database as well as non-linear transformations from the estimated retention time space to each individual run in the database. These values encode much of the relevant information for determining when a phosphopeptide will appear in an LC gradient, and can be directly used as iRT values when analyzing a new run (Fig. 4.2a). We trained our model on all peptides and split 25% of the detections from peptides observed in at least 4 files into a held out test set, from which we calculated the per peptide average loss. We observed a similar loss distribution across the gradient, with errors being slightly more pronounced in the middle of the gradient (Fig. S4.4a). This is likely due to greater

flexibility of the model at the ends of the gradient, allowing it to conform to the individual files more easily.

We wanted to further characterize the quality of RT estimation, and produce a metric which could be used for all peptides. An attractive option for the task was RT prediction using the sequence of a phosphopeptide. We describe the outlier detection algorithm in detail in the Methods, but we provide a brief overview here. We implemented a cross validation scheme which splits the phosphopeptides into 10 distinct sets, allowing us to hold each out in turn and train a predictor on the rest (Fig. S4.4b). This allowed us to build an unbiased outlier score based on the deviation of

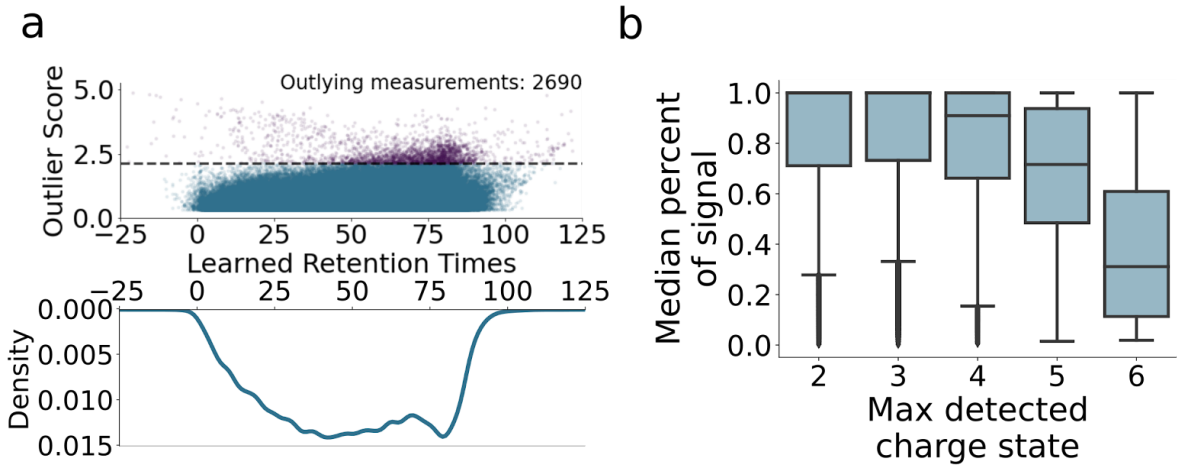


Figure 4.2. a) Density of learned retention times for phosphopeptides detected in our database (bottom) and outlier scores for individual peptides within the database vs their learned retention times (top). Learned retention times are produced using a non-linear alignment approach which estimates an average retention time while accounting for non-linear shifts in gradients between samples. The outlier score is based on the error of the common scale retention time to the theoretical retention time of the peptide estimated from a simplified sequence to retention time model. The outlier threshold is set at the 90th percentile of the theoretical distribution. **b)** The median percentage of signal across files for the most detected charge state for each phosphopeptide. Every phosphopeptide detection in every file was matched to as many MS1 features as possible, and the proportion of signal for each charge state was calculated as the signal for the feature divided by the sum of all features in the given file mapping to the given peptide.

each RT prediction vs the database's reported retention time on the held out test sets (Fig. S4.4c). The advantage of an approach like this is that it acts as a general strategy to identify outliers in sequence to retention time predictions, regardless of the underlying model. It was important that this step could be implemented as part of the Phosphopedia pipeline, so we utilized a multilayer perceptron based model, which provided adequate predictions quickly and without specialized hardware. We discuss the use of recurrent neural networks for RT prediction below.

Plotting our outlier scores against the estimated retention times reveals that most phosphopeptides with poor retention time estimates are concentrated near the high end of the RT space. According to this metric, a total of 2798 peptides would be considered as having outlying retention times, which only constitutes ~1% of the database (Fig. 4.2a). This information could be useful as individuals build assays, as it would give them information about whether to trust the database recorded RT value. Phosphopedia thus supplies the outlier score alongside the RT value, allowing users to judge the accuracy of individual values themselves.

4.4.4 Peptide charge state

In order to determine which charge states would be most effective to target, Phosphopedia provides spectral counts for every peptide in all detected charge states. Previously we showed that this statistic outperformed counting charge acceptors for this task. Here, we further evaluate this metric against the MS1 signal of detected peptides. We used Dinosaur to attempt to quantify the amount of signal for charge states 1-6 for every peptide detected in each run.[106] Consistently, this gave charge distributions for 74.6% of peptides in each file (Fig. S4.5a). However, it was clear that a subset of

peptides were consistently poorly quantified across files, even though some were consistently detected (Fig. S4.5b).

We compared the statistics from the qualifications to what was observed from detections. Interestingly, the majority of peptides could be quantified at least once occupying more than two charge states across files, which contrasts with the detection distribution discussed above (Fig. S4.5c). However, when users query the database information, it is desired that the claimed best charge states will have the most signal in real data. This appears to be the case for our database, as the max charge state according to spectral counts consistently contains at least 50% of the signal for charges 2-5 (Fig. 4.2b).

4.4.5 Agreement of database statistics with new runs

We wanted to determine how well the information contained within Phosphopedia would transfer to new instruments and biological contexts. Thus, we turned to a public phosphoproteomic dataset with runs which were not contained within the human Phosphopedia database to test the resource's predictions. Furthermore, we wanted a set of samples which would provide a high depth of coverage of available peptides in multiple charge states, a situation in which DIA mass spectrometry is particularly adept. Thus, we chose the human phosphoproteomic dataset MSV000082956 from Searle *et al.*, which contains paired DDA and DIA runs and is likely to have a high overlap in detections with Phosphopedia, making it ideal for testing peptide statistics in our database.[107]

We first set out to determine how well retention time predictions transfer to the new data. In order to do this, we needed calibration samples to build a map from the

database's iRT space to the new gradients. This was facilitated by selecting a set of 3 paired DDA and DIA runs, where each DDA run was performed right before its counterpart and could be used for calibration (Fig. S4.6a). All DDA files were searched together with a standard database search pipeline, and the DIA files were searched library free using the DirectDIA method from Spectronaut. For the DDA data, this resulted in an average of 8,134 detected phosphopeptides per run, with an average 7,914 (97.3%) overlapping with the database. As expected, the DIA produced a higher coverage of the underlying phosphoproteome with an average of 13,299 unique phosphopeptides detected, of which an average of 9,362 (70.4%) overlapped with the database predictions (Fig. S4.6b). The difference in overlap percentage is notable since it highlights that Phosphopedia is nearing saturation with respect to phosphopeptides detected by DDA mass spectrometry, and it shows that future work may benefit from integrating DIA mass spectrometry data into the database.

For our current purposes, both searches produced an abundance of data which could be used for validation purposes. For calibration, we chose the retention time of the best PSM from each DDA run and then took a random sample of 25 peptides from each run to fit a linear model between the database iRTs for each peptide and their experimental values. Using the calibrated predictions, we counted what percentage of peaks within the DIA data occurred at increasing time increments from the predicted values. The random calibration was repeated 10 times to determine the variability of the estimate. This revealed that >80% of peptides could be found within 3 minutes of the prediction and >90% of peptides could be found within 5 minutes (Fig. 4.3a). When looking at errors based on the predicted elution order, outliers tended to happen across

the gradient, but it is clear that bias occurs at the ends of the gradient which are harder to model (Fig. S4.8c). Given that a potentially large number of calibration peptides can be used, this effect could possibly be mitigated with more flexible calibration models.

When users select a peptide from Phosphopedia, they often have the choice between multiple charge states for analysis. The hypothesis is that by ranking these charge states by spectral counts within the database, a user can determine which charge state would be most likely to be seen. We wanted to determine whether a phosphopeptide could be detected in a charge state given that Phosphopedia predicted it to be present. Thus, we counted how many peptides in the DIA datasets were detected at least once in the best charge states according to Phosphopedia's spectral counts. This showed that >80% of validation peptides were detected in the top predicted charge state, and >95% were detected in one of the top two predicted charge states (Fig. 4.3b).

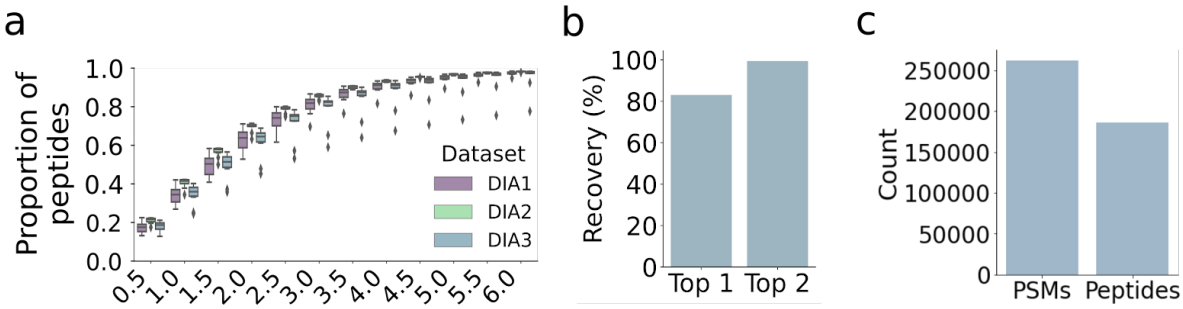


Figure 4.3. a) Proportion of detected peptides detected in 2 hr human phosphoproteome DIA data from MSV000082956 when cutting off at increasing experimentally measured retention time errors from their predicted values. Phosphopeptide retention times were calculated by calibrating database retention times to the experimental gradient. Calibration was performed using 25 random peptides from proceeding DDA runs, and repeated 10 times to generate box plots. **b)** Proportion of detected peptides in the same data which were detected in the top charge state or in one of the top two charge states according to Phosphopedia. Charge state ranks were determined by database spectral counts. **c)** Total number of phospho-PSMs and phosphopeptides in the final high resolution spectral library generated from Phosphopedia's detections.

4.4.6 Spectral library generation

While the phosphopeptide statistics within the new Phosphopedia database provide a detailed map with which individuals can build targeted phosphoproteomics assays, Phosphopedia also provides an important resource for improving the analysis of acquired data. By refining Phosphopedia's millions of PSMs into a core spectral library with high quality representations for multiple charge states per peptide, individuals can potentially greatly improve their power to analyze DIA datasets. The analysis of this type of data benefits greatly from a high quality spectral library. However, it can be difficult to achieve high depth of phosphoproteome coverage in a single study, a situation where Phosphopedia can potentially help.

In order to refine Phosphopedia to a core set of high quality spectra, we took the top scoring high resolution spectra per charge state and per peptide. After filtering out the bottom 5% of peaks, spectra contained a median of 72 peaks with a median of 12 of these peaks being able to be annotated to b and y fragment ions (Fig. S4.7a). One worry with integrating data from this number of different instruments and runs is PPM drift, which may be difficult to correct during later analysis. Thus, we used annotated peaks to determine a median PPM error of the theoretical masses of the annotated species against the experimentally determined mass. Across spectra, 90% of spectra had an absolute median PPM error less than 3.63, which shows that much of the data has high concordance with theoretical predictions (Fig. S4.7b). After filtering and quality control, we produced a final spectral library with 262,010 phosphorylated PSMs across 186,258 phosphopeptides (Fig. 4.3c). This high PSM to peptide ratio means that 66,812 (36.9%) phosphopeptides within the final library are represented by spectra in more

than two charge states, which provides a boost in versatility when analyzing those targets (Fig. S4.7c).

4.5 Discussion

Recently, the ProteomeExchange project reported that 13,461 proteomics datasets had been submitted to the PRIDE and MassIVE repositories as of June 2019.[8] Since then, submissions have continued in the hundreds per month, making it more important than ever to synthesize these datasets and make tools available for the broader proteomics community. With Phosphopedia 2.0, we sought to answer this challenge and bridge a fundamental gap in phosphoproteomics – continuously cataloging information in a way that allows researchers to quickly and confidently build new assays.

Our primary mechanism for this is an automated and extensible data synthesis pipeline. This has allowed us to bring together 4,521 human phosphoproteomics runs into a single repository, with the anticipation of continued growth in the coming years. Not only does this provide information about identifiable phosphosites and the physical properties of their representative phosphopeptides, but our library of more than 20 million PSMs provides a valuable dataset in of itself. Spectral libraries increase the power of computational analyses for both DDA and DIA data,[108,109] and researchers can download a spectral library containing representative high resolution spectra from our web portal. The full set of PSMs also provides an ample training dataset for individuals seeking to predict the spectra of phosphopeptides, and can be mined to study the fragmentation mechanisms of phosphopeptides across instruments.

As has been seen in other large scale analyses of DDA mass spectrometry data, even though the total number of phospho-PSMs within the database was nearly tripled, the number of phosphopeptides and phosphosites saw a much smaller increase. As will be seen in the subsequent chapter, we do not need to let this be a limitation. Instead, we can directly use Phosphopedia 2.0's core knowledge base to produce a resource which can handle any phosphopeptide of interest. This work goes beyond simply producing a core analysis of the human phosphoproteome by situating itself in an important thread of modern research, the democratization of public data. Many more individuals have use for public data than have the computational capacity to bring it together. By directly addressing this limitation, we open the door for researchers from across the field to build deep analyses of the human phosphoproteome.

Chapter 5. Predictive modeling to power targeted proteomics

Author contributions: This chapter was a joint collaborative effort between Anthony Barente, Alexis Chang, and Ricard Rodriguez. Alexis Chang collected the experimental data for the PRM validation assays, wrote the corresponding methods sections, and designed the analysis methodology for the resulting data. Ricard Rodriguez provided guidance in developing and analyzing the PRM validation assays. Anthony Barente built the Phosphopedia 2.0 web resource, performed all predictive modeling and initial validation, planned the PRM validation assays and designed the list of targets, and wrote up the results for all experiments.

5.1 Abstract

Mining public phosphoproteomics data to produce a central knowledge base of phosphopeptide detections provides a treasure trove of data, with every detection containing information on the relationship between amino acid sequence and the properties of a peptide. By predicting these properties from peptide sequences with deep learning, we can further extend the database to include novel peptides. Here, we have taken these principles and built Phosphopedia 2.0, a web resource which combines automated data synthesis with deep learning to power targeted and DIA phosphoproteomics experiments.

5.2 Introduction

As discussed in the previous chapter, the production of central repositories of phosphopeptide information to inform targeted phosphoprotomics assays is a powerful

technique to mitigate the upfront cost of targeted assay development. This said, solely relying upon information about peptides from experimental data has a fundamental limitation, i.e. that there is no direct information about peptides which have not been previously detected. Thus, when a mass spectrometrist is interested in understudied organisms or even single amino acid variants, they still need to perform a great deal of up front experimental work to generate information about those peptides. However, each detection of a peptide within a dataset contains information about how a specific amino acid sequence behaves in an instrument. By learning the connection between sequence and its physical properties, a mass spectrometrist can derive information about any phosphopeptide of interest, without the necessity of ever having seen it before.

Efforts to predict phosphopeptide properties from sequence have reached a fever pitch in recent years, especially with the success of deep learning methods.[110] Several of the most successful models to predict retention time from sequence initially focused on unmodified or oxidized peptides, due to the abundance of data for these species.[111,112] Interestingly, a single class of architecture, namely the class of recurrent neural networks, has worked exceedingly well at predicting several properties including retention time, charge state distribution, and fragmentation.[113] These models read individual amino acids one by one and create a final representation for the whole sequence which can be used for prediction. Often, unmodified and modified versions of an amino acid are represented as distinct entities, thus requiring that the analyst compile enough examples of the impact of the modified peptide on the property of interest to learn a quality representation. One effort to overcome this limitation was put

forward in DeepLC, which encoded the chemical composition of amino acids into the network itself, allowing the network to share information between different types of modifications and even predict retention times for modifications which were not included in the training set.[114]

In our case, we have produced a large database of phosphopeptides which provides ample high quality training data to build deep predictive models of phosphopeptide retention time and charge state distribution. We use these capabilities to extend the Phophosphedia 2.0 web resource, freeing key points in the assay production interface from reliance on database predictions. Thus, users are now able to dynamically choose between building assays from experimental data or predictions, and even combine the two sources.

5.3 Methods

5.3.1 Peptide property prediction via deep learning

Overview: In order to provide predicted targeting information for peptides which are not detected in the database, we turn to deep learning methods to predict properties directly from sequence. We chose to build all of our models using the Keras submodule of the Tensorflow deep learning framework and each model was trained using GPUs provided by the Google collab notebook system.[115] For the models discussed in the paper, all detected peptides within the human Phosphopedia database are split into train, validation, and test sets randomly with a proportion of 80%, 10%, 10%, respectively. We provide all training data and notebooks on our github

(github.com/AnthonyOfSeattle/PhosphopediaNotebook), where final model weights can also be accessed.

Predictors: For a model architecture we turned to gated recurrent unit neural networks, which provide state-of-the-art predictions while remaining comparatively simple.[116] The input to each model is the peptide sequence encoded as integers, with modified amino acids being given a unique integer so that the network could treat them differently than their unmodified counterparts. Within the model, the integer encodings are first mapped to 8 dimensional embeddings before being fed to a single layer single directional gated recurrent unit (GRU) layer. For the prediction of statistics which involve the entire peptide sequence, there is no expected benefit to doing bi-directional networks, as the network is able to encode all the necessary information with a single pass. As part of model selection, we evaluated adding recurrent dropout to the GRU layer. After the GRU step, the encodings pass to a fully connected layer which either maps the values to a single dimension in the case of retention time or to soft-max transformed 5 dimensional space in the case of charge state prediction to predict the probability of charge states 1-5.

For our retention time model training, we focused on predicting the common retention times directly. As discussed in the previous chapter, Phosphopedia's retention time alignment holds out a test set of PSMs in order to produce an estimate of modeling error. Training peptides were retained if they were detected at least 4 times in the database and their error did not exceed the 90th percentile. However, it should be noted that all peptides were retained in the test set to evaluate the performance of the model

in an unbiased fashion. The model was trained for this task using a squared error loss function.

In the case of charge state modeling, we first filter the train, validation, and test datasets to only retain peptides for which the number of files which a peptide is quantified in does not fall below 10% of the number of files a peptide is detected in. The final charge state profiles are normalized by dividing the charge profile for a peptide in a given file by the sum of the charge state intensities. Then the median value for each charge state is calculated for each peptide across runs before renormalizing the charge distribution. This process leaves many zeros within the dataset, which the networks will focus on in training. Thus, while we use a squared error loss function, we reweight the values so that the charge states with measurements count 4 times higher than the charge states without values.

