Developing Proteomic Methods to Assay Function of Proteoforms

Ian Robert Smith

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

Devin Schweppe

Shao-En Ong

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Developing Proteomic Technologies to Assay Function of Proteoforms

Ian Robert Smith

Chair of the Supervisory Committee:

Judit Villén

Department of Genome Sciences

The human genome encodes a repertoire of ~20,000 proteins. The unique functional roles of these proteins diversify through mechanisms of allelic variation, post-translational modifications (PTMs), alternative splicing, and protein cleavages. The modified versions of a protein, or proteoforms, comprise a vast collection of molecules originating from a single protein-coding gene. Proteoforms can perform different functions and warrant extensive characterization for their involvement in normal cell function and disease. Modern advances in genomic sequencing and mass spectrometry (MS) have accelerated our capacity to identify proteoforms at scale. However, a vast majority of proteoforms lack functional annotation. In this thesis work, I developed novel MS-based proteomic methods to assess the impact of PTMs, missense mutations, and protein cleavages on protein function at scale. I chose to measure thermal stability and turnover of proteins to probe functional differences caused by protein modifications.

Measuring these protein properties is feasible at the proteome scale and they encapsulate many generalized functions of PTMs, mutations, and protein cleavages. To capture proteoform functions, I performed our molecular selections at the protein-level, thus retaining proteoform-specific turnover and stability information and MS readout at the peptide-level. I identified functional proteoforms by comparing proteoform-specific peptides to their cognate peptides from their unmodified proteoform. For the PTM phosphorylation, we identified 253 and 68 phosphorylated proteoforms that alter protein turnover or thermal stability in the yeast proteome, respectively. For protein cleavages, we implemented protein turnover and thermal stability assays to identify and functionally characterize substrates of the NSP5 protease from SARS-CoV-2. For protein mutations, we piloted a peptide barcoding method to represent each protein variant with a short peptide and to enable pooled functional screens for protein variant libraries by MS. Lastly, my thesis work expanded to implementing novel MS acquisition strategies and developing software to improve peptide detection and quantitative precision. Collectively, these high-throughput proteomic methods enabled biochemical interrogation of proteoforms, accelerating the functional characterization of the modified proteome.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I want to express my deepest gratitude for the amazing people in my life who have supported me in the completion of my degree.

First, I would like to thank my advisor Dr. Judit Villén, who without her I would not be the scientist I am today. I am grateful to have had the opportunity to work for an incredibly brilliant and creative scientist as well as a great person. In the lab, I am grateful for her time and guidance creating a great research program for me, teaching me how to effectively communicate my science, and how to think critically about my projects. I appreciate the scientific freedom she has offered me to explore projects I am excited and passionate about. I am thankful for Judit's support through the times when graduate school was particularly difficult for me. Also, I greatly appreciate the mentorship regarding my career and guiding me towards my long-term career goals.

Also, I am thankful for the collaborative environment in the Department of Genome Sciences, and I have enjoyed my interactions with faculty, students, and staff over the years. I am grateful to the members of my committee who have offered valuable guidance and expertise to my projects and advice and support towards my career goals.

Throughout my time in graduate school, I have been incredibly fortunate to work with an amazing team of talented and kind scientists in the Villén Lab. I really enjoyed my experiences in the lab and will greatly miss the incredible individuals who have made my experience here in Seattle so fun. I particularly want to thank Ricard Rodriguez for our exciting discussions over the years and for your mentorship. Also, I want to thank Jimmy Eng who has been a great mentor to me, and I truly enjoyed our collaboration on projects over the years. Also, I particularly want to

acknowledge Bianca Ruiz, Mario Leutert, Kyle Hess, Anthony Barente, and other Villén lab colleagues for their support and guidance with my projects and for the fun times. In the lab, I have had the blessing of having truly incredible friends and I am so thankful for you all. I will miss you all dearly.

Lastly, I could not have completed graduate school without the loving support from my family. I want to thank my mother and father for always loving and supporting me. When times are good or bad, you both have always been there for me with a smile and a bright perspective. Thank you for everything you do for Evan and I to allow us to achieve our goals. I want to thank my brother, Evan, for always reaching out and supporting me since we were young. I want to thank the Ruiz family who have become a second family to me during my time in graduate school. To my partner Bianca, thank you so much for your love and support every step of the way. Your caring heart, laughter, and positivity have carried me through graduate school and I look forward to exploring our next adventures together. Also, Clifford, you are truly the best and thanks for bringing us so much happiness.