5.3.2 Cell culture

HeLa S3 cells were cultured at 37°C and 5% $C0_2$ in Dulbecco's modified Eagle's medium (DMEM) supplemented with 4.5 g/L glucose, L-glutamine, and 10% fetal bovine serum and 0.5% streptomycin/penicillin. To generate bulk phosphopeptides for method comparisons, cells were grown to 80% confluency, incubated in serum-free medium for 6 hours prior to treatment with pervanadate (1 mM) for 15 min, followed by addition of 10% FBS for 15 min. At the time of harvest, cells were rinsed three times quickly with ice-cold phosphate-buffered saline (PBS), resuspended in minimal volume PBS, transferred to conical tubes, centrifuged at 2000 x g for 5 min at 4°C. PBS was discarded and cells were flash frozen in liquid nitrogen prior to storage at -80°C.

5.3.3 Sample preparation

Cell lysis was performed in 8 M urea, 50 mM HEPES, 75 mM NaCl, pH 8.0. Cell resuspension was sonicated 6 x 20 sec pulses of 5-6 watts with 20-30 sec rests, all while submerged in ice. Lysate was clarified by centrifugation at 7197 x g for 25 min at 20°C (not 4°C, to avoid precipitating urea). Protein content was assayed using the bicinchoninic acid method (Pierce). Proteins were reduced with 5 mM dithiothreitol (DTT) for 30 min at 55°C, alkylated with 15 mM iodoacetamide for 15 min at room temperature in the dark, then quenched with 5 mM DTT for 15 min at room temperature. Protein extracts were diluted to 5 mg/mL in 8 M urea, 50 mM HEPES, 75 mM NaCl, pH 8.0 and processed via an optimized R2-P1 protocol in 96-well plates on the King Fisher™ Flex robot. Phosphopeptide enrichment was performed via an optimized R2-P2 protocol on the same instrument.[11]

5.3.4 Mass spectrometry acquisition

Phosphopeptide-enriched samples were resuspended in 4% FA, 3% ACN and subjected to liquid chromatography on an EASY-nLC 1200 system equipped with a 100 µm inner diameter (ID) x 2 cm precolumn packed with Reprosil C18 3 µm beads (Dr. Maisch GmbH), and separated by reverse-phase chromatography on a 100 µm x 35 cm analytical column packed with Reprosil C18 1.9 µm beads (Dr. Maisch GmbH) and housed into a column heater set to 50°C. All separations were performed at a flow rate of 350 nL/min on an Orbitrap Eclipse.

DDA. DDA runs were performed using a multi-step gradient by varying ACN concentration in 0.1% FA as follows: 2.4% to 20.8% ACN from min 0 to 85, 20.8% to

40% ACN from min 85 to 103, 40% to 76% ACN from min 103 to 104, 76% ACN from min 104 to 109, 76% - 2.4% ACN from min 109 to 110, 2.4% ACN from min 110 to 120. Injections were performed in triplicate, each with 4 µL corresponding to 50 µg of peptide. Full MS scans were acquired from 350 to 1500 m/z at 70,000 resolution with AGC target set to standard (the instrument sets the recommended target in an automated fashion per scan type and user defined settings) and maximum injection time set to auto (system calculates maximum amount of injection time available to maximize sensitivity while maintaining maximum scan rate). The most abundant ions on the full MS scan were selected for fragmentation using 1.6 m/z precursor isolation window and beam-type collisional-activation dissociation (HCD) with 30% HCD collision energy for a cycle time of 3 s. MS/MS spectra were collected at 30,000 resolution with AGC target set to standard and maximum injection time set to auto given cycle time of 3 s. Fragmented precursors were dynamically excluded from selection for 30 seconds.

PRM. Parallel reaction monitoring runs were performed using a multi-step gradient by varying ACN concentration in 0.1% FA as follows: 2.4% to 20.8% ACN from min 0 to 43, 20.8% to 76% ACN from min 44 to 45, 76% ACN from min 45 to 49, 76% to 2.4% ACN from min 49 to 50, 2.4% ACN from min 50 to 60. The PRM methods consisted of a full MS scan (350 - 1500 m/z, 120,000 resolution, 8e5 AGC target, max injection time mode: auto) followed by up to 20 targeted MS/MS as defined by a time-scheduled inclusion list (50,000 resolution, 200% or 300% (1e5 or 1.5e5) AGC target, max injection time mode: auto, 1.6 m/z isolation window, 30% HCD collision energy). PRM assays #1 through #4 were performed in duplicate each with 4 µL injections corresponding to 58.8 µg of peptide and AGC target set to 200% (1e5). PRM

assay #5 was performed in duplicate with 12 uL injections corresponding to 176.5 ng of peptide and AGC target set to 300% (1.5e5).

Phosphopedia 2.0's web interface includes curated lists of biologically interesting phosphosites which can be used to build targeted assays for representative phosphopeptides. Using these lists, five different validation PRM assays were built. Assay 1 targeted 69 phosphopeptides representing cell growth control signaling. Assay 2 targeted 53 phosphopeptides curated from a combined list of phosphopeptides representing AKT and MAPK signaling. Assay 3 targeted 40 phosphopeptides representing transcriptional control signaling. Assay 4 targeted 50 phosphopeptides covering tyrosine kinase activation loops, all of which are phosphorylated on tyrosine residues. Assay 5 replicated peptide targets from Lawrence *et al.* (2018) to serve as a direct comparison along with 8 additional phosphopeptide targets based on new detecctions.[10] Assays 1 - 4 were acquired with 5 min windows on either side of predicted retention time. Assay 5 was acquired with 3 min windows on either side of predicted retention time. To calibrate PRM scheduling, an initial pilot DDA run was conducted with a 2 uL injection corresponding to 29 ng of peptide prior to assays 1 - 4 and a 1 uL injection corresponding to 14.7 ng of peptide prior to assay 5. Using Phosphopedia 2.0's web interface, a retention time calibration model was then built using 100 random peptides sampled from the middle 90% of detections from this run, excluding any phosphopeptides targeted in any PRM run. Suitability of an alternative low cost calibration matrix was tested by running 1 uL of unmodified yeast peptides corresponding to 250 ng immediately after HeLa phosphopeptide calibration matrix for

later analysis. Validation PRM assays were then run sequentially, with assays 1 - 4 performed on day 1 and assay 5 performed the following day with a new calibration.

5.3.5 Data processing and analysis

DDA. Raw DDA data files were converted to mzML and searched using Comet (version 2019.01.2) against either sp_iso_HUMAN_4.9.2015_UP000005640.fasta for human data or UP000002311_saccharomyces_cerevisiae_2020_03_22.fasta for yeast. Carbidomethylation was set as a constant modification on Cysteines, and Phosphorylation of STY, Oxidation of M, and n-terminal acetylation were set as variable modifications. Trypsin (KR|P) full digestion was selected allowing for up to 4 missed cleavages. Precursor mass tolerance was set to 50 ppm, and fragment ion tolerance to 0.02 Daltons. Final analyses were performed on the Percolator results filtered at a 1% FDR at the PSM and Peptide level.

PRM. For spectrum centric analysis of PRM mass spectrometry results, we used the same search pipeline as for DDA described above, except search results were not filtered to reach a 1% false discovery rate at any level (PSM, peptide or protein) in order to capture any potential hits since spectrum-centric searches served as comparison to manual precursor and product ion interrogation using Skyline.

Peptide-centric analysis was performed using Skyline. Trypsin digestion (KR|P) was set to max 2 missed cleavages. Peptide length was constrained to 8 to 25 amino acids without excluding any N-terminal amino acids. Carbamidomethylation of cysteine was set as a structural modification while up to 3 variable modifications and 1 loss were allowed. Signal extraction was performed on precursor charges +2, +3 and +4 and b and y product ion charges +1, +2 and as necessary +3. N-terminal to proline was

included as a special ion. Precursor and product mass were set to monoisotopic. Instrument transition settings were set for 150 to 1500 m/z with a method match tolerance of 0.055 m/z. Full-scan transition settings included 3 isotope peaks with mass accuracy of 10 ppm on centroided data. MS/MS filtering settings were “targeted” with mass accuracy of 10 ppm on centroided data. Retention time filtering was set to include all matching scans.

For each assay, modified peptides from inclusion lists were pasted into Skyline as targets. Precursor charges not targeted were removed. Explicit retention time and explicit retention time window were specified to narrow and expedite chromatographic peak picking. Raw files for both replicates for each assay were imported into the same Skyline document. Peptide identifications were further refined by manual interpretation using several criteria including product ion mass accuracy, correlation of precursor and product ion peak shapes and isotopic envelope ratios. To consider a peptide localized, we required at least 1-site diagnostic ion, although in many cases, more than one diagnostic ion was needed to rule out phospho-localization on other S, T, and Y residues within the peptide. After peak boundaries were set, product ions were filtered for the top 20 ranked ions, favoring longer fragments if a shorter fragment contributed marginally more peak area.

5.4 Results

5.4.1 Retention time prediction

The Phosphopedia database provides a valuable source of information for building targeted assays for human phosphopeptides, but is limited to peptides which

have been detected in the conditions and cell types analyzed. This said, every detection within the database is a direct link between the sequence of peptides and how they behave in the instrument. Thus, we set out to use this information to build predictive models which could provide information for any phosphopeptide a user was interested in. In order to do this, we turned to gated recurrent unit (GRU) neural networks.[117]

We first set out to build a predictor for Phosphopedia's peptide retention times. All peptides in the database were split into training, validation, and test sets and the training and validation sets were further filtered to focus on high quality retention time estimates. This provided 95,230 phosphopeptides to train a GRU network until convergence (Fig. S5.1ab). Evaluation of the model on the test data revealed that the error was highly dependent on the number of detections for a peptide (Fig S5.2a). This is expected, as the retention times in Phosphopedia are themselves averages between multiple runs. As detections increase, we are better able to average the errors between samples and the predictions become better. This is recapitulated in the fact that the pearson correlation of the test predictions is .984 for all peptides in the test set, but rapidly increases to .995 for peptides with at least 5 detections (n=12,001), which is comparable to state of the art (Fig. 5.1a).

While performance on held out data gives a good sense of the performance of the model, users will be most interested in the performance of the model on new runs. Thus, we analyzed a series of 9 previously published yeast 90 minute IMAC runs which were analyzed on our instruments.[11] In order for the model to transfer to the new data, it must first be recalibrated. We compared 3 different models for doing this step. In the first two models, we transformed the GRU network's 1 dimensional outputs, first by

scaling and shifting and then second by fitting a degree 3 polynomial with ridge regression. In the third model, we removed the final linear layer of the network and refitted the 512 dimensional output to the data with ridge regression as well. To compare the models, we performed a 10 fold cross validation with an increasing number of recalibration peptides. This revealed that the polynomial model performed better than the scale and shift for all numbers of recalibration peptides, reaffirming the importance of taking into account the non-linear deviations in the data (Fig. S5.2b). Interestingly, the polynomial model does better than retraining of the final layer of the network until ~2500 peptides are used for refitting.

In order to give a sense of performance in the worst case scenario, we choose to recalibrate with 25 random calibration peptides from our calibration run and then predict the retention times for the remaining 8 runs in the series. This revealed that errors only increased slightly across the runs with 95% of peptides falling within 4.77-5.47 minutes of the prediction and 75% of peptides able to be detected within 2.62-3.27 minutes (Fig. 5.1b).

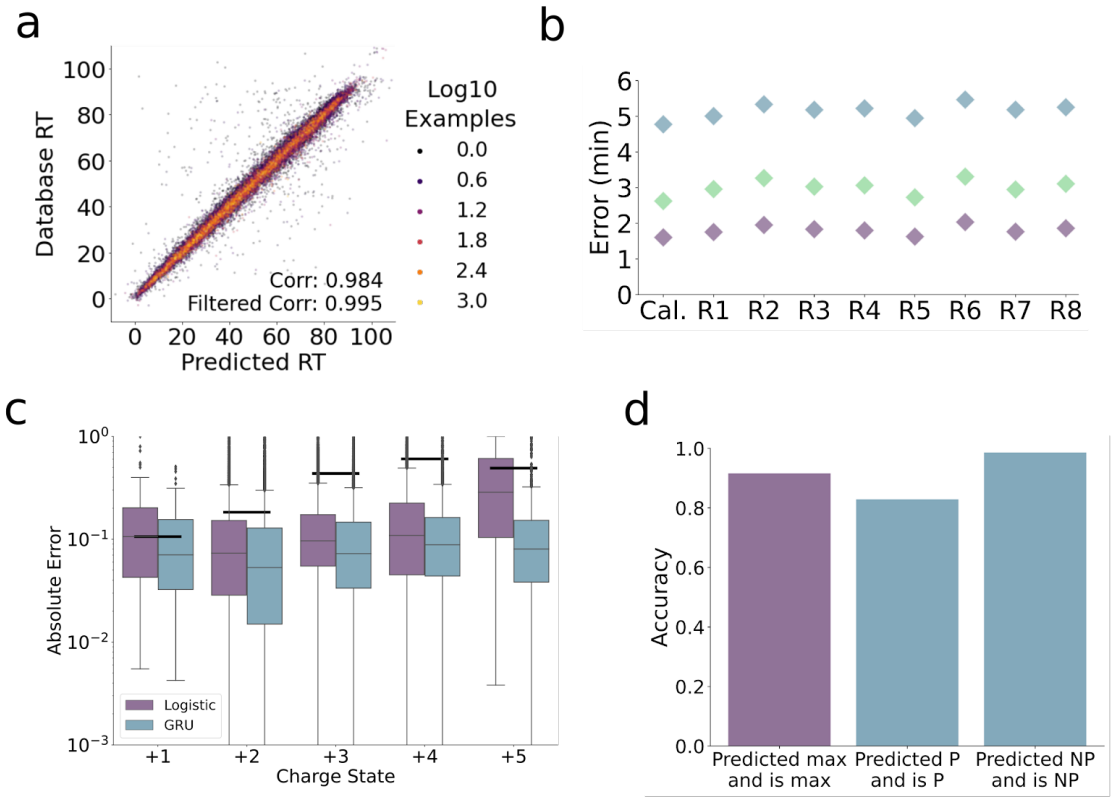


Figure 5.1. Training GRU networks to predict phosphopeptide properties from sequence. a) Predicted retention times vs the common scale retention times recorded in the Phosphopedia for the test set of phosphopeptides. Data is colored by the number of individual examples in the database available for estimating the retention time values. Two Pearson correlations are listed, one for the all phosphopeptides in the test set (Corr.) and one for all phosphopeptides with at least 5 hits (Filtered Corr.). **b)** Retention time predictive performance after recalibration across a series of 1.5 hr Yeast IMAC samples. Retention time predictions were recalibrated by fitting an order 3 polynomial to 50 randomly selected phosophopeptides from the first run, before applying the model to all subsequent runs. Colors show the 50th (red), 75th (green), and 95th (blue) percentiles of absolute predictive error for each file. **c)** Absolute error in predicted charge state proportion vs actual measured charge state proportion for non-zero phosphopeptide quantificiations in the test set. Results for logistic regression and GRU prediction are shown compared to the baseline results obtained from averaging the quantification values for each charge state in the training set. **d)** Utilizing the MS1 features determined by Dinosaur, we measured the accuracy of the GRU model to predict important aspects of a peptide's charge distribution. First, we measured the ability of the network to predict the max intensity charge state. Then, we measured the agreement between whether the network predicted that a charge state should have an MS1 feature and whether one was present in the test data, and the agreement when the network predicted that a MS1 feature was not there. The presence and absence of a feature or predicted charge state was determined by whether it exceeded a cutoff of 0.05 as a proportion of total signal.

5.4.2 Charge state prediction

Deep models can require a large amount of high quality data in order to obtain low errors. Above, we discussed the generation of MS1 quantification measurements for each peptide detection, which provide a higher sensitivity readout of the charge distribution of a peptide than spectral counts. This data also provides information about the +1 charge state, which is usually avoided in DDA acquisition. We thus decided to model the relationship between sequence and charge state using features with charge 1-5, excluding the +6 charge state due to its low abundance in the database. We partitioned data into the same train, validation, and test sets as our retention time data and filtered for quantifiability, leaving 166,631 training phosphopeptides (Fig S5.1a). Finally, we trained two different model classes to convergence on the data, a linear model which predicts charge distribution from amino acid counts and a GRU similar to the one used for retention time (Fig S5.1b). For comparison purposes, we also put together a baseline model, which is defined as the average of all charge distributions.

For most charge states, the predictions of the GRU model achieve quite low error, with the median absolute error being less than 0.1 (Fig. S5.2b). However, most of this likely comes from the high abundance of missing charge states in the data. We thus wanted to decouple the model's ability to predict presence from its ability to predict relative affinity for charge states. Thus, we plotted errors only for those charge states which have detected Dinosaur features. This revealed that the GRU model does better than both logistic regression as well as the baseline for all charge states, achieving median absolute errors of less than 0.09 for all charge states (Fig. 5.1c).

Often the network will predict values for charge states which are low enough that they should be interpreted as undetectable. We decided to set this threshold at 5%, and measured the accuracy relative to how often we actually detect a feature. We found that predicting that a charge state is not present is relatively simple, with the predictor achieving 98.6% accuracy. As stated above, this makes sense due to the overall abundance of missing values. Predicting presence is overall more difficult, with the network achieving 82.9% accuracy (Fig. 5.1d). Users are also interested in being able to differentiate between which charge states to target, so we looked at how well the network could predict the max charge state of data. The network did well in this regard, predicting the charge of the max intensity feature with 91.6% accuracy (Fig. 5.1d). Breaking this apart by charge state, most of this accuracy comes from charge states 2-4, with poorer performance for charge state 5 (Fig. S5.2d). Unfortunately, the +1 charge state doesn't occur as the max enough in the test dataset to provide sufficient evaluation here.

5.4.3 Agreement of predictions with new runs

Similarly to the database statistics presented in the previous chapter, we wanted to determine how well Phosphopedia's predictions transferred to unrelated phosphoproteomics runs. We thus chose the yeast phosphoproteome dataset PXD013453 from Leutert *et al.*, which is likely to have low overlap with the training dataset for our deep models, and thus makes a good testbed for Phosphopedia's ability to predict properties of new peptides.[11] In this case, we were able to find a single DDA run, which we used to calibrate the predicted retention times, and a subsequent set of three DIA runs which were used to validate predictions (Fig S5.3a).

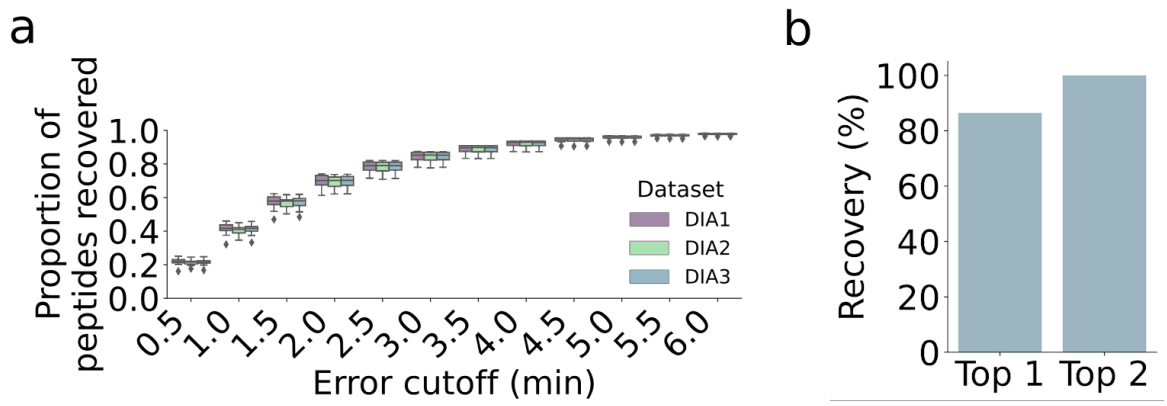


Figure 5.2. Validation of phosphopeptide retention time and charge state predictors on yeast phosphoproteomics data. a) Proportion of detected peptides detected in the 1.5 hr yeast phosphoproteome DIA data from PXD013453 when cutting off at increasing experimentally measured retention time errors from their predicted values. Phosphopeptide retention times were calculated by calibrating predicted retention times to the experimental gradient. Calibration was performed using 25 random peptides from the proceeding DDA run, and repeated 10 times to generate box plots. **b)** Proportion of detected peptides in the same data which were detected in the top charge state or in one of the top two charge states according to Phosphopedia. Charge state ranks were determined by Phosphopedia's sequence to charge state occupancy predictor.

The DDA run was searched with the same database search pipeline that was used in chapter 4, except with a yeast reference proteome, and the DIA files were searched library free using the DirectDIA method from Spectronaut. This resulted in 8,602 unique phosphopeptides for the DDA data, and an average of 16,344 unique phosphopeptides detected in the DIA data (Fig. S5.3b). Given that we were able to generate a prediction for every single phosphopeptide detected, evaluations could be performed on the entire dataset.

For our current purposes, we wanted to evaluate the situation where users only provide a small amount of recalibration data. Thus, we again chose the retention time of the best PSM from each DDA run and then took a random sample of 25 peptides from each run to fit a linear model between the predicted iRTs for each phosphopeptide and

their experimental values. Using the calibrated predictions, we counted what percentage of peaks within the DIA data occurred at increasing time increments from the predicted values. The random calibration was repeated 10 times to determine the variability of the estimate. Similarly to the database predictions, this revealed that >80% of peptides could be found within 3 minutes of the prediction and >90% of peptides could be found within 5 minutes (Fig. 5.2a). The distribution of error across the gradient was also similar, with outliers tending to happen across the gradient but bias concentrated near the ends (Fig S5.3c).

For charge state prediction, users are directly provided with a predicted proportion of MS1 signal. This should be directly related to the detectability of a phosphopeptide in the MS2, so we wanted to show it had a similar capacity as spectral counts to predict the presence of a species in the DIA data. Thus, we counted how many peptides in the DIA datasets were detected at least once in the best charge states according to Phosphopedia's charge distribution predictions. Similarly to the spectral counts in chapter 4, this showed that >80% of validation peptides were detected in the top predicted charge state, and >95% were detected in one of the top two predicted charge states (Fig. 5.2b).

5.4.4 Overview of the Phosphopedia 2.0 web resource

Phosphopedia 2.0 was built as a 3 component system, which adheres to modern design principles in building smart applications. Users interact with a java front end server, which dynamically handles retrieval of phosphopeptide information from a MySQL database for experimental information or from predictive modeling. Instead of directly integrating predictive modeling as part of the java web application, we built a

secondary REST API microservice, called DeepPepServer, which runs independently and provides predictions through HTTP requests. We built DeepPepSever to provide a common interface to serve any standard sequence based predictor, allow scaling predictive resources separately from front end and database resources, and provide application independence so that we could recycle the predictive server into new applications in the future.

When building our resources' new web interface, we wanted to seamlessly integrate database detections and model predictions to power two key phases of assay construction. First, when building a list of targets, users can specify phosphosites of interest and obtain retention time and charge distribution information for representative peptides regardless of previous detection, the latter being supplied by predictive models. This provides a list of phosphopeptide targets which can be mapped to the user's gradient to finalize an assay. A user can then supply a list of peptide-retention time pairs from a recent calibration run, which first uses our deep retention time model to predict where in the database's RTs the user's peptides would fall, and then fits a linear model to map the database's RTs to the user's gradient. This has the effect of allowing the user to calibrate with most standard unmodified and phosphorylation containing peptides, without the requirement that they be detected in the database.

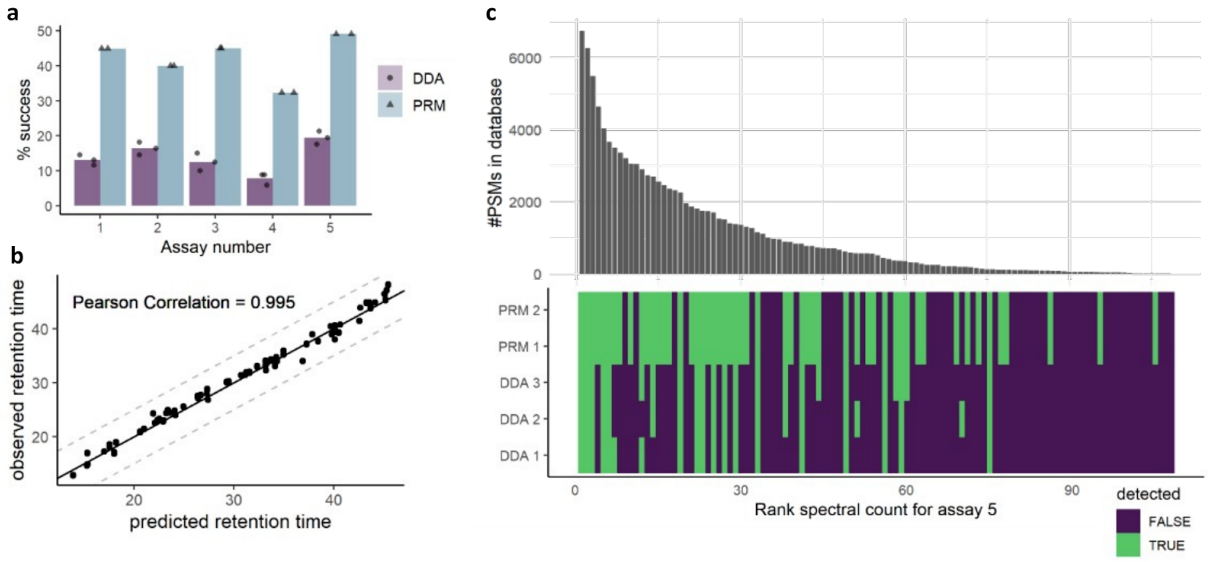


Figure 5.3. PRM validation experiments of the Phosphopedia 2.0 web resource. a) Percent recall (successful detections out of total targets included in each assay) for custom 60 min PRM assays run in duplicate and the percent recall for the same target lists in a series of triplicate 120 min DDA acquisitions. PRM assays 1-4 used 5 min RT windows on either side of the predicted retention time, whereas assay 5 used 3 min RT windows. **b)** Correlation between predicted and observed retention time for successfully identified targets in assays 1 to 4 following PRM acquisition each using 5 min windows on either side of predicted retention time. y=x (black line) and +/- 5 min (dashed gray line) are shown for reference. **c)** Target recall for assay 5 acquired by duplicate 60 min PRM assays using 3 min windows on either side of predicted retention time are compared to recall of the same target list following 120 min DDA acquisition in three technical replicates. Target peptides are ordered by the total number of database spectral counts (PSMs) within Phosphopedia.

5.4.5 Building custom PRM assays

In order to test our online interface, we selected 4 lists of phosphosites to build targeted assays. We then chose 100 random peptides from a 60 minute reduced volume DDA assay in order to calibrate the database retention times. This provided 4 final PRM runs with 5 minute windows that were performed in duplicate. An example elution profile is provided in Fig. S5.4a. Notably, calibration with 100 peptides from a 60 minute DDA yeast proteome sample yielded assays whose windows only differed by a median of 16.2 seconds (Fig. S5.4bc). In addition to our PRM runs, we also ran a

triplicate series of 120 minute DDA runs in order to compare depth of coverage. Targeted runs ranged in recovery from 32.4% to 42.0%, and detected between 2.3 and 5.1 times as many peptides on average compared to DDA of the same sample (Fig. 5.3a). This is notable given that these assays were derived directly from the database and performed on unstimulated cells. Assay 4 in particular contains entirely phosphotyrosine peptides, showing the potential to build assays for normally difficult to detect peptides without phosphotyrosine specific enrichment. When comparing predicted retention times vs retention times at peak intensity, we observed a high Pearson correlation of .995 with 75% of detections falling within 1 min of predictions and all falling within 2.9 min (Fig. 5.3b).

These results encouraged us to attempt a more challenging PRM run. Thus, we built a single run with 109 phosphopeptides which spanned the range of detectability according to the database. Another calibration DDA was performed, and a final PRM run with 3 minute windows and three times the input material was performed in duplicate. Detections from this assay correlated well with the number of PSMs per peptide in the database (Fig. 5.3c). Analysis of precursor and fragment intensities showed that peptides solely identified in the PRM run had intensity on average 3.26 fold lower (2.47-4.06; 95% t-test conf. Int. df=103.95) and 3.39 fold lower (2.66-4.12; 95% t-test conf. Int. df=103.98) respectively than those identified in both the PRM and DDA runs (Fig. S5.5b). Additionally, evidence for detected peptides could be found in both PRM runs, which contrasts with stochastic detection of the DDA analyses. Overall, 49.1% of targeted peptides were detected in the PRM run, with missing peptides being concentrated near the end of the gradient during our acetonitrile spike (Fig. 5.3a; Fig.

5.5c). It is likely that with direct stimulation of the targeted pathways, and a gradient which specifically targets more hydrophobic peptides, many of the remaining peptides would be recovered.

5.5 Discussion

The diminishing returns of the previous chapter, where increased data did not translate into substantially increased identifications, pushed us to explore how deep learning models for peptide properties could be seamlessly integrated into Phosphopedia to allow users to obtain information for peptides which are not present in the database. This integration also solves an issue with the cataloged information for phosphopeptides, where estimated retention times and charge distributions for low detection peptides may have higher variance than peptides detected across the database. For these peptides, users have the option to exchange experimental data for predictions and potentially obtain higher quality results.

One limitation with predictions is that there is no guarantee that the desired peptide is detectable even if a phosphosite exists. While predictors for proteotypicy exist for unmodified peptides,[118] their production for modified peptides has been less well studied due to the added problem of PTM site occupancy. However, since Phosphopedia allows for rapid production of assays, users can take more risks in incorporating phosphopeptides which have never been detected.

An interesting tactic our resource makes possible is the ability to optimize draft assays without having to explicitly monitor targets of interest. We showed that users can measure diluted samples or potentially unrelated samples and visualize how changes to

parameters such as LC gradients affect a list of peptides of interest without having to measure those peptides themselves. This opens many avenues for further research in both human and model organism systems. Given a protein or pathway of interest, a researcher could make an assay which targets all potential phosphosites to gather evidence for their existence and analyze their dynamics across conditions. Or an individual phosphoproteomic assay could be made for several organisms of interest, each targeting the same pathway but taking into account amino acid variants specific to each organism.

The integration of deep learning into proteomics promises to transform our ability to tackle biological systems.[110] Predictors which can predict physical properties of peptides, modified or not, directly from sequence are versatile tools in of themselves. However, it is not until they find integration into user oriented tools that their potential for the broader field can be realized. Over the last 2 chapters, we presented one vision for this integration, Phosphopedia 2.0, which provides a powerful integration of experimental data and predictions from deep models.

Chapter 6. Conclusion

6.1 Impact of presented work

6.1.1 A detailed overview for handling large scale datasets

Within this dissertation I provided two major case studies detailing the computational challenges of scaling up phosphoproteomics datasets either across large panels of treatments (Chapter 2) or across public repositories (Chapters 4 & 5). These situations are becoming increasingly more common in the literature as public repositories continue to grow and sample throughput continues to increase.[8,88] Discussion about the challenges and reward that these datasets bring is important so that future endeavors are conducted as efficiently as possible.

In particular, this dissertation addresses the data heterogeneity intrinsic to discovery phosphoproteomics through two different lenses. In the analysis of the Yeast Phosphoproteome Atlas, the heterogeneity in both phosphosite detection and quantification required careful correction in order to correctly model the underlying phosphoregulation. Here, the logic behind each step and its consequences were detailed to give a blueprint for future analyses of large perturbation screens. Similarly, during the production of the Phosphopedia 2.0 web resource, I provide a detailed look into all the steps necessary to combine data from multiple public datasets. This includes steps to limit the false discovery of phosphosites, and steps to correct heterogeneous phosphopeptide statistics within the database, such as retention time.[119]

Chapter 4 also provides a detailed look into the process of building a reproducible data pipeline for analyzing DDA phosphoproteomics data. By basing the

pipeline in the Snakemake workflow management system, I built a system which can scale to 1000s of data files and re-run only essential components as new data is added.[41] While the main product of this work was Phosphopedia 2.0 itself, the modularity of the underlying data integration pipeline means that other researchers have a starting point for building similar databases for other organisms and post-translational modifications (PTMs). Included in this is the necessity that all underlying technologies process data flexibly and efficiently. This challenge is addressed in chapter 3 with pyAscore, which details a python package providing fast and flexible PTM localization scoring, and essential step in building a database like Phosphopedia 2.0.

6.1.2 Increasing the usability of public phosphoproteomics data

A major theme of this dissertation was addressing a fundamental gap in phosphoproteomics, which lies in the usability of the continually increasing number of ProteomeXchange datasets.[8] As others have noted, there is a great deal of interest in the field to utilize the information contained within these repositories for future analyses.[120] Within chapters 4 and 5, I detail a complete vision for the Phosphopedia 2.0 web resource, which provides two distinct avenues to providing public data to build targeted phosphoproteomics assays.

The first avenue exists as a large-scale classic database of experimentally derived detections. I see this as the gold standard for phosphoproteomic information in the database, and users can easily build targeted assays from the maps of charge state and retention time contained within the data. Interestingly, chapter 4 details that the database is nearing saturation with respect to phosphopeptide and site detections from human tryptic samples. At a current count 259,745 phosphopeptides of 116,937

phosphosites, this database stands a middle ground between conservative estimates of the extent of phosphorylation derived from evolutionary arguments,[121] and liberal estimates from literature curation.[7]

Saturation of peptide detections presents a unique challenge, since it is likely that individuals will want to target phosphosites which are not in the database. This presents the second avenue for making use of the public data, since Phosphopedia 2.0's database can be used to train deep models which predict peptide properties straight from sequence.[110] Through integration of these technologies at multiple points in the assay production pipeline, I was able to build a resource which can mix experimental observations and predictions when building targeted assays, as well as calibrate to a user's liquid chromatography gradient with a broad range of peptides.

Together this represents a valuable resource of refined public data presented directly to the experimentalist, without the need that they curate the information for themselves. By using Phosphopedia 2.0, users can rapidly prototype new assays and use the extra flexibility to include targets which may be predicted to exist but never before detected. Once users have acquired data, either PRM or DIA, they can also take advantage of Phosphopedia's 262,010 high resolution phospho-PSMs across 186,258 phosphopeptides to perform powerful peptide centric analyses.

6.2 Looking forward

6.2.1 Data reuse at the forefront of phosphoproteomics

As the number and size of phosphoproteomics studies continues to grow, it is becoming increasingly clear that the planning phase needs to include a consideration of

data lifetime. For early single perturbation studies, authors could generally address most of the conclusions generated by their data in a single publication. In contrast, within some of the large phosphoproteomics datasets from the last decade and a half, the amount of questions generated exceeds what can easily be discussed in a single manuscript.[4] Given the far reaching impact that global studies of the phosphoproteome can have on signaling biology, deciding how to make results available to as wide of an audience as possible is imperative.

The proteomeXchange partner repositories provide one straightforward solution, which is to make the all data publicly available for others to reprocess on their own.[8] While access to these data are becoming increasingly simplified through programmatic interfaces,[104,122] building pipelines which can process the data is still specialized knowledge. Certainly the repositories themselves have pushed to bridge the gap with end users, especially with the creation of tiered submissions depending on if users provide raw or processed data. However, there is an imperative in the field to go further and provide the data in a form which is directly relevant to use cases within the field.

This can be accomplished through the production of online applications which allow exploration of the data. For phosphoproteomics and proteomics in general, several studies have provided successful examples. Currently, the human protein atlas exists as an interactive web application,[20] and a study of kinase activities across experiments allows users to compare and contrast between the included cell lines.[29] Another example is Phosphopedia 2.0, which provides its data directly in a form that is most relevant to experimentalists in phosphoproteomics.

Many of these attempts have been the purview of labs with dedicated computational times, but that is slowly changing with the growing popularity of R Shiny. Given that users have at least a basic understanding of the R programming language, the Shiny package provides a way to build interactive visualizations for users to ask tailored questions. Given the range of expertise which may be interested in the data, it is important that these applications hide most of the modeling logic from the user. To take an example from the Yeast Phosphoproteome Atlas, the server would handle all of the batch correction and linear modeling, while allowing a predefined range of analyses. In the end, a well designed application can allow rapid permeation of results into the field, and potentially greatly boost the lifetime of data.

6.2.2 Integration of predictive modeling into phosphoproteomic analysis

A related concept exists in the realm of predictive modeling as it applies to phosphoproteomics. As was discussed several times in this thesis, individuals have applied predictive modeling to understand several aspects of proteomics, including predicting peptide properties and predicting underlying signaling network architecture.[17,112] While the production of models is an interesting exercise in of itself, methods for making the model available beyond the original study should be given careful consideration. At the very least, the weights of the model should be made publicly available, and interaction could be further facilitated through the dissemination of Google Colab notebooks.

A burgeoning direction in the field is to integrate models directly into applications, in order to improve the next generation of experiments. Groups have already begun to integrate deep learning technologies into PSM rescoring applications which improve the

analysis of DDA data and applications which improve the distinction between signal and noise in DIA proteomics data.[112,123] The analysis of phosphoproteomics data comes with unique challenges, such as modification localization, which the integration of predictive models may greatly aid.[124] For designing new experiments, applications which allow modeling signaling states or applications like Phosphopedia 2.0 that aid in designing assays can fuel the production and validation of hypotheses.