Chapter 1. INTRODUCTION

From the Human Genome Project, a genetic architecture composed of ~20,000 protein-coding genes was reported to blueprint all cellular functions[1]. Regulation during transcription and translation underlie gene expression control to generate protein molecules which act as the cell's basic functional units[2]. Composed of an evolutionarily-tuned sequence of amino acid building blocks, protein molecules fold into complex structural elements to inform function. Despite a discrete number of protein-coding genes, the functions performed in cells greatly outnumber the number of genes[3]. Also, individual protein-coding genes can perform 10's to 100's of functions making the functional inference for each protein elusive. Over the years, we have uncovered that mechanisms to diversify protein sequence and structure can help inform the link between a single gene and the myriad of molecular phenotypes.

Molecular information encoded within a single gene is greatly expanded through allelic variation, alternative splicing, protein cleavages, and post-translational modifications (PTMs), yielding different proteoforms[4]. To date, researchers have been able to catalog 1,000,000s of proteoforms in humans[5], however the impact of most of these variations on protein function is limited. While genomic technologies have provided evidence for many proteoforms, protein technologies that detect and assay proteoforms directly will be essential to infer molecular functions. The objective of this thesis is to develop novel protein technologies addressing the functional roles of proteoforms.

1.1 MEASURING PROTEOFORMS AT SCALE BY MASS SPECTROMETRY

Mass spectrometry (MS) has emerged as the leading technology to identify proteins and proteoforms at scale, or proteome-wide[6]. Ideally, measuring full-length proteoforms by mass

spectrometry, or top-down proteomics[7], is preferred, however top-down approaches currently lack the throughput required to identify 1,000,000's of proteoforms. By intentionally digesting proteins to peptides, we can identify 10,000's of proteoforms in a single experiment via a signature mass difference from a protein's unmodified peptide mass. With this bottom-up proteomics[6,8] approach, proteoforms are fragmented and require error-prone protein inference[9] to reconstruct fully-length proteoforms. However, proteoform-specific peptides can serve as molecular barcodes for all full length protein molecules of that proteoform, termed peptidoform[10]. We can compare a proteoform-specific peptide to the matching unmodified peptide that represents the canonical proteoform to elucidate differences between proteoforms (Figure 1.1: demonstrates phosphorylated proteoform comparison). By performing molecular selections at the protein-level then digesting the proteome into peptides, biological information is embedded at the peptide-level. Functional comparisons between proteoforms can thus be performed at the peptide-level with proteoform-specific peptides following protein-level selections.

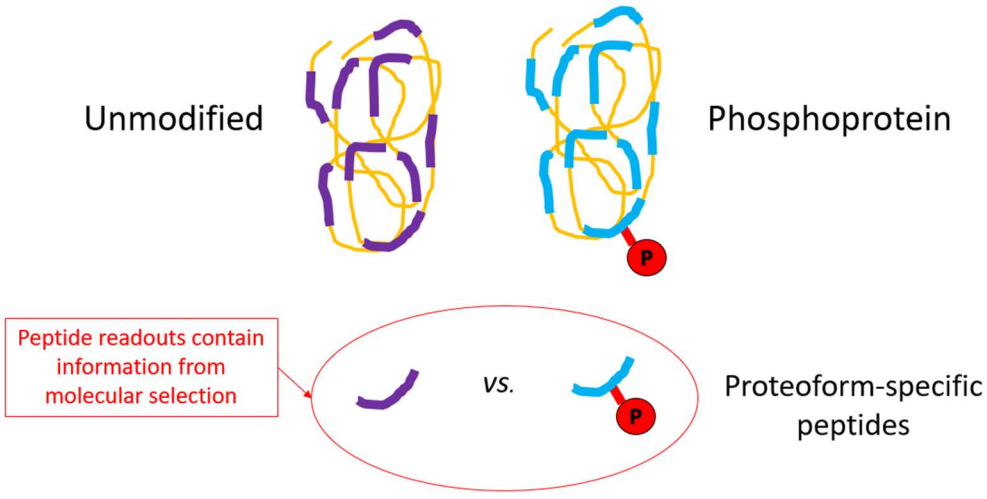


Figure 1.1: Comparing proteoforms with a phosphorylated proteoform example. In a bottom-up proteomics approach, proteoform-specific peptides can serve as molecular identifiers for molecules that contain the proteoform modification. Comparing proteoform-specific peptides with biological readouts can proxy functional differences due to a proteoform difference.