6.3 Closing remarks

In summary, I have presented studies which detail methods for handling the increasing size and complexity of phosphoproteomics datasets. The techniques detailed in the contained chapters act as a guide for addressing the unique challenges presented by heterogeneous data, and provide an overview of the rewards which can come from undertaking large-scale studies. I expect that the field will benefit from these analyses as datasets continue to grow, and that the contained products will facilitate hypothesis and assay generation for years to come.

References

1. Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **80**, 225–236 (1995).
2. Cohen, P. The origins of protein phosphorylation. *Nature Cell Biology* vol. 4 E127–E130 (2002).
3. Bodenmiller, B. *et al.* Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. *Sci. Signal.* **3**, rs4 (2010).
4. Wilkes, E. H., Terfve, C., Gribben, J. G., Saez-Rodriguez, J. & Cutillas, P. R. Empirical inference of circuitry and plasticity in a kinase signaling network. *Proceedings of the National Academy of Sciences* vol. 112 7719–7724 (2015).
5. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.* **12**, (2019).
6. Hijazi, M., Smith, R., Rajeeve, V., Bessant, C. & Cutillas, P. R. Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. *Nat. Biotechnol.* **38**, 493–502 (2020).
7. Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261–70 (2012).
8. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2020: enabling ‘big data’ approaches in proteomics. *Nucleic Acids Res.* **48**, D1145–D1152 (2020).
9. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
10. Lawrence, R. T., Searle, B. C., Llovet, A. & Villén, J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. *Nat. Methods* **13**,

431–434 (2016).

11. Leutert, M., Rodríguez-Mias, R. A., Fukuda, N. K. & Villén, J. R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. *Mol. Syst. Biol.* **15**, e9021 (2019).
12. Humphrey, S. J., Karayel, O., James, D. E. & Mann, M. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. *Nat. Protoc.* **13**, 1897–1916 (2018).
13. Huttlin, E. L. *et al.* A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Cell* vol. 143 1174–1189 (2010).
14. Sharma, K. *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **8**, 1583–1594 (2014).
15. Dokládal, L. *et al.* Phosphoproteomic responses of TORC1 target kinases reveal discrete and convergent mechanisms that orchestrate the quiescence program in yeast. *Cell Rep.* **37**, 110149 (2021).
16. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* vol. 9 (2008).
17. Terfve, C. D. A., Wilkes, E. H., Casado, P., Cutillas, P. R. & Saez-Rodriguez, J. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun.* **6**, 8033 (2015).
18. Li, J. *et al.* TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J. Proteome Res.* **20**, 2964–2972 (2021).
19. Lundby, A. *et al.* Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nature Communications* vol. 3 (2012).
20. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
21. Studer, R. A. *et al.* Evolution of protein phosphorylation across 18 fungal species. *Science* **354**, 229–232 (2016).

22. Wang, B. *et al.* An overview of kinase downregulators and recent advances in discovery approaches. *Signal Transduct Target Ther* **6**, 423 (2021).
23. Morris, M. K., Saez-Rodriguez, J., Sorger, P. K. & Lauffenburger, D. A. Logic-Based Models for the Analysis of Cell Signaling Networks. *Biochemistry* vol. 49 3216–3224 (2010).
24. Tabb, D. L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **9**, 761–776 (2010).
25. Čuklina, J. *et al.* Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol. Syst. Biol.* **17**, e10240 (2021).
26. Jin, L. *et al.* A comparative study of evaluating missing value imputation methods in label-free proteomics. *Sci. Rep.* **11**, 1760 (2021).
27. Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **15**, 1116–1125 (2016).
28. Casado, P. *et al.* Kinase-Substrate Enrichment Analysis Provides Insights into the Heterogeneity of Signaling Pathway Activation in Leukemia Cells. *Science Signaling* vol. 6 (2013).
29. Ochoa, D. *et al.* An atlas of human kinase regulation. *Molecular Systems Biology* vol. 12 888 (2016).
30. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research* vol. 43 D512–D520 (2015).
31. Čuklina, J., Pedrioli, P. G. A. & Aebersold, R. Review of Batch Effects Prevention, Diagnostics, and Correction Approaches. *Methods Mol. Biol.* **2051**, 373–387 (2020).
32. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
33. Boedigheimer, M. J. *et al.* Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics* **9**, 285

(2008).

34. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29–39 (2016).
35. Leek, J. T., Evan Johnson, W., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* vol. 28 882–883 (2012).
36. Prasad, T. S. K., Kandasamy, K. & Pandey, A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* **577**, 67–79 (2009).
37. Bradley, D. & Beltrao, P. Evolution of protein kinase substrate recognition at the active site. *PLoS Biol.* **17**, e3000341 (2019).
38. Yu, K. *et al.* qPhos: a database of protein phosphorylation dynamics in humans. *Nucleic Acids Res.* **47**, D451–D458 (2019).
39. Schmidt, T. *et al.* ProteomicsDB. *Nucleic Acids Res.* **46**, D1271–D1281 (2018).
40. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
41. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
42. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
43. The, M., Matthew The, MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of the American Society for Mass Spectrometry* vol. 27 1719–1727 (2016).
44. Fondrie, W. E. & Noble, W. S. mokapot: Fast and Flexible Semisupervised Learning for Peptide Detection. *J. Proteome Res.* **20**, 1966–1971 (2021).
45. Hu, J. X., Zhao, H. & Zhou, H. H. False Discovery Rate Control With Groups. *Journal of the*

American Statistical Association vol. 105 1215–1227 (2010).

46. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).
47. Shteynberg, D. D. *et al.* PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. *J. Proteome Res.* **18**, 4262–4272 (2019).
48. Mohammadi, S., Saberidokht, B., Subramaniam, S. & Grama, A. Scope and limitations of yeast as a model organism for studying human tissue-specific pathways. *BMC Syst. Biol.* **9**, 96 (2015).
49. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).
50. Lanz, M. C. *et al.* In-depth and 3-dimensional exploration of the budding yeast phosphoproteome. *EMBO Rep.* **22**, e51121 (2021).
51. Soufi, B. *et al.* Global analysis of the yeast osmotic stress response by quantitative proteomics. *Mol. Biosyst.* **5**, 1337–1346 (2009).
52. Kanshin, E., Kubiniok, P., Thattikota, Y., D'Amours, D. & Thibault, P. Phosphoproteome dynamics of Saccharomyces cerevisiae under heat shock and cold stress. *Mol. Syst. Biol.* **11**, 813 (2015).
53. MacGilvray, M. E. *et al.* Phosphoproteome Response to Dithiothreitol Reveals Unique Shared Features of Stress Responses. *J. Proteome Res.* **19**, 3405–3417 (2020).
54. Kanshin, E., Bergeron-Sandoval, L.-P., Isik, S. S., Thibault, P. & Michnick, S. W. A cell-signaling network temporally resolves specific versus promiscuous phosphorylation. *Cell Rep.* **10**, 1202–1214 (2015).
55. Landry, C. R., Levy, E. D. & Michnick, S. W. Weak functional constraints on phosphoproteomes. *Trends Genet.* **25**, 193–197 (2009).
56. Leutert, M., Entwisle, S. W. & Villén, J. Decoding post translational modification crosstalk

with proteomics. *Mol. Cell. Proteomics* 100129 (2021).

57. Grassetti, A. V., Hards, R. & Gerber, S. A. Offline pentafluorophenyl (PFP)-RP prefractionation as an alternative to high-pH RP for comprehensive LC-MS/MS proteomics and phosphoproteomics. *Anal. Bioanal. Chem.* **409**, 4615–4625 (2017).
58. Villén, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* **3**, 1630–1638 (2008).
59. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
60. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
61. Barente, A. S. & Villén, J. A Python Package for the Localization of Protein Modifications in Mass Spectrometry Data. doi:10.1101/2022.04.04.487044.
62. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
63. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
64. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 57 289–300 (1995).
65. Buttrey, S., Samuel, Buttrey, E., Lyn & Whitaker, R. treeClust: An R Package for Tree-Based Clustering Dissimilarities. *The R Journal* vol. 7 227 (2015).
66. Schubert, O. T. *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10**, 426–441 (2015).
67. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based

Metabolomics Data. *Sci. Rep.* **8**, 663 (2018).

68. Zweytick, D. *et al.* Contribution of Are1p and Are2p to steryl ester synthesis in the yeast Saccharomyces cerevisiae. *Eur. J. Biochem.* **267**, 1075–1082 (2000).
69. Romanov, N. *et al.* Identifying protein kinase–specific effectors of the osmostress response in yeast. *Science Signaling* vol. 10 (2017).
70. Dogic, D. *et al.* The ADP-ribosylation factor GTPase-activating protein Glo3p is involved in ER retrieval. *Eur. J. Cell Biol.* **78**, 305–310 (1999).
71. Zhang, C.-J., Bowzard, J. B., Anido, A. & Kahn, R. A. Four ARF GAPs in Saccharomyces cerevisiae have both overlapping and distinct functions. *Yeast* **20**, 315–330 (2003).
72. Tu, J. & Carlson, M. REG1 binds to protein phosphatase type 1 and regulates glucose repression in Saccharomyces cerevisiae. *EMBO J.* **14**, 5939–5946 (1995).
73. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
74. Thomson, M. & Gunawardena, J. Unlimited multistability in multisite phosphorylation systems. *Nature* **460**, 274–277 (2009).
75. Casado, P. *et al.* Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* **6**, rs6 (2013).
76. Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9**, 4632–4641 (2009).
77. Chalkley, R. J. & Clauser, K. R. Modification site localization scoring: strategies and performance. *Mol. Cell. Proteomics* **11**, 3–14 (2012).
78. Schweppe, D. K. *et al.* Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J. Proteome Res.* **19**, 2026–2034 (2020).
79. Röst, H. L., Schmitt, U., Aebersold, R. & Malmström, L. pyOpenMS: a Python-based

interface to the OpenMS mass-spectrometry algorithm library. *Proteomics* **14**, 74–77 (2014).

80. Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry–based proteomics. *Nature Biotechnology* vol. 31 557–564 (2013).
81. Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V. & Gorshkov, M. V. Pyteomics--a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J. Am. Soc. Mass Spectrom.* **24**, 301–304 (2013).
82. Matheron, L., van den Toorn, H., Heck, A. J. R. & Mohammed, S. Characterization of biases in phosphopeptide enrichment by Ti(4+)-immobilized metal affinity chromatography and TiO2 using a massive synthetic library and human cell digests. *Anal. Chem.* **86**, 8312–8320 (2014).
83. Ressa, A. *et al.* A System-wide Approach to Monitor Responses to Synergistic BRAF and EGFR Inhibition in Colorectal Cancer Cells. *Mol. Cell. Proteomics* **17**, 1892–1908 (2018).
84. Hogrebe, A. *et al.* Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.* **9**, 1045 (2018).
85. Svinkina, T. *et al.* Deep, Quantitative Coverage of the Lysine Acetylome Using Novel Anti-acetyl-lysine Antibodies and an Optimized Proteomic Workflow. *Mol. Cell. Proteomics* **14**, 2429–2440 (2015).
86. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
87. Lundby, A. *et al.* Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nat. Commun.* **3**, 876 (2012).
88. Riley, N. M. & Coon, J. J. Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. *Anal. Chem.* **88**, 74–94 (2016).
89. Osinalde, N., Aloria, K., Omaetxebarria, M. J. & Kratchmarova, I. Targeted mass spectrometry: An emerging powerful approach to unblock the bottleneck in

phosphoproteomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **1055-1056**, 29–38 (2017).

90. Osinalde, N. *et al.* Nuclear Phosphoproteomic Screen Uncovers ACLY as Mediator of IL-2-induced Proliferation of CD4+ T lymphocytes. *Mol. Cell. Proteomics* **15**, 2076–2092 (2016).
91. Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D. A. & White, F. M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5860–5865 (2007).
92. de Graaf, E. L. *et al.* Signal Transduction Reaction Monitoring Deciphers Site-Specific PI3K-mTOR/MAPK Pathway Dynamics in Oncogene-Induced Senescence. *J. Proteome Res.* **14**, 2906–2914 (2015).
93. Abelin, J. G. *et al.* Reduced-representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-scale Comparison of Drug-induced Phenotypes. *Mol. Cell. Proteomics* **15**, 1622–1641 (2016).
94. Soste, M. *et al.* A sentinel protein assay for simultaneously quantifying cellular processes. *Nat. Methods* **11**, 1045–1048 (2014).
95. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **11**, 1475–1488 (2012).
96. Bourmaud, A., Gallien, S. & Domon, B. Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications. *PROTEOMICS* vol. 16 2146–2159 (2016).
97. Bodenmiller, B. *et al.* PhosphoPep--a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.* **26**, 1339–1340 (2008).
98. Gnad, F., Gunawardena, J. & Mann, M. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res.* **39**, D253–60 (2011).

99. Dinkel, H. *et al.* Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res.* **39**, D261–7 (2011).

100. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–20 (2015).

101. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

102. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

103. Bittremieux, W. spectrum_utils: A Python package for mass spectrometry data processing and visualization. doi:10.1101/725036.

104. Fondrie, W. E., Bittremieux, W. & Noble, W. S. ppx: Programmatic Access to Proteomics Data Repositories. *J. Proteome Res.* **20**, 4621–4624 (2021).

105. Krug, K. *et al.* A Curated Resource for Phosphosite-specific Signature Analysis. *Mol. Cell. Proteomics* **18**, 576–593 (2019).

106. Teleman, J., Chawade, A., Sandin, M., Levander, F. & Malmström, J. Dinosaur: A Refined Open-Source Peptide MS Feature Detector. *J. Proteome Res.* **15**, 2143–2151 (2016).

107. Searle, B. C., Lawrence, R. T., MacCoss, M. J. & Villén, J. Thesaurus: quantifying phosphopeptide positional isomers. *Nat. Methods* **16**, 703–706 (2019).

108. Zhang, X., Li, Y., Shao, W. & Lam, H. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics* **11**, 1075–1085 (2011).

109. Zi, J. *et al.* Expansion of the ion library for mining SWATH-MS data through fractionation proteomics. *Anal. Chem.* **86**, 7242–7246 (2014).

110. Meyer, J. G. Deep learning neural network tools for proteomics. *Cell Rep Methods* **1**, 100003 (2021).

111. Ma, C. *et al.* Improved Peptide Retention Time Prediction in Liquid Chromatography

through Deep Learning. *Anal. Chem.* **90**, 10881–10888 (2018).

112. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).

113. Guan, S., Moran, M. F. & Ma, B. Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. *Mol. Cell. Proteomics* **18**, 2099–2107 (2019).

114. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **18**, 1363–1369 (2021).

115. Paramasivam, K., M., P. & Sharif, H. Heterogeneous Large-Scale Distributed Systems on Machine Learning. *Deep Neural Networks for Multimodal Imaging and Biomedical Applications* 47–68 (2020) doi:10.4018/978-1-7998-3591-2.ch004.

116. Cho, K., van Merrienboer, B., Bahdanau, D. & Bengio, Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (2014) doi:10.3115/v1/w14-4012.

117. Cho, K. *et al.* Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014) doi:10.3115/v1/d14-1179.

118. Serrano, G., Guruceaga, E. & Segura, V. DeepMSPeptide: peptide detectability prediction using deep learning. *Bioinformatics* **36**, 1279–1280 (2020).

119. Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111–1121 (2012).

120. Martens, L. & Vizcaíno, J. A. A Golden Age for Working with Public Proteomics Data. *Trends Biochem. Sci.* **42**, 333–341 (2017).

121. Kalyuzhnyy, A. *et al.* Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation. doi:10.1101/2021.04.14.439901.

122. Reisinger, F., del-Toro, N., Ternent, T., Hermjakob, H. & Vizcaíno, J. A. Introducing the PRIDE Archive RESTful web services. *Nucleic Acids Res.* **43**, W599–604 (2015).

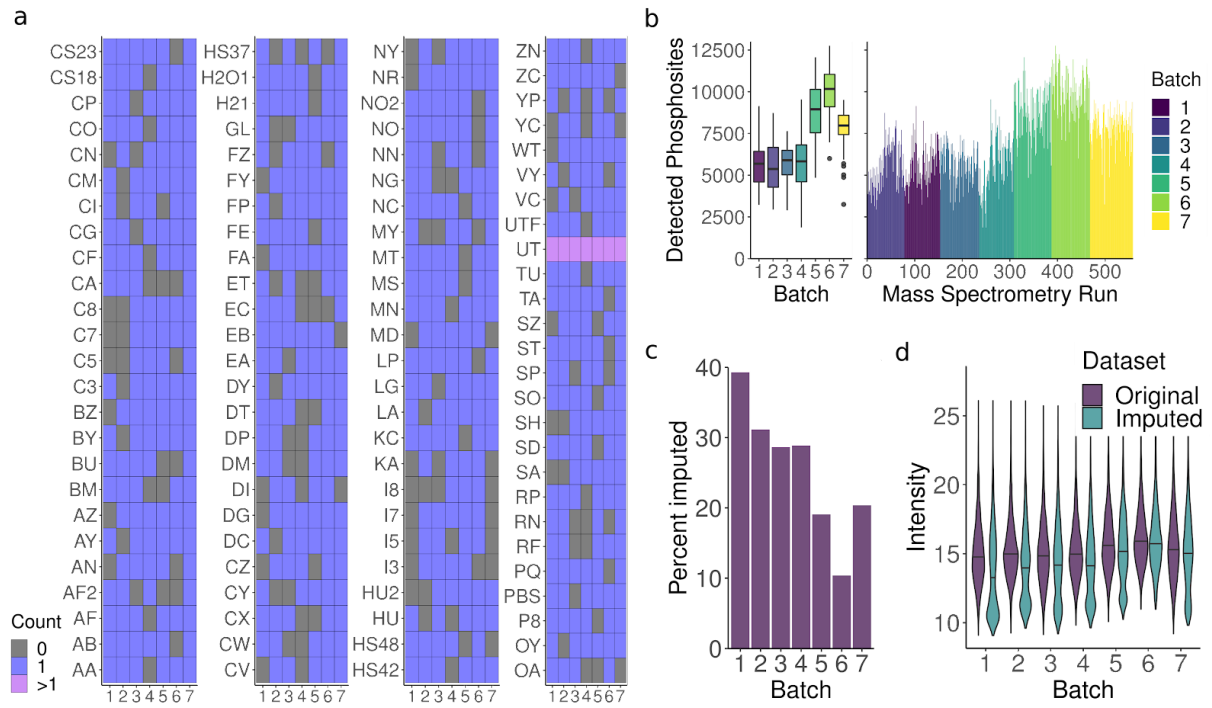
123. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods* vol. 17 41–44 (2020).

124. Lou, R. *et al.* DeepPhospho accelerates DIA phosphoproteome profiling through in silico library generation. *Nat. Commun.* **12**, 6685 (2021).

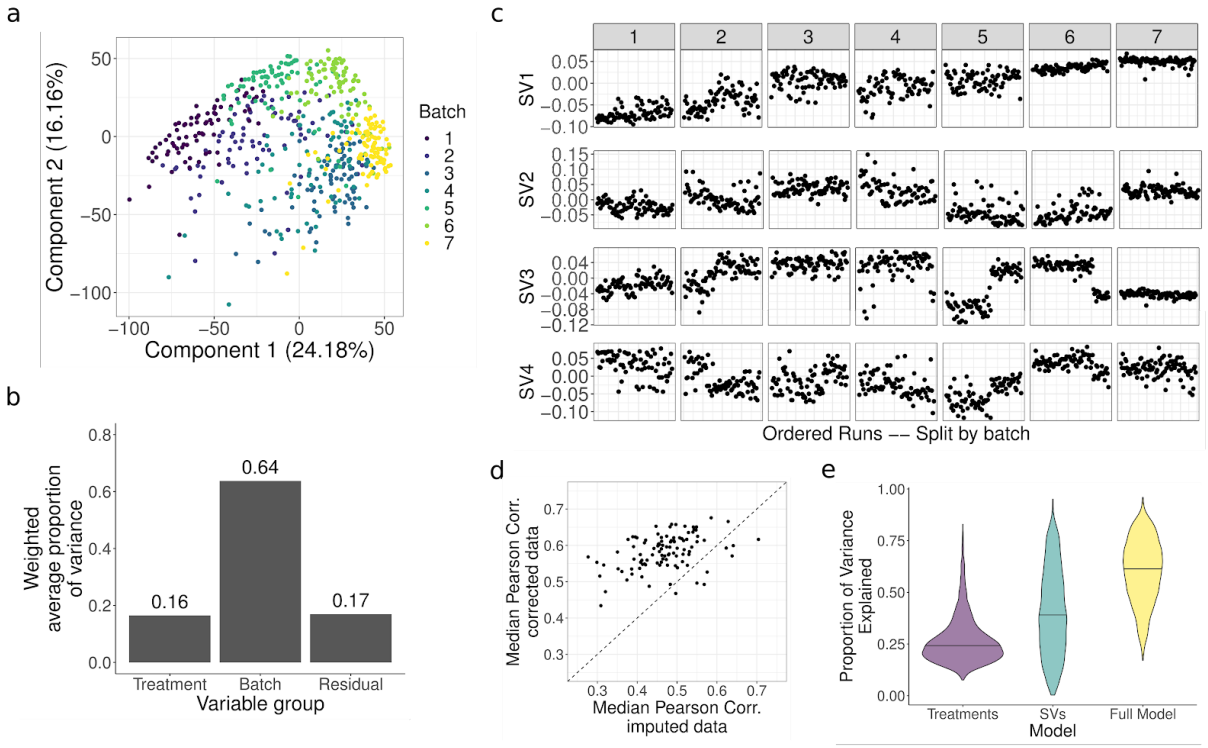
Appendix A

Supplementary information for chapter 2.

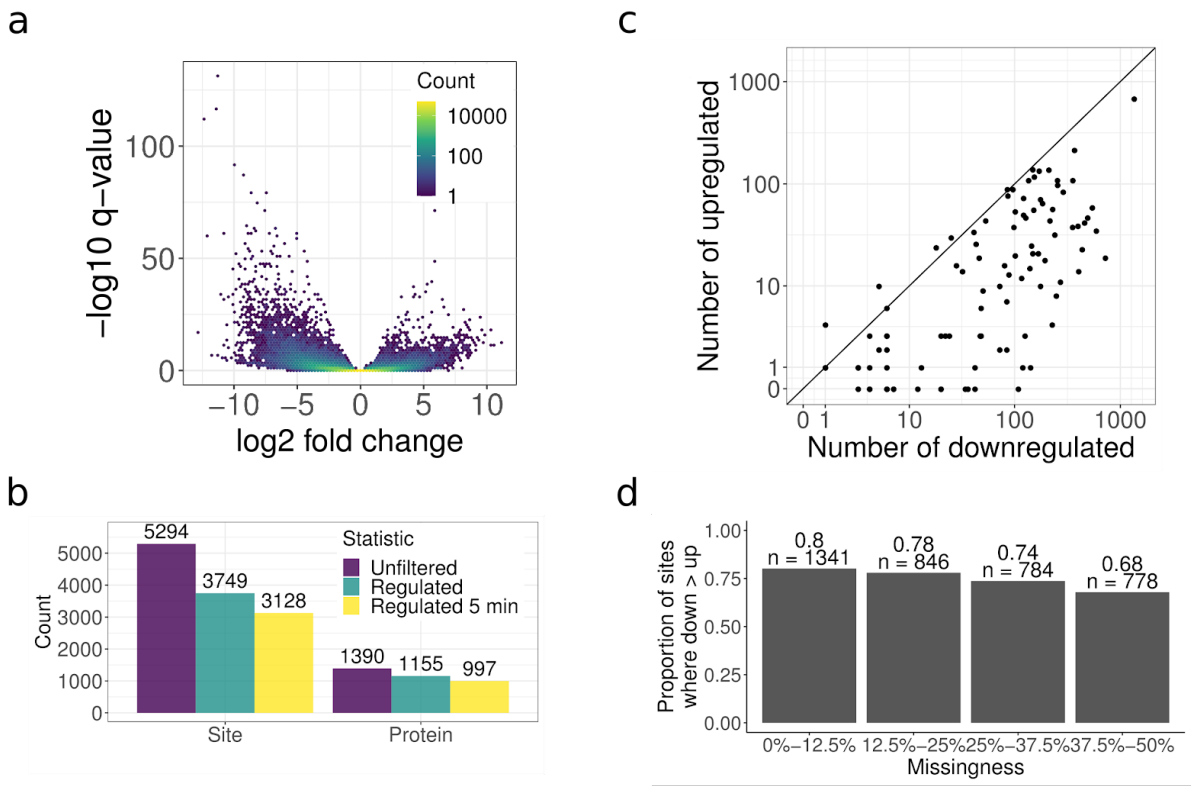
Supplementary Figures



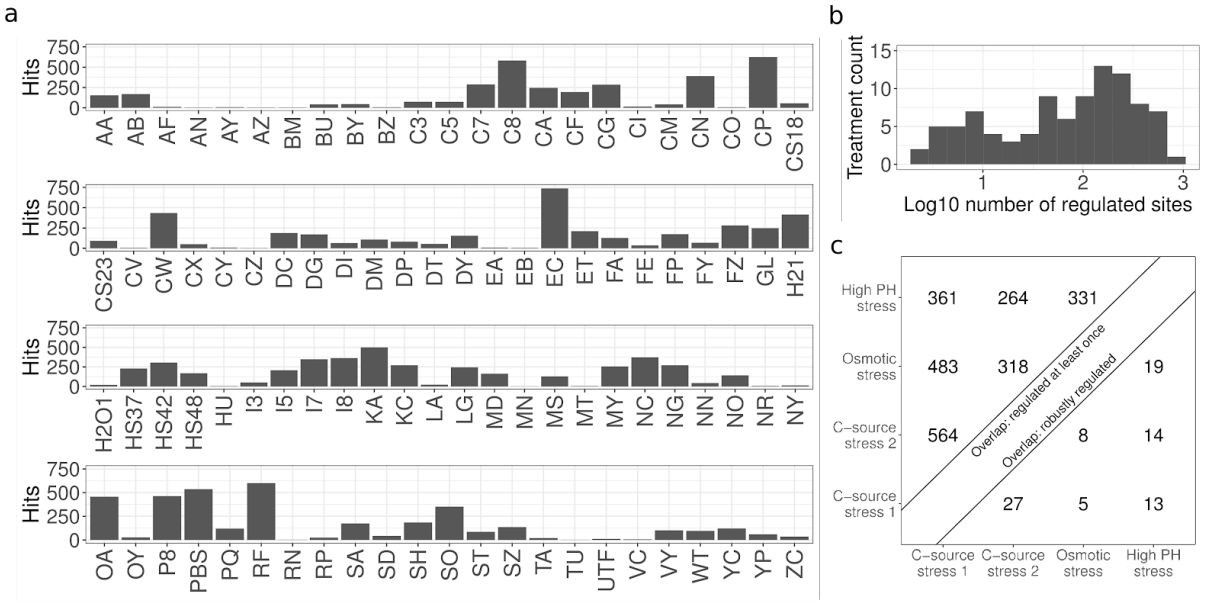
Supplementary Figure 2.1. Yeast phosphoproteome atlas detection statistics across sample batches. a) Distribution of all treatments, including treatments with durations longer than 5 minutes, among the 7 R2-P2 batches for the quantitative DIA samples in the yeast phosphoproteome atlas. Treatment definitions can be found in table 4.1. **b)** Proportion of all samples where a phosphosite was detected vs the median log2 intensity of the site in the remaining samples before imputation and batch correction. Color indicates the number of phosphosites falling within a 2d hexagonal bin. The observed weak logistic trend, implies that peptide abundance is in part at play for missing quantifications. **c)** A box-plot of the number of quantified phosphosites per sample across 7 batches. **d)** Barplot representing the percent of values within each R2-P2 batch which are imputed after filtering for sites which appear in 50% of the samples in the dataset. **e)** The change in intensity distribution per batch after QRILC imputation.



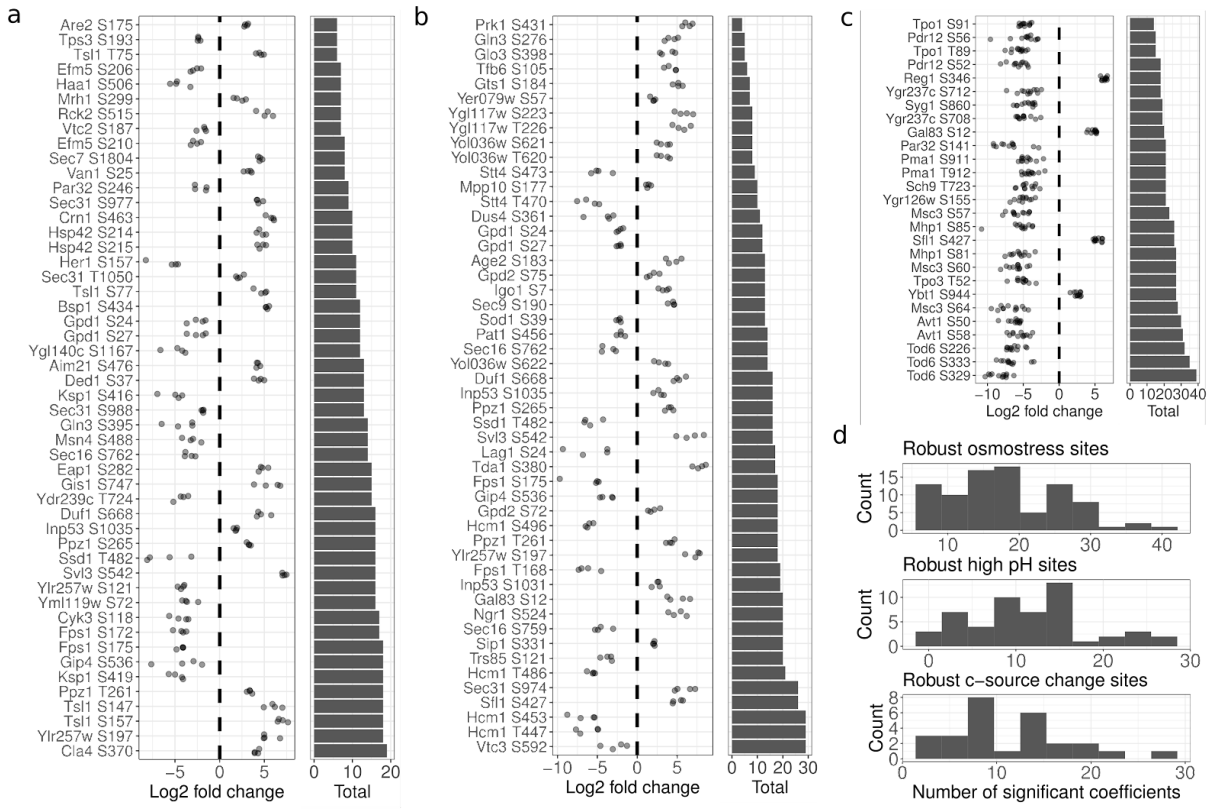
Supplementary Figure 2.2. Modeling batch effects for differential expression analysis. a) A PCA of phosphosite quantifications per sample colored by sample batch after phosphosites were filtered for missingness, and **b)** a Principal Variance Component Analysis (PVCA) demonstrating the relative amount of signal explainable by the treatment applied to cells, the sample batch, and a residual component. **c)** Plots of the 4 surrogate variables determined by the SVA package for individual samples plotted by order of sample acquisition and split by sample batch. **d)** Scatter plot demonstrating the median Pearson correlation between different samples from the same treatment across batches before (imputed data) and after subtracting out the fitted signal of the surrogate variables (corrected data). In order to perform correction on the phosphosite intensities, a model with solely an intercept and the surrogate variable values was fit per phosphosite with LIMMA, and the fitted signal of the surrogate variables was subsequently subtracted out. **e)** The proportion of variance explained across phosphosites was compared for 3 different models, one which included solely an intercept and the treatment variables (Treatments), one which contained an intercept and the surrogate variables (SVs), and a final model which included an intercept, treatment variables, and surrogate variables (Full Model). Each model was fit to the uncorrected imputed data using the LIMMA package.



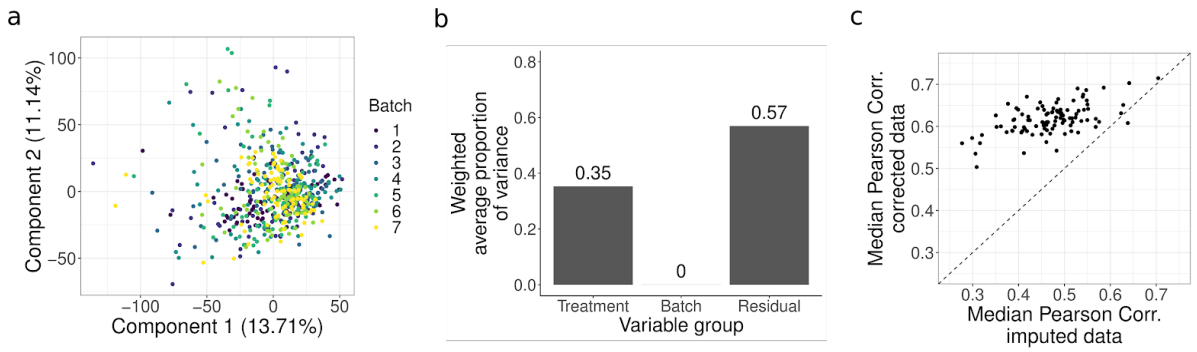
Supplementary Figure 2.3. Global statistics from differential expression analysis with LIMMA. a) Volcano plot displaying the negative log10 of the Benjamini-Hochberg corrected p-values by the log2 fold change determined by LIMMA. Each p-site in each treatment was tested for differential expression against the untreated controls, and all tested coefficients are displayed in the volcano plot. **b)** Counts of the number of sites and proteins with at least 1 significant (global FDR ≤ 0.01 and absolute log fold change ≥ 1) differential expression event (regulated) in either all treatments modeled by LIMMA or the 5 minute treatments. **c)** Number of significant negative coefficients (downregulated) vs positive coefficients (upregulated) for each treatment modeled by LIMMA. **d)** Bar chart with sites binned by percent missing data displaying the percent of sites where the number of significant down regulated coefficients numbers upregulated coefficients.



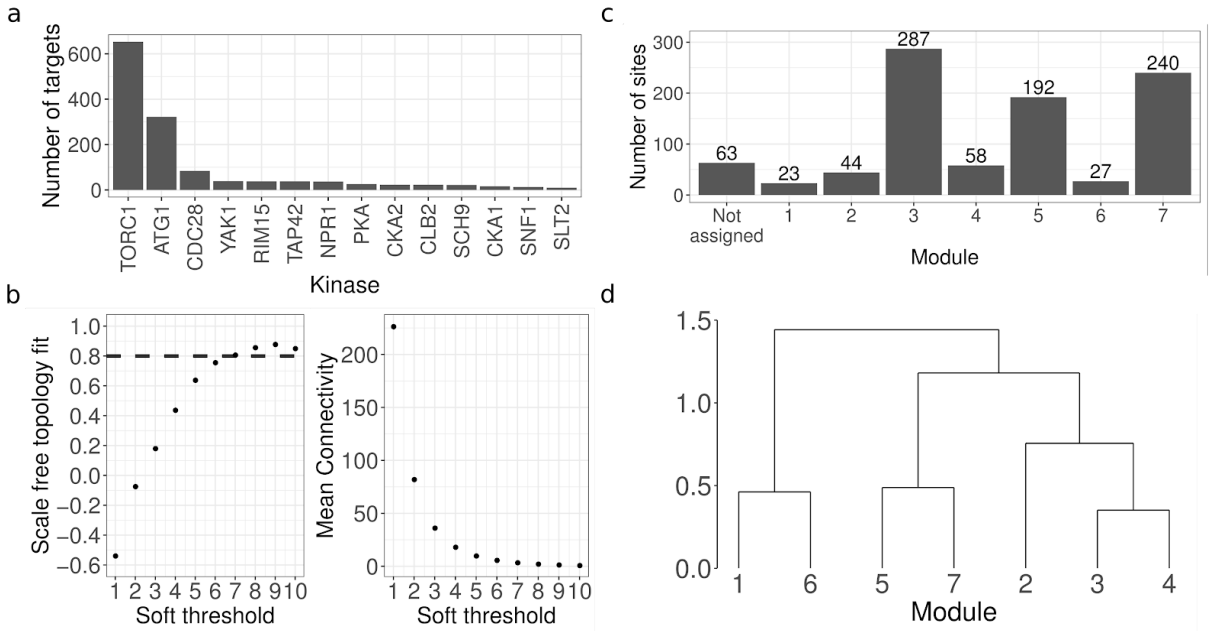
Supplementary Figure 2.4. Treatment specific statistics from differential phosphorylation analysis. a) Barplot showing the number of significant differentially phosphorylated sites in each treatment according to LIMMA linear modeling. Each p-site in each treatment was tested for differential expression against the untreated controls. **b)** Histogram of the number of significant differentially phosphorylated sites in each treatment. **c)** Overlap between 4 groups of treatments in the number of sites regulated in at least 1 treatment in a group (top triangle) or the number of sites regulated in all treatments in a group (bottom triangle). The three groups are defined as the osmotic stresses–NaCl, KCl, CaCl2, and Sorbitol–the high pH stresses–C7, C8, I7, I8–carbon source stress 1–EC, OA, CP, CN, CG, NG, KA, PBS, and SA–and carbon source stress 2–LG, GL, and RF.