1.2 PROTEOFORMS MODULATE PROTEIN FUNCTION

To annotate proteoform functions, we need informative MS-compatible assays to probe biological properties that reflect protein molecular functions. To determine the assays to use, we must understand the variety of different functions proteins can have. For instance, many proteins' functional role in the cell is to act as enzymes that catalyze biochemical or metabolic reactions. Also, proteins provide structure to cells, respond to external stimuli, conduct cellular transport, and interact with other proteins to assemble complexes for complex tasks[11]. Some protein-coding genes can perform multiple functions, while others only perform a function in certain contexts.

Proteoforms are thought to serve as modulators that diversify protein-coding genes' many protein functions. For instance, post-translational modifications, such as phosphorylation or ubiquitination, can modulate protein activity, alter protein binding affinity, elicit subcellular localization change, and promote protein degradation[12,13]. Also, PTMs can act as molecular switches[14] to turn a protein's function "on" or "off". This switch behavior enables rapid protein to protein communication to quickly propagate signals and modulate protein functions in response to cellular stress. Allelic variation and missense protein variants, like in p53, have been observed to result in protein loss-of-function or act in a dominant negative behavior with new non-canonical functions[15]. Protein cleavages can alter protein subcellular residency by eliminating peptide transit sequences or promote protein destabilization and degradation[16–18]. Natural proteoforms are essential for cellular homeostasis and phenotypes, while aberrant regulation or presence of specific proteoforms can result in disease phenotypes[5].

To identify proteoforms that drive protein function, biological and biochemical selections must be designed to stratify the different proteoforms in their different functional states. The

optimal selection assays must be easy to analyze proteome-wide by mass spectrometry and encompass a broad range of molecular functions of proteoforms.

1.3 PRIORITIZING PROTEOFORMS FOR FUNCTION BY MS

Herein, we have identified the biological and biochemical properties of protein thermal stability and protein turnover as valuable proxies to prioritize proteoforms for function. We propose that these properties can unbiasedly assess functional differences driven by proteoforms, including but not limited to the PTM phosphorylation, protein cleavages, and missense mutations.

1.3.1 PROTEIN THERMAL STABILITY

In order to assay protein thermal stability proteome-wide, Savitski *et al.* developed a method called Thermal Proteome Profiling (TPP)[19], which couples a thermal shift assay with mass spectrometry. TPP has been applied to identify protein-small molecule/drug interactions[20], protein-protein interactions[21], and subcellular localization changes[22]. We can utilize thermal stability assays, similar to TPP, to probe thermal stability differences between a modified proteoform and its unmodified proteoform counterpart. We anticipate a MS thermal stability assay can identify proteoform events that modulate protein interactions and alter protein conformational stability.

When applied to PTMs, we expect to identify substrates of phospho-binding proteins (e.g. 14-3-3) that will be more thermally stable upon binding[23]. We will also identify phosphorylation events that induce long-range conformational changes, like observed in glycogen phosphorylase[24]. Finally, phosphorylation commonly targets intrinsically disordered

regions[25], altering the order/disorder properties and likely the protein's thermal stability. For protein cleavages, altered peptide-level thermal stability readouts across a discrete breakpoint along the length of a protein sequence could denote a protein cleavage event. Missense variants with altered thermal stability could implicate a mutation altering protein interactions or structural stability.

1.3.2 PROTEIN TURNOVER

Protein turnover is the balance between protein synthesis and degradation, and helps maintain properly functioning proteins. The combination of dynamic SILAC (Stable Isotope Labeling with Amino Acids in Cell Culture) with mass spectrometry allows to measure protein turnover proteome-wide[26,27]. Our group has recently found that protein-interactions slow protein turnover, while degradation motifs, and protein activation accelerate protein turnover[28,29]. Since phosphorylation can facilitate these three functions, we reason that we could identify functional phosphosites by measuring protein turnover differences between a phosphoprotein and its unmodified counterpart. Additionally, protein cleavages have been observed to destabilize proteins promoting degradation[30], and altered protein turnover readouts surrounding a breakpoint could suggest protein cleavage[31]. Lastly, mutations have also been observed to alter protein stability resulting in decreased steady state abundance, likely due to decreased stability and increased protein degradation[32]. Protein variants with altered turnover expressed under the same promoter could indicate missense mutations that alter protein degradation and thus increase protein turnover. Molecular functions related to protein interactions and protein destabilization could explain observed changes in proteoform protein turnover.