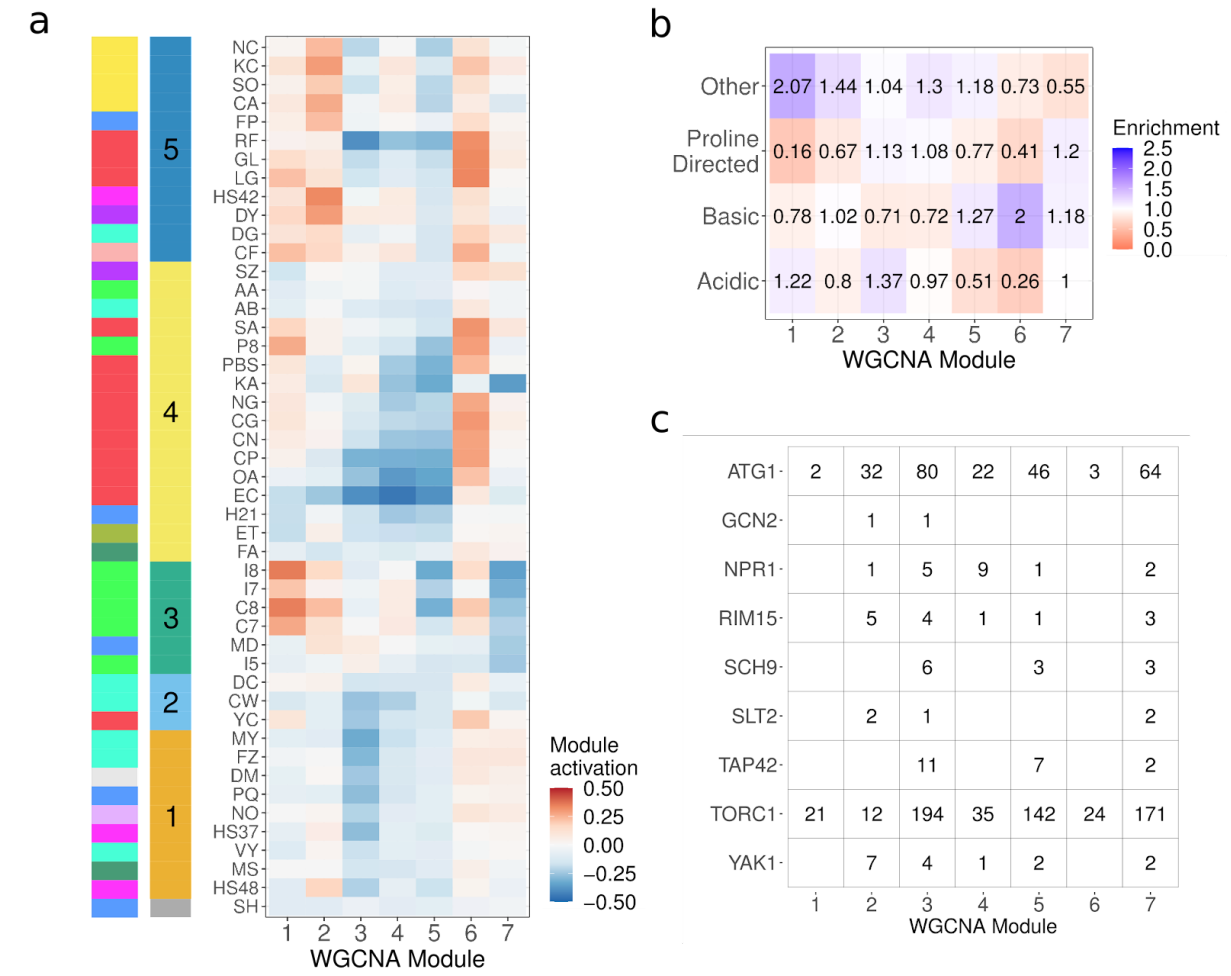
Supplementary Figure 2.5. Analysis of robust differential expression. A selection of robustly regulated sites, i.e sites differentially express in **a)** the osmotic stresses, NaCl, KCl, CaCl2, and Sorbitol **b)** the high pH stresses, C7, C8, I7, I8 **c)** and the carbon source stresses, EC, OA, CP, CN, CG, NG, KA, PBS, SA, LG, GL, and RF. Sites were ordered based on the total number of treatments that they were differentially expressed in. **d)** Histogram of the total number of treatments in which robustly regulated sites are differentially expressed for the 3 groups.



Supplementary Figure 2.6. Global COMBAT correction of dataset batch effects. a) A PCA of phosphosite quantifications per sample colored by sample batch after phosphosites were filtered for missingness and batch effects were corrected out by COMBAT, and **b)** a PVCA demonstrating the relative amount of signal explainable by the treatment applied to cells, the sample batch, and a residual component. Due to the removed batch signal, the PVCA analysis did not converge, which is acceptable as we are mainly focused on demonstrating the results of the correction. **c)** Scatter plot demonstrating the median Pearson correlation between different samples from the same treatment across batches before (imputed data) and after batch correction (corrected data).



Supplementary Figure 2.7. Weighted Gene Correlation Network Analysis (WGCNA) of the TORC1 cascade targets. a) Number of phosphosites out of 934 downstream of each kinase in the TORC1 cascade–TORC1, SCH9, SLT2, YAK1, RIM15, TAP42, NPR1, GCN2, and ATG1–as well as other kinases which were also labeled to at least 10 of the included sites. **b)** Signed R^2 of the scale free topology model fit (left) and the mean connectivity (right) of the underlying adjacency matrix produced by raising the Pearson correlation matrix to a soft threshold power. **c)** Number of phosphosites in the TORC1 cascade assigned to each module discovered by WGCNA. **d)** Hierarchical clustering of the Pearson correlation between modules.

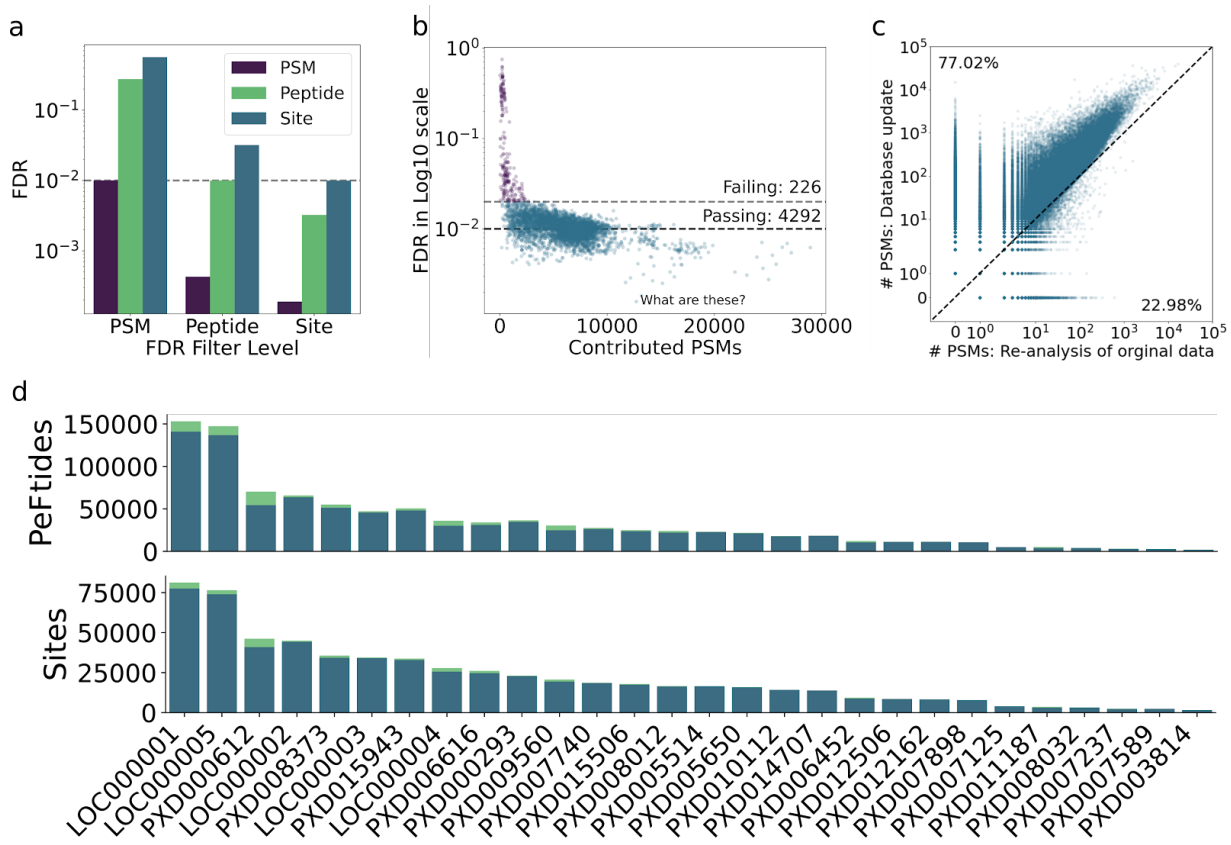


Supplementary Figure 2.8. Analysis of TORC1 Cascade WGCNA modules. a) Relative activation against untreated conditions of the eigensites generated from each module from the WGCNA analysis of the TORC1 cascade. Indicator colors (left 2 columns) of treatment category and treatment cluster are taken from Fig 2.2. **b)** Enrichment of prominent motifs amongst all sites assigned to each WGCNA module. **c)** Number of phosphosites downstream of each kinase in the TORC1 cascade assigned to each WGCNA module.

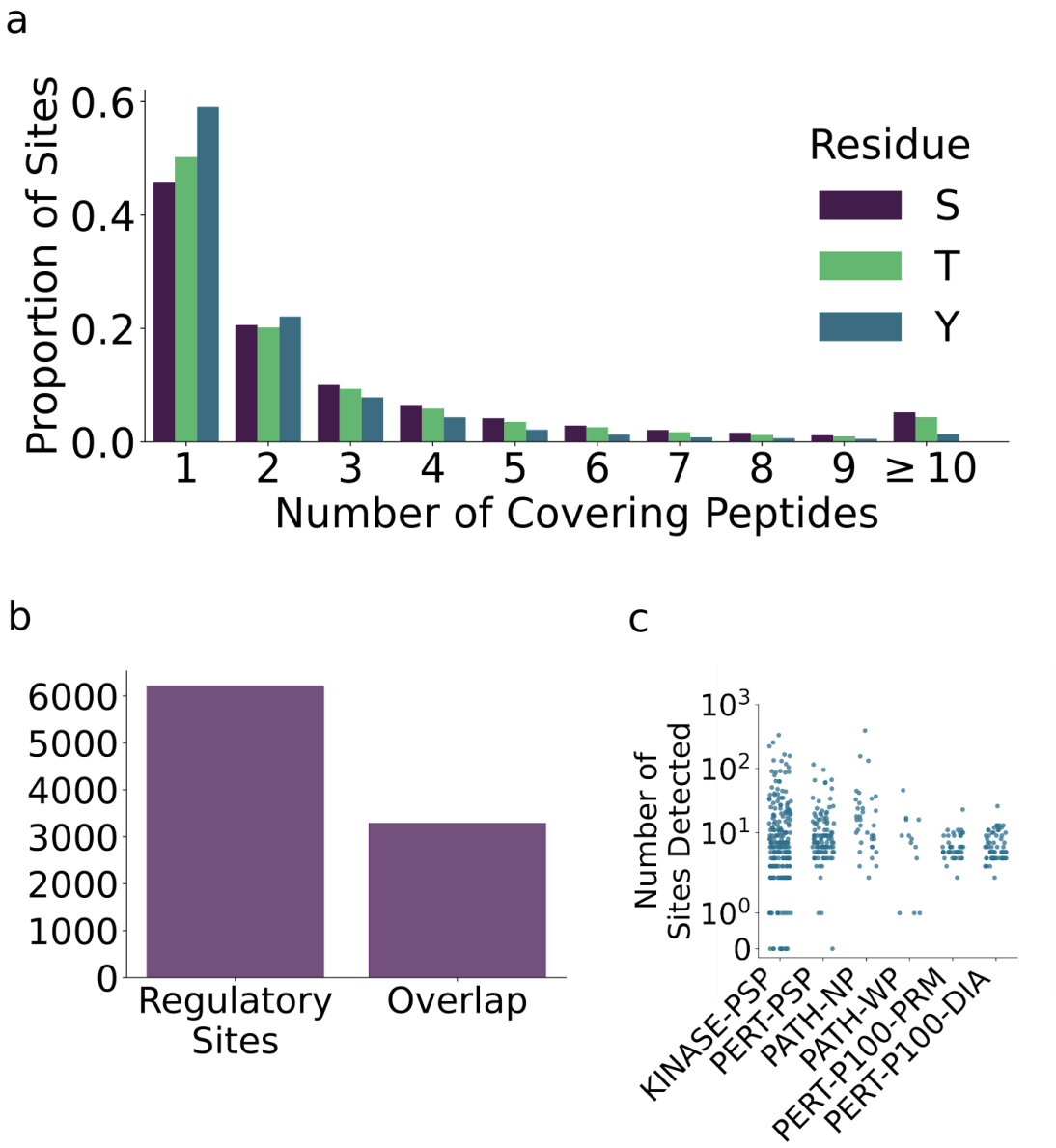
Appendix B

Supplementary information for chapter 4

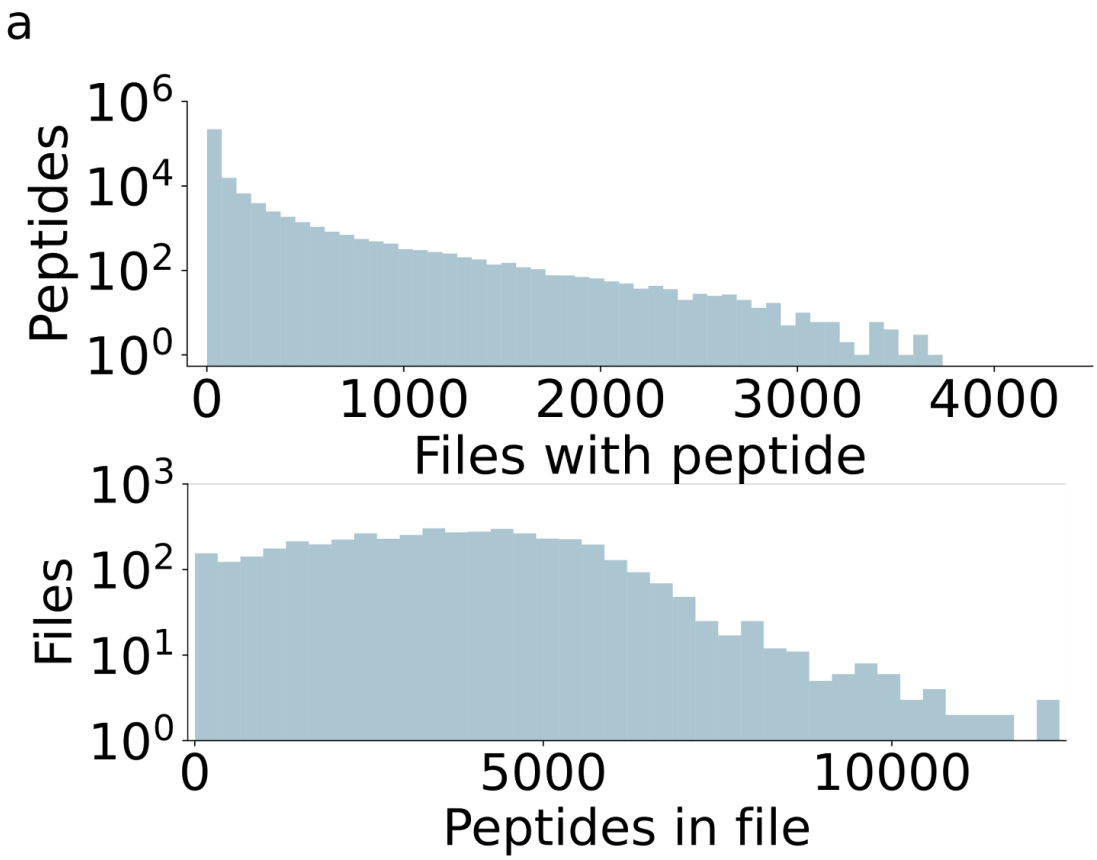
Supplementary Figures



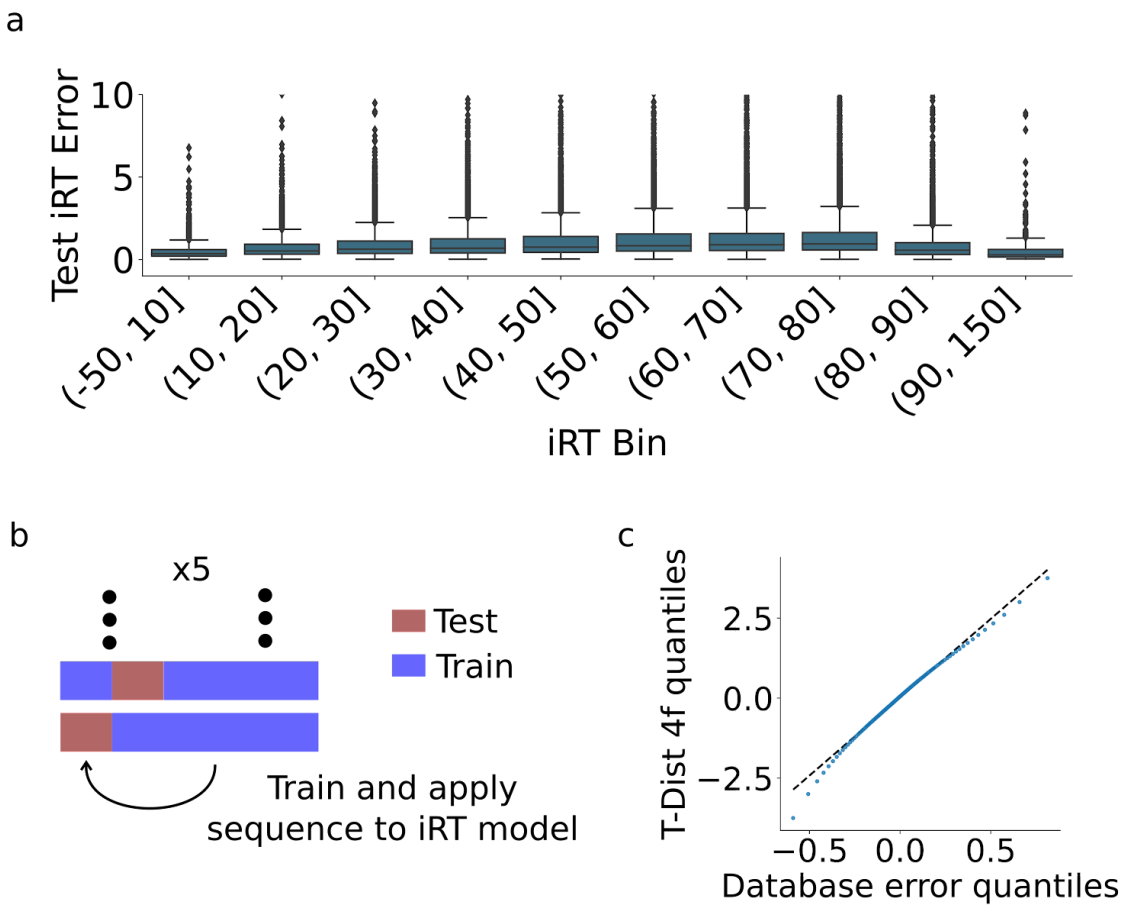
Supplementary Figure 4.1. Overview statistics of the human Phosphopedia 2.0 database. a) Final database FDR at the phospho-PSM, phosphopeptide, and phosphosite level when filtering all levels at the Percolator score corresponding to a 1% FDR at the PSM, peptide, or site level. **b)** True phospho-PSM FDR for individual files calculated as the number of decoy phospho-PSMs vs target phospho-PSMs passing a 1% phospho-PSM FDR calculated across files plotted against the number of phospho-PSMs the file contributes to the final database. **c)** Total number of detections for individual phosphopeptides when reanalyzing the samples in the original phosphopedia with Phosphopedia's new pipeline vs the number of detections with our database update. **d)** Occurrence of phosphopeptides and phosphosites from our database update within individual datasets with colors describing whether those peptides are uniquely found in the given dataset. Datasets labeled with LOCXXXXX were produced by our own lab's instruments, whereas all PXDXXXXX datasets are external data.