1.4 Organization of Dissertation

Following this introductory chapter (Chapter 1), the thesis is composed of several mass spectrometry-based proteomic technologies to assay proteoform's impact on protein function. First, protein thermal stability (Chapter 2) and protein turnover (Chapter 3) are leveraged to prioritize phosphorylation events that likely impact protein function. In Chapter 4, protein turnover and thermal stability assays are used to identify protein cleavage events proteome-wide, piloted with the overexpression of SARS-CoV-2 NSP5 protease in HEK293T cells. Chapter 5 outlines the strategy and initial progress in the development of a peptide barcoding approach to functionally characterize protein variants in parallel by mass spectrometry. To improve our functional assays, the final project describes a novel MS method to coisolate and fragment peptide pairs for peptide-spectral matching and MS/MS-based quantification (Chapter 6). Finally, the thesis is summarized in Chapter 7 with concluding remarks and implications of this work for the field of protein biology.

Chapter 2. IDENTIFICATION OF PHOSPHOSITES THAT ALTER PROTEIN THERMAL STABILITY

This chapter is adapted from the following work:

Ian R. Smith, Kyle N. Hess, Anna A. Bakhtina, Anthony S. Valente, Ricard A. Rodríguez-Mias and Judit Villén. Identification of phosphosites that alter thermal stability. *Nature Methods*, 18:760-762, 2021.[33]

2.1 PREFACE

Post-translational modifications (PTMs) act as dynamic regulators of cellular function. More than 200 different PTMs have been described with phosphorylation being the most studied, having over 200,000 phosphorylation sites in the human proteome[34]. Despite extensive mass spectrometry efforts, only 2.9% of observed post-translational modifications have functional annotation. Computational predictions of functional phosphosites have used criteria like conservation, localization on protein interfaces or hot spots, and presence of multiple nearby PTMs[35]. However, our group observed that the conservation criterion is limited with most phosphosites evolving rapidly and many "young" sites being functional[36]. Because computational approaches are limited in accuracy, experimental approaches are preferred.

For experimental validation of phosphosites, the gold standard method is to mutate the site to a phospho-mimetic (Glu/Asp) and a phospho-inhibitory (Ala) amino acid and assay the mutants for gain or loss of function[37]. This approach is labor intensive and does not scale for functionally annotating 100,000s of PTMs. Therefore, the field requires new experimental methods that can annotate the function of PTMs at a proteome-scale. In Chapter 2 and 3, we

address this major bottleneck by developing proteomic methods to measure the effects of phosphorylation on protein properties that associate with function. Among these properties, we explore protein thermal stability (Chapter 2) and protein turnover (Chapter 3) to prioritize phosphorylation events that play a functional role on proteins.

2.2 MAIN

Proteomic methods have enabled the discovery of 100,000s of protein phosphorylation sites[34], however we lack methods to systematically annotate their function. Phosphorylation has numerous biological functions that biochemically involve changes in protein structure and interactions. These biochemical changes can be detected by measuring the difference in thermal stability between the protein and the phosphoprotein. Building on recent work, we present a proteomic method to infer phosphosite functionality by reliably measuring such differences.

Recently, Huang et al.[38] developed the Hotspot Thermal Profiling (HTP) method to identify phosphosites that alter protein thermal stability, reporting 719 out of 2,883 (25%) measured human phosphoisoforms with significant effects. The reported melting temperatures (T_m) for phosphoisoforms and their corresponding proteins showed low correlation ($R^2 = 0.18$) (Figure 2.1a), implying that phosphorylation largely functions by structurally reshaping the proteome. However, the poor T_m reproducibility between replicates (mean pairwise $R^2 = 0.42$ for proteins and 0.21 for phosphoisoforms, Figure 2.1b) suggests that the large effects observed may be due to technical variation, given that critical steps of the method are conducted separately for phosphopeptides and proteins and for the different temperature sample points (Appendix A Supplementary Discussion).

The HTP workflow consists of phosphopeptide enrichment followed by separate isotopic labeling and mass spectrometric analysis to derive T_m values for phosphoisoforms and proteins, respectively. To assess technical variation between the two samples, we used unmodified peptides identified in the phosphopeptide samples, which are expected to yield the same T_m as the protein (Appendix A Supplementary Discussion). Disconcertingly, our re-analysis revealed significant stability effects on 1,223 out of 3,074 (40%) of the co-enriched unmodified peptides, a larger percentage than phosphoisoforms (29%, 774 out of 2,656 in our reanalysis), calling into question the reported hits (Figure 2.1c, Dataset S1). Additionally, the T_m correlation of these unmodified peptides measured in the phosphopeptide samples with their protein was similarly low ($R^2 = 0.18$) to the T_m correlation between phosphoisoforms and their proteins (Extended Data Figure 2.1a). This suggests that the separate enrichment and independent labeling prior to mass spectrometric analysis of peptide and phosphopeptide samples could have introduced substantial technical error precluding the comparison between phosphoisoform and protein, and perhaps that the reported hits arise from a lack of stringency in the applied statistical analysis. Indeed, we show that reimplementing the t-test considering unequal variances for proteins and phosphoisoforms, consolidating technical (mass spectrometry reanalysis) replicates, and applying multiple testing correction dramatically decreases the number of significant hits (Extended Data Figure 2.2, Appendix A Supplementary Discussion).