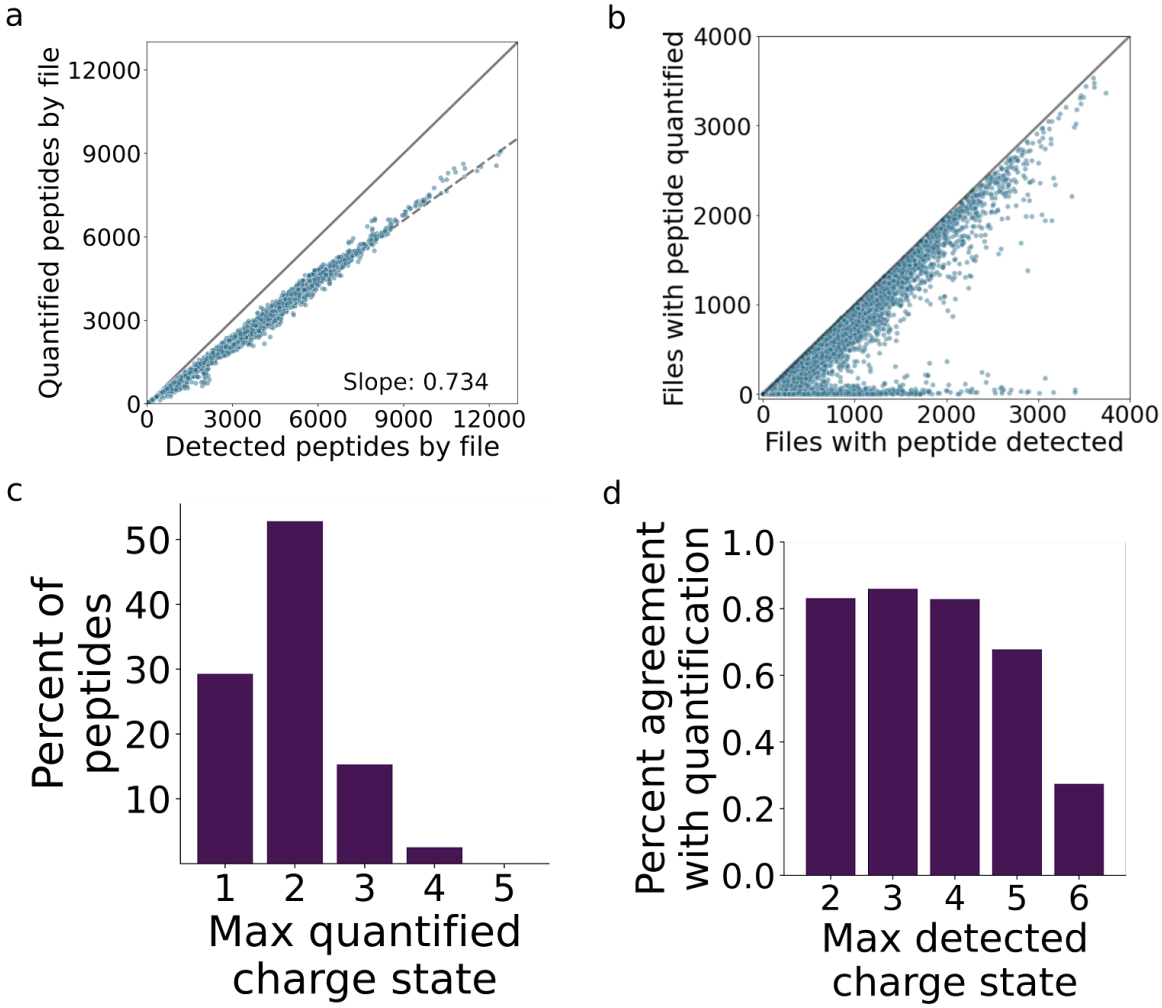
Supplementary Figure 4.2. Phosphosite statistics for the human Phosphopedia 2.0 database. a) Histogram showing the number of unique phosphopeptides in our updated database with at least one phosphorylation event mapping to each phosphosite with colors breaking down the distributions for S, T, and Y sites. **b)** Number of reported human regulatory sites within PhosphositePlus and the number of sites in our update database overlapping with the PhosphositePlus set. **c)** Overlap of sites within the Phosphopedia database and individual curated sets from Krug *et al.* broken down by category.



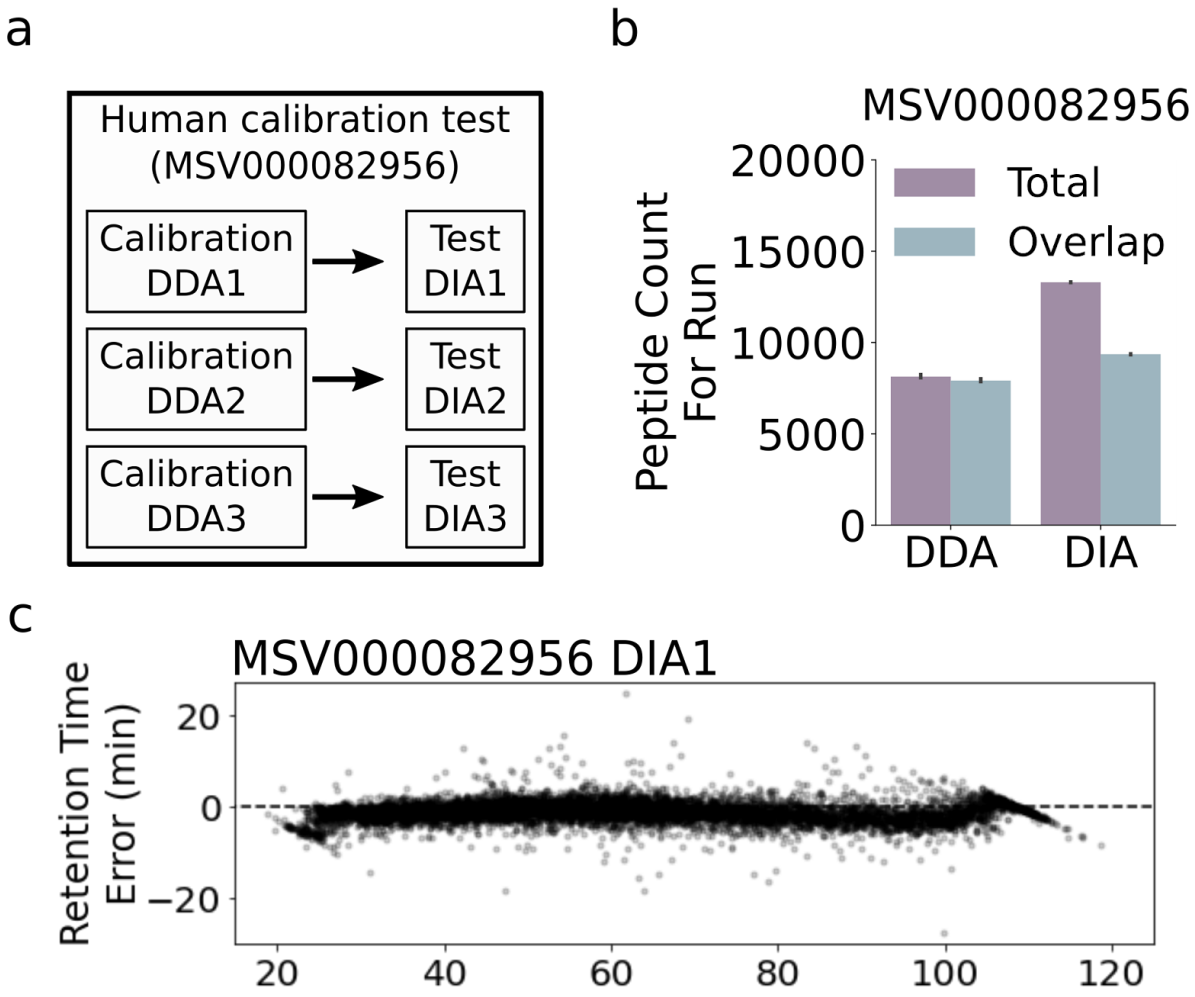
Supplementary Figure 4.3. Reproducibility of phosphopeptide detections. a) Histogram showing number of files that a given database peptide was detected in (top) and a histogram showing the total number of unique peptides per file within the database (bottom).



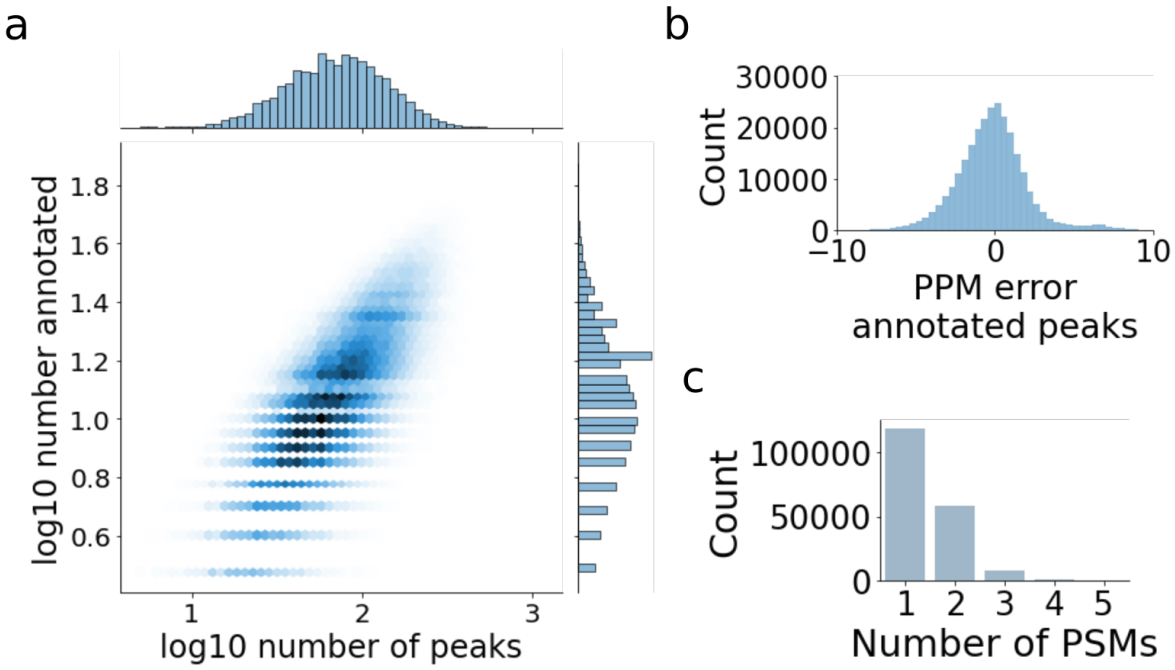
Supplementary Figure 4.4. Evaluating learned retention times in phosphopedia. a) Mean loss for estimated peptide common retention times against their experimentally measured retention times. These are calculated from a model which learns a retention time for each peptide and a non-linear mapping of the common retention time space to each experimental gradient. Errors are calculated for peptides with at least 4 high quality detections in the database based on a 25% held out test set and then binned based on the learned retention time. **b)** Cross validation scheme used for training sequence to iRT multi-layer perceptron models, which is used as the first step in iRT outlier detection in Phosphopedia's pipeline. Briefly, data was split into five chunks and each chunk was held out in turn while an MLP model was trained to predict iRT from the counts of each amino acid in a peptide. The iRT was then predicted for each peptide in the held out data, and the squared error was calculated. Scores from all held out sets were concatenated to achieve predictive errors for all peptides in the database. **c)** Quantile-quantile plot for the iRT predictive errors vs the best fitting theoretical student's t-distribution. In order to find the best fit, the middle 50% of the iRT errors was matched against t-distributions with increasing degrees of freedom.



Supplementary Figure 4.5. Quantification of MS1 intensity across charge states for detections in Phosphopedia 2.0. a) Number of unique phosphopeptides detected in a given file vs the number of phosphopeptides which can be matched against at least 1 Dinosaur MS1 feature. The solid line shows 1:1 while the dashed line shows the line of best fit for the data points. **b)** Number of files where a given database phosphopeptide can be detected vs the number of files where the phosphopeptide can be matched against at least 1 Dinosaur MS1 feature. **c)** Histogram showing the maximum number of charge states across files that a given database phosphopeptide matches to. **d)** Proportion of times that the maximum detected charge state agrees with the max charge state from quantifications. The mode of the max charge state per file for each database phosphopeptide was used as the max quantification charge.



Supplementary Figure 4.6. Validation of Phosphopedia 2.0's statistics on an outside dataset. a) Relationship between DDA samples used to calibrate retention time predictions and DIA samples used for testing predictions against the human phosphoproteome dataset, MSV000082956. **b)** Number of phosphopeptide detections in MSV000082956 after library free search with Spectronaut's library free Direct DIA search, and the overlap of detected phosphopepides. These overlap peptides are specifically used for prediction validation. **c)** Residual errors of phosphopepeptide retention time predictions across the predicted retention time gradient.

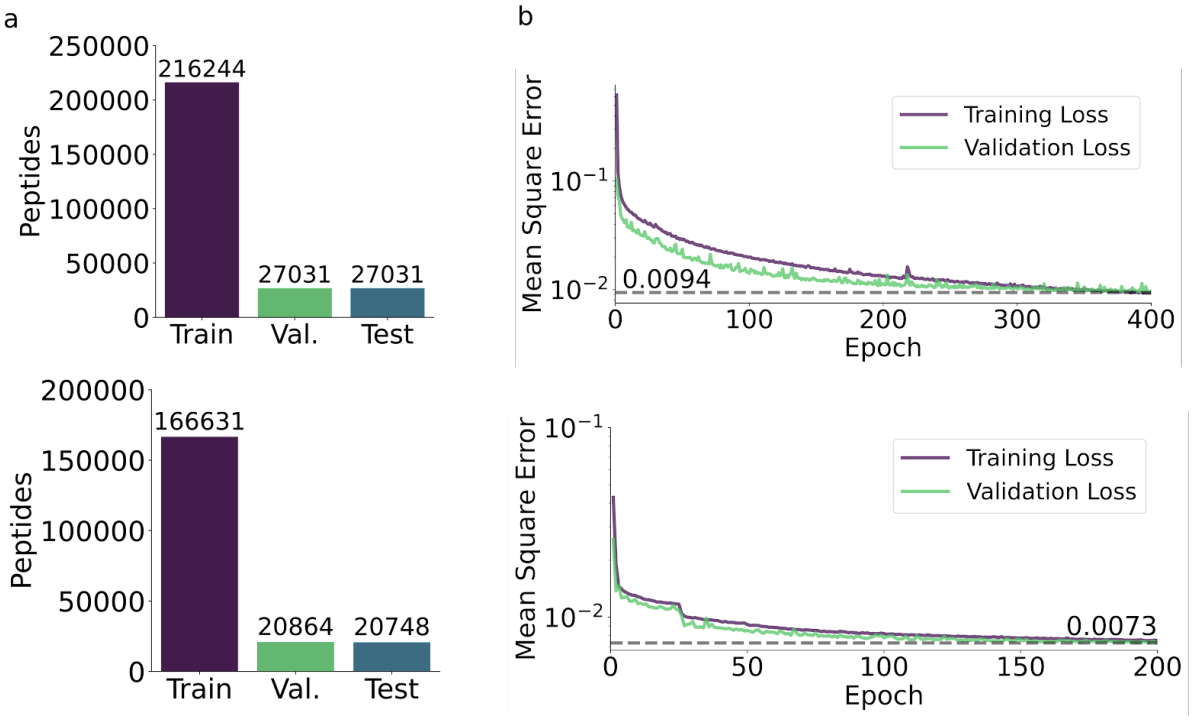


Supplementary Figure 4.7. Building a spectral library from Phosphopedia 2.0. a) A 2 dimensional density plot breaking down the high resolution spectral library generated from Phosphopedia's new pipeline by the number of peaks for each individual PSM as well as the number peaks with a b or y ion annotation. **b)** The median PPM error of the theoretical masses of the annotated peaks within each PSM to their experimentally measured masses. **c)** The final number of PSMs per peptide within the high resolution spectral library.