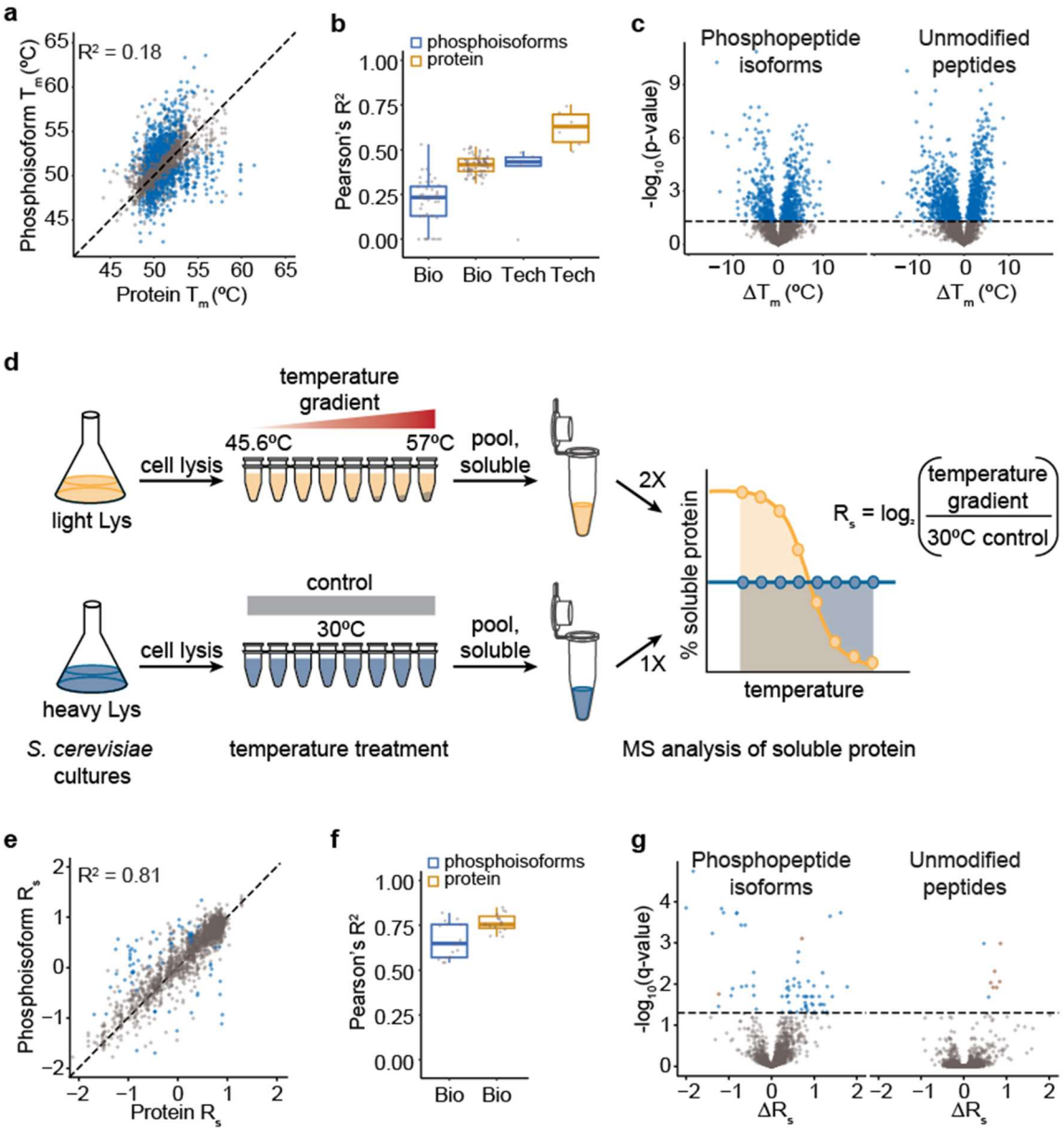


Figure 2.1. Most phosphosites have little effect on protein stability. a, Scatter plot and Pearson correlation between T_m of phosphopeptide isoforms (n=10) and T_m of the corresponding protein (n=12). Mean T_m values were obtained from Huang et al.. Significant phosphopeptide isoforms in blue. **b,** Boxplot depicting pairwise Pearson correlations between biological and technical replicates for HTP's T_m values. The line represents the median, the box designates the interquartile range (IQR), and the whiskers define 1.5*IQR from the box ends. **c,** Volcano plots showing differences in protein thermal stability (T_m) between phosphopeptide isoforms (n=10, left panel) or unmodified peptides observed in the phosphopeptide enriched sample (n=10, right panel) and their corresponding protein (n=11) (two-sided Student's t-test, significant hits at p-value <0.05 in blue), from Huang et al. data reanalysis. **d,** Dali workflow depicting SILAC labeling of yeast, the gradient temperature treatment of the protein extract, the inclusion of a 30ºC control, the quantification of soluble protein by mass spectrometry, and the calculation of relative stability (R_s). **e,** Scatter plot and Pearson correlation as in (a) using Dali's R_s data. **f,** Boxplot as in (b) with Pearson correlations between biological replicates for Dali's R_s values. **g,** Volcano plots as in (c) with x-axis showing R_s values. Here n=6, using a two-sided Welch's t-test, p-values were Benjamini-Hochberg corrected and significant hits at q-value < 0.05. Significant hits in blue, significant hits found in proteins with known cleavage or splicing events are in orange.

To minimize technical noise derived from sample preparation, peptide samples should be labeled and mixed prior to phosphopeptide enrichment (Appendix A Supplementary Discussion and accompanying manuscript[39]). Because scaling-up isobaric chemical labeling increases reagent costs substantially, we have developed an alternative approach to identify phosphosites that alter protein thermal stability, that we call Dali (Figure 2.1d). Dali applies the Proteome Integral Stability Alteration (PISA) method[40], a simplified version of thermal proteome profiling[19], in which the soluble protein from the different temperature points are combined to provide an estimation of the area under the protein melting curve. To reliably compare phosphoisoforms to proteins, we normalize each measurement to a 30ºC treated proteome reference that is labeled with heavy lysine, obtaining a relative stability (R_s) measurement for phosphoisoforms and proteins. This 30ºC reference is mixed in with the temperature gradient treated samples prior to protein digestion, and it is present during phosphopeptide enrichment and mass spectrometry (MS) measurement of peptides and phosphopeptides.

We applied Dali to the *S. cerevisiae* proteome and found that the stability of phosphoisoforms correlates well with the stability of their respective proteins ($R^2 = 0.82$ for mean R_s comparisons) (Figure 2.1e), suggesting that most phosphosites do not alter protein stability, as also observed in human cells[39]. We obtained reproducible R_s measurements for proteins (average $R^2 = 0.77$) and phosphoisoforms (average $R^2 = 0.66$) (Figure 2.1f). As expected, the stability of unmodified peptides present in the phosphopeptide-enriched samples also correlated well with their proteins ($R^2 = 0.92$ for mean R_s comparisons), indicating that R_s measurements in the phosphopeptide samples and protein samples can be reliably compared (Extended Data Figure 2.1b). Finally, our analysis yielded 68 phosphopeptide isoforms out of 1,978 (3%) with significantly different thermal stability than the unmodified protein (Figure

2.1g, Dataset S2). We detected a few unmodified peptides with significant differences in stability, but this set constituted a much smaller fraction than the fraction of significant phosphoisoforms (Dataset S3). Most of these peptides (6 out of 8) were cases where the protein is known to be post-translationally processed via cleavage (e.g. RPS31[41]) or splicing (e.g. VMA1[42]), resulting in proteins and/or proteoforms of different thermal stability as our method reliably measured (Extended Data Figure 2.3).

Among phosphosites that decreased protein thermal stability, we identified four sites located at protein interfaces (Ser56 on PUP2, Ser59 on ARO8, Ser79 on TPI1 and Ser201 on GAPDH) (Figure 2.2a, Extended Data Figure 2.4) that may act by disrupting protein-protein interactions. For example, PUP2 is the alpha 5 subunit of the 20S proteasome, and Ser56 is a known Cdc28 substrate[43] located at the protein interaction interface with PRE6, the 20S proteasome alpha 4 subunit (Figure 2.2a). The stability measured for the phosphopeptide spanning Ser56 is significantly lower than the stability of PUP2, which is similar to other proteins in the 20S proteasome, suggesting Ser56 phosphorylation may dissociate PUP2 from the 20S proteasome. We identified stabilizing phosphosites that may play a role in the protein translation process. For example, we found that phosphorylation at Ser38 on ribosomal protein RPL12/uL11 significantly increased protein stability (Figure 2.2b). This phosphosite is an evolutionarily-conserved Cdc28 substrate[43] that is regulated during the cell cycle[44] and has been reported to be depleted in polysomes and influence mitotic translation[45]. Considering RPL12 location at the ribosome P-stalk and the proximity of residue Ser38 to elongation factor 2, we hypothesize that this phosphosite may modulate the interaction with EF-2 to aid ribosomal translocation during protein synthesis, and the change in conformation and binding may stabilize RPL12. We also identified a stabilizing phosphorylation on NEW1 at Thr1191 (delta R_s = 1.23)

(Figure 2.2c). A T1191A mutant has growth defects[46] suggesting that phosphorylation is important for NEW1 function (Appendix A Supplementary Discussion). Additionally, we identified phosphosites that may modulate the kinetics of key glycolytic enzymes (Thr331 on PGK1 destabilizing, Ser149 on GAPDH stabilizing) (Extended Data Figure 2.5, Appendix A Supplementary Discussion).

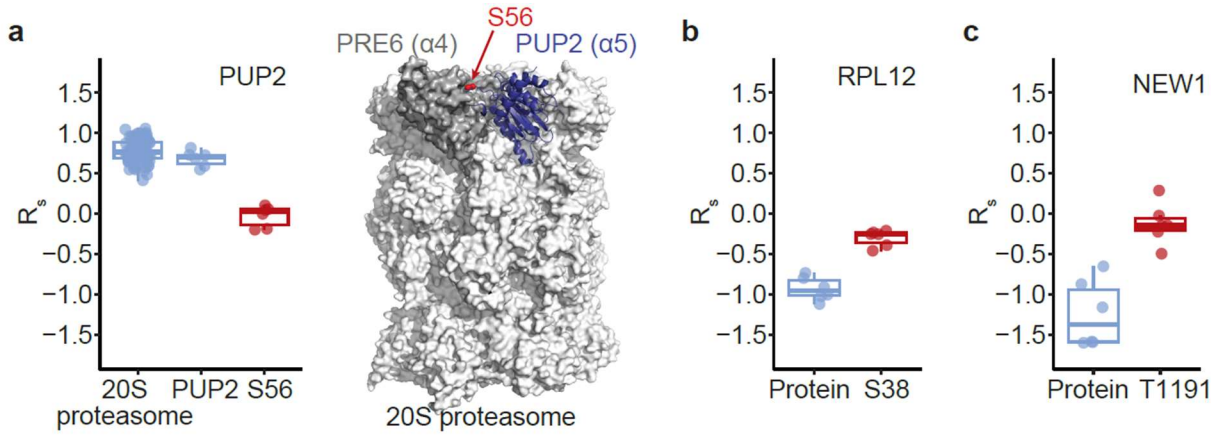


Figure 2.2. Examples of phosphosites that alter protein thermal stability. a, Boxplot of R_s values and distributions for the phosphopeptide containing PUP2 Ser56, PUP2 protein, and all the proteins in the 20S proteasome. Shown at the right is the structure of the 20S proteasome with PUP2 in blue, Ser56 in red, and PRE6 in grey. **b,** R_s values for RPL12 and RPL12 S38 phosphopeptide isoform. **c,** R_s values for NEW1 and phosphopeptide isoform containing NEW1 T1191. Boxplots show results from 6 biological replicates, the line represents the median, the box designates the interquartile range (IQR), and the whiskers define 1.5*IQR from the box ends.