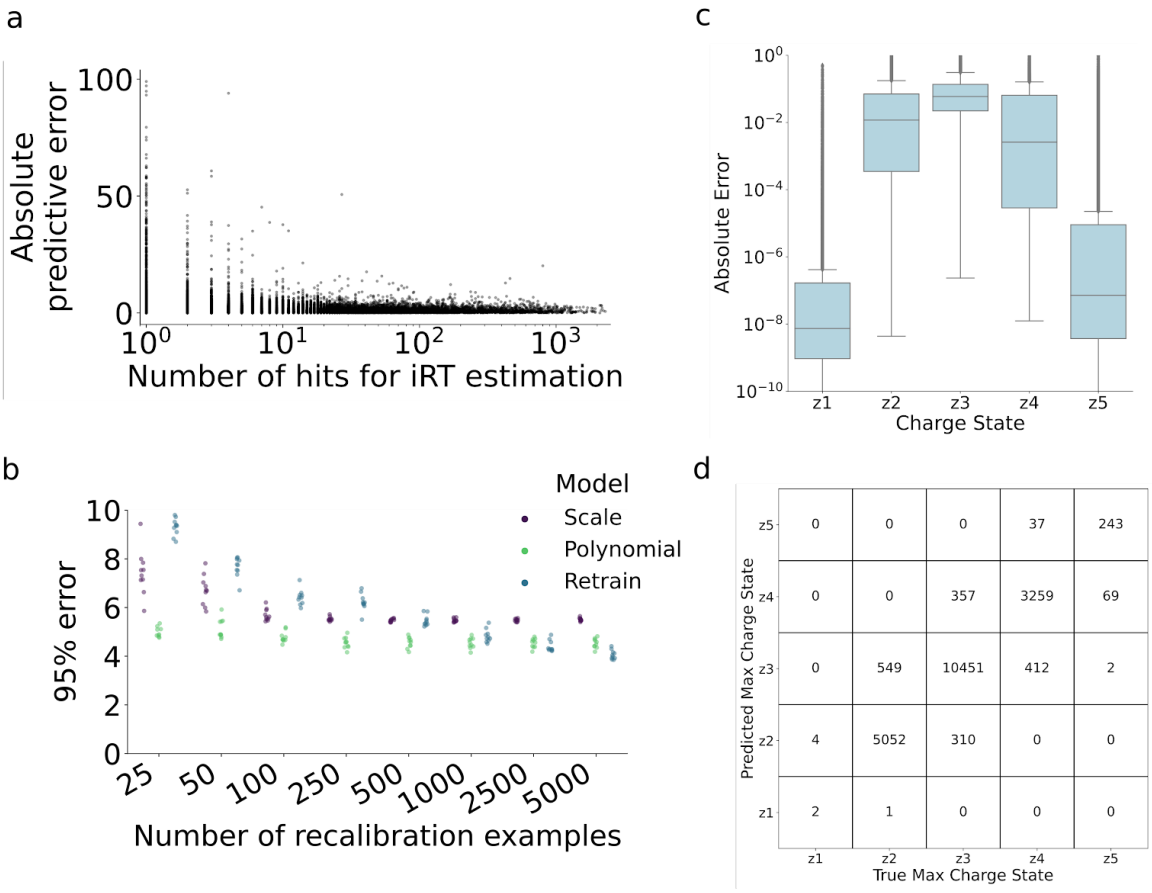
Appendix C

Supplementary information for chapter 5

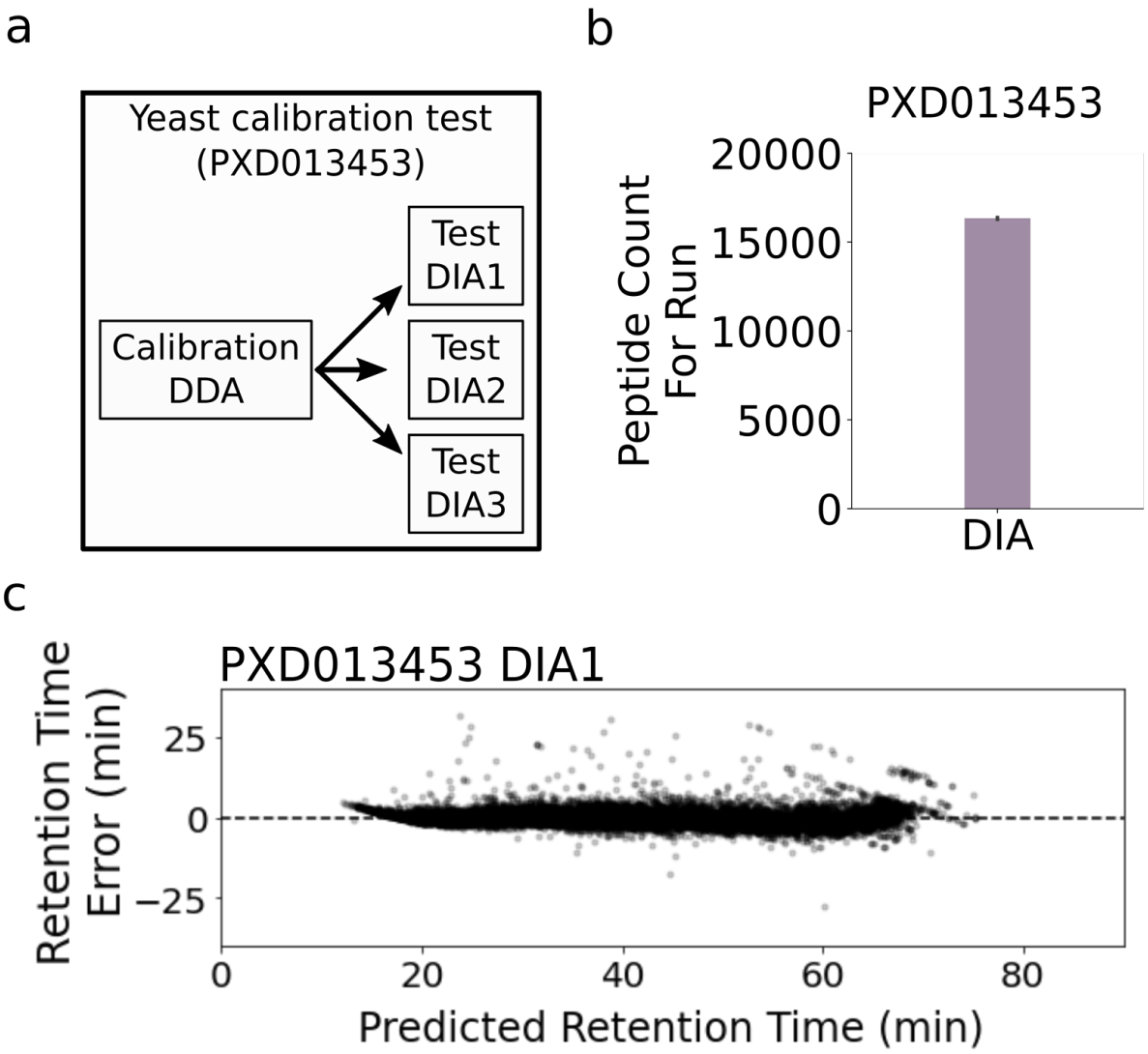
Supplementary Figures



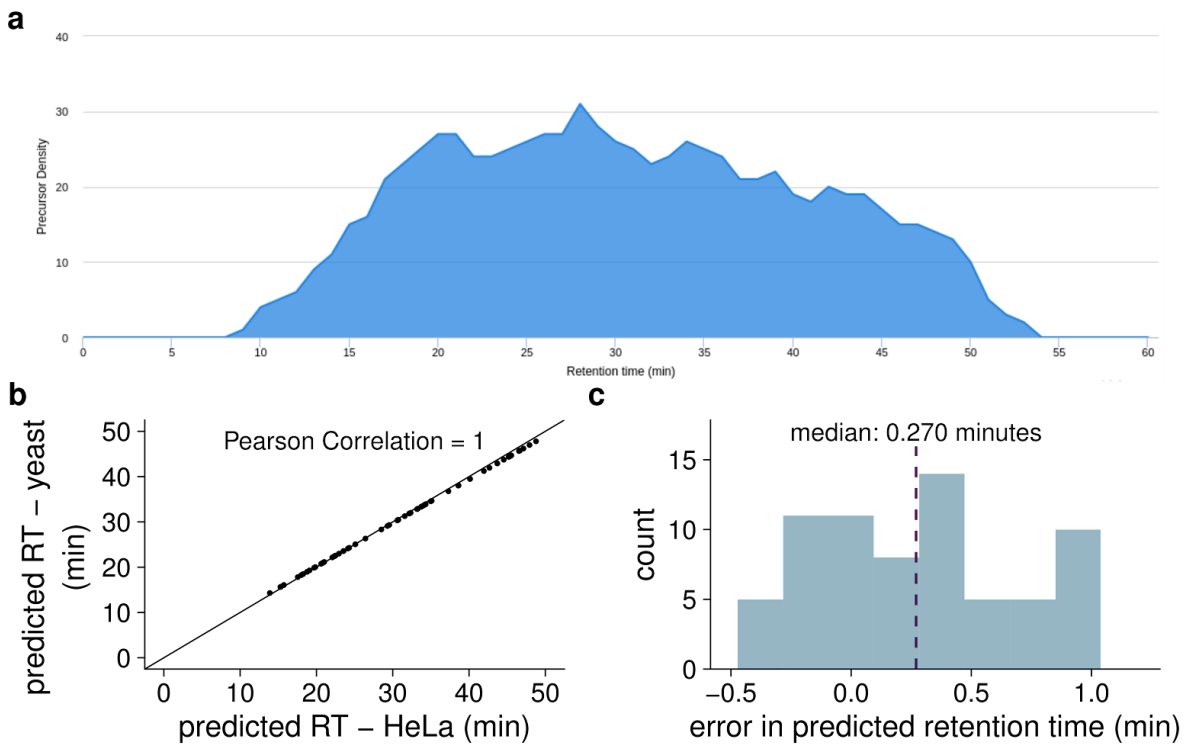
Supplementary Figure 5.1. Training of phosphopeptide retention time and charge state models. **a)** Number of database phosphopeptides in the training, validation, and test sets for retention time prediction (top) and charge state prediction (bottom). The charge state sets of phosphopeptides are subsets of their respective retention time sets, with difficult to quantify phosphopeptides removed. The training set for retention time prediction was further filtered to focus on high quality training examples. **b)** Loss curves for the training of gated recurrent unit neural networks for the prediction of retention time (top) and charge state (bottom).



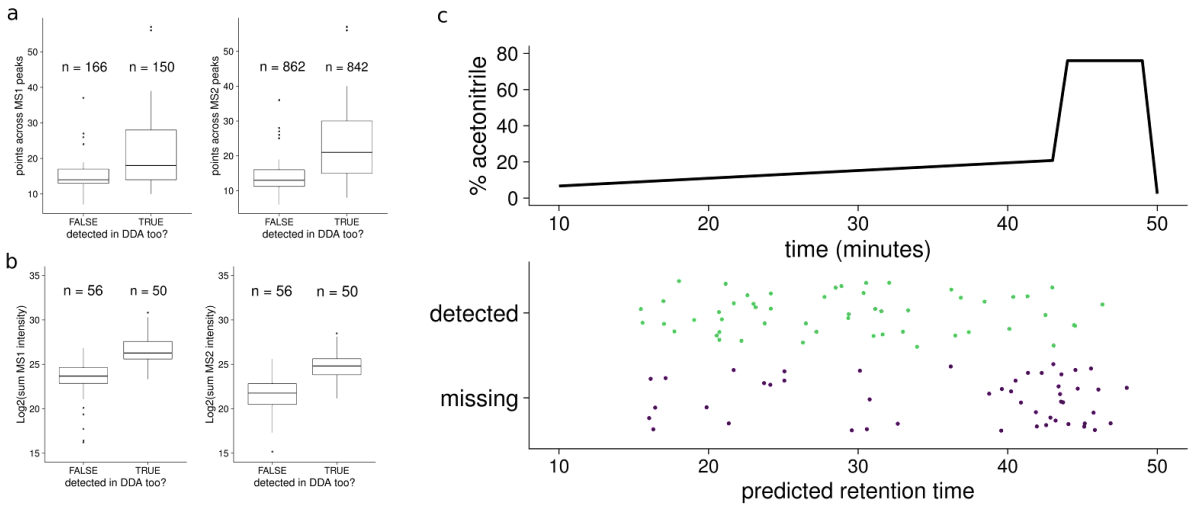
Supplementary Figure 5.2. Evaluation of Phosphopedia 2.0's phosphopeptide property models. **a)** Improvement of absolute predictive error for Phosphopedias retention time predictor with increasing number of database examples used to estimate common scale retention times. Each data point represents one Phosphopeptide from the test set. **b)** 10-fold cross validation for 3 types of models used to calibrate Phosphopedia's retention time predictions to a new *S. cerevisiae* IMAC sample analyzed with a 1.5 hr gradient. CV splits were the same for each model, including the increasing number of calibration peptides that were taken from the training data. **c)** Absolute error in predicted charge state proportion vs actual measured charge state proportion for phosphopeptides in the test set. **d)** Confusion matrix showing the predicted max charge state as determined by the GRU charge state predictor vs the true max charge state as determined by Dinosaur MS1 features.



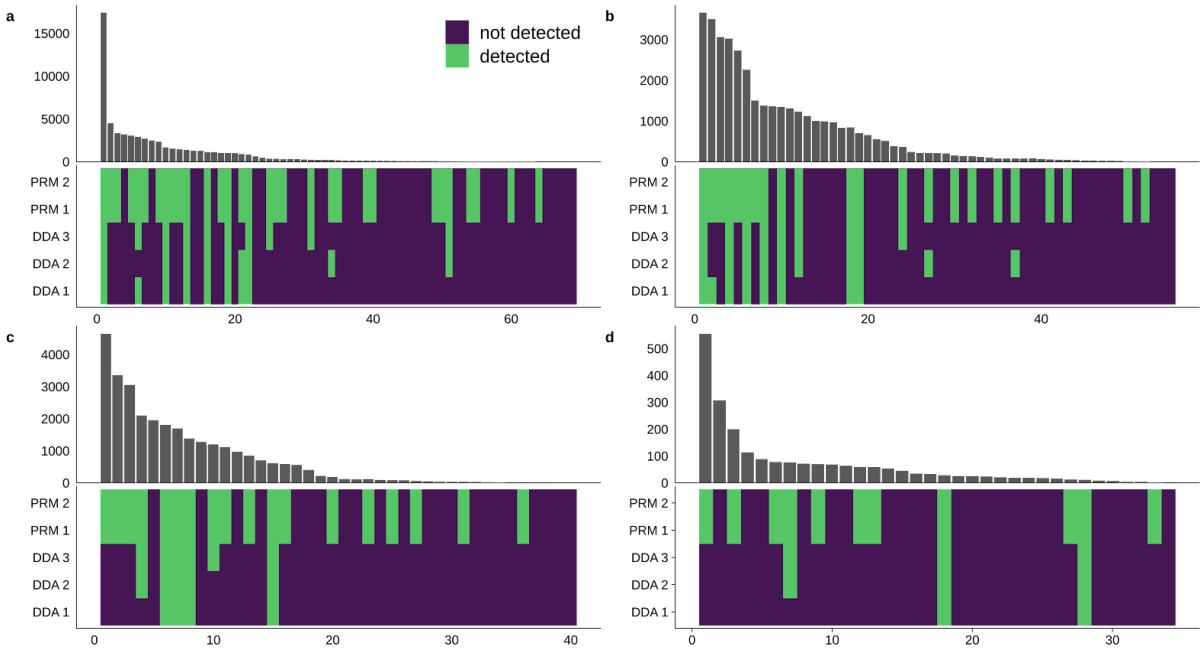
Supplementary Figure 5.3. Validation of phosphopedia's phosphopeptide property models on an outside dataset. a) Relationship between DDA sample used to calibrate retention time predictions and DIA samples used for testing predictions against the yeast phosphoproteome dataset, PXD013453. **b)** Number of phosphopeptide detections in PXD013453 after library free search with Spectronaut's library free Direct DIA search. Since the deep retention time predictor can provide a theoretical retention time based solely on sequence, all phosphopeptides can be used for calibration and testing. **c)** Residual errors of phosphopepeptide retention time predictions across the predicted retention time gradient.



Supplementary Figure 5.4. Core steps in building targeted phosphoproteomic runs with Phosphopedia 2.0. a) Elution profile for custom PRM assay 1 which contained X targets with 5 minute windows and was calibrated to an initial DDA human phosphoproteome run **b)** Correlation of the predicted retention time for assay 1 generated by calibrating to either a DDA human phosphoproteome run or a yeast total proteome. **c)** Difference in predicted retention times for assay 1 generated by calibrating to a DDA human phosphoproteome run vs the same assay calibrated with a yeast total proteome.



Supplementary Figure 5.5. Results of validation PRM runs for Phosphopedia 2.0. a) Number of measured points across precursor ion peaks (left) and product ion peaks (right) within peak boundaries defined at peptide level are separated by their intersection of MS acquisition methods. Points across peaks data is sourced solely from PRM acquisition, from technical duplicates. Dark lines within box plots represent median points across ion peaks, bottom and top of boxes represent 25th and 75th quantiles, length of whiskers represents 1.5 times the IQR, points represent outliers and sample size (n) refers to numbers of individual ions, counting replicates separately. **b)** Area under precursor ions peaks (left) and product ion peaks (right) were summed for each peptide to yield relative MS1 and MS2 peptide intensities, respectively. Boxplots are separated based on whether peptides were detected solely in PRM acquisition or in both PRM and DDA acquisition. Dark line within box plots represents median of the log2-transformed summed intensities, bottom and top of boxes represent 25th and 75th quantiles, respectively; length of whiskers represents 1.5 times the interquartile range (IQR), points represent outliers and sample size (n) refers to numbers of peptides with summed peak areas. Replicates are counted separately because area under curves may differ between replicates. **c)** Chromatography elution profile showing percentage acetonitrile used to elute peptides zoomed into time when peptides elute (top). Detected (green) and missed (purple) targets are plotted with respect to predicted retention time (bottom). Points are spread vertically only to aid visualization. Acetonitrile changes slope at 43 min, 44 min, 49 min and 50 min; times 0 – 10 min and 50 – 60 are not shown.



Supplementary Figure 5.5. Target recall from PRM versus DDA acquisition relative to database spectral matches. Peptide targets are ranked by their associated spectral count in the Phosphopedia 2.0 database (x-axis in all plots). Targets successfully detected from PRM technical duplicate injections or DDA technical triplicate injections are colored green while missed targets are colored purple. The y-axis of all bar plots represent the summed spectral counts from Phosphopedia 2.0 for precursors charged +2, +3 and +4 for each target peptide. Total PSMs for charges +2, +3, +4 were summed for bar plot; however, assays targeted precursors with only a charge +2, +3 or +4 without precursor sequence redundancy. The y-axis for the heatmap indicates mass spectrometry acquisition method and technical replicate number.

a. Assay 1 targeted 69 peptides from the Phosphopedia 2.0 Cell Growth Control assay.
b. Assay 2 targeted 55 peptides focused on Akt signaling.
c. Assay 3 targeted 40 peptides from the Phosphopedia 2.0 Transcriptional Control assay.
d. Assay 4 targeted 34 peptides containing only phosphorylation of tyrosine residues on tyrosine kinase activation loops.