In this communication, we have detailed some technical issues in the HTP method and suggested changes to the experimental workflow and statistics to improve its reproducibility and reliability. We have also outlined a novel proteomic method that enables robust thermal stability comparison between proteins and phosphorylated proteoforms. Our method identified 3% phosphosites in the *S. cerevisiae* proteome that significantly changed protein melting behavior. Additional experiments will be needed to precisely characterize the function of these phosphosites. One potential limitation of our method is that the sensitivity to detect changes in thermal stability is lower for proteins with extreme melting temperature, which can be circumvented by performing the experiment using different temperature gradients. Our method

can be extended to other model organisms and cell culture systems, as well as to other post-translational modifications, expanding the proteomic toolkit to functionally annotate dynamic protein modifications at scale.

2.3 METHODS

Yeast strains

All yeast experiments were conducted on the *Saccharomyces cerevisiae* haploid strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0).

S. cerevisiae growth, stable isotope labeling, and cell harvest

Two overnight yeast cultures were grown at 30ºC in synthetic complete media (SCM) containing 6.7g/L yeast nitrogen base, 2g/L of synthetic complete mix minus lysine, 2% glucose, and supplemented with either regular lysine (light culture) or $^{2}H_{4}$-lysine (heavy culture) at 0.872 mM final concentration. These cultures were used to seed three 50mL cultures of each light and heavy at OD_{600} 0.15, which were grown at 30ºC and 45mL were harvested at OD_{600} ~ 1 by centrifugation at 7,000 x g for 10min. Yeast pellets were washed by resuspension in 1.5mL ice-cold sterile water and centrifugation in 2mL screw cap tubes at 21,000 x g for 10min, and then snap-frozen in liquid nitrogen and stored at -80ºC.

Cell lysis and protein extract temperature treatment

Frozen yeast cell pellets were resuspended in 700µL of non-denaturing lysis buffer (50mM HEPES pH 7.0, 75mM NaCl) containing 0.5X protease inhibitors (Pierce) and phosphatase inhibitors (50mM β-glycerophosphate, 10mM sodium pyrophosphate, 50mM of NaF, 1mM sodium orthovanadate) on ice. Cells were lysed by bead beating with 0.5mm

zirconia/silica beads for 4 cycles of 60sec of mechanical agitation followed by 90sec rest on ice. Lysates were clarified by sequential centrifugation, first at 1,200 x g for 1min to remove the beads and then at 21,000 x g for 10min at 4ºC to remove cell debris. To bring all protein extracts to the same concentration, extract volumes were adjusted to 1 OD_{600} unit from a 45mL culture in 1mL.

Each cell extract was aliquoted into 2 strips of 8 PCR tubes each (1x8 for the temperature gradient and 1x8 for the 30ºC) dispensing 50µL of protein extract per tube. All samples were initially equilibrated to 30ºC for 5 min. Temperature gradient samples were subjected to 45.6ºC, 46.8ºC, 48.3ºC, 50ºC, 52ºC, 53.6ºC, 54.9ºC, and 57ºC, one tube to each temperature, for 5min. In parallel, controls were subjected to an additional 30ºC temperature treatment for 5min. All samples were cooled down to room temperature for 10min. For each replicate, temperature gradient samples were all pooled into one tube and 30ºC controls were pooled into a separate tube prior to centrifugation at 21,000 x g for 30min at 4ºC. The soluble protein fractions for the temperature gradient and 30ºC controls were combined 2:1, three replicates with the temperature gradient labeled heavy and the 30ºC controls labeled light, and three additional replicates with the labels swapped. We generated additional controls where heavy and light 30ºC controls were combined to assess potential differences in protein expression due to the different labeling. Protein concentration was measured by the BCA assay.

Protein reduction, alkylation, LysC digestion, and desalting

Samples were diluted 2-fold with a buffer containing 8M urea, 50mM HEPES pH 8.9, 75mM NaCl, 1mM sodium orthovanadate, 50mM β-glycerophosphate, 10mM sodium pyrophosphate, 50mM NaF. Protein samples were subjected to reduction with 7.5mM

dithiothreitol (DTT) for 30min at 55ºC and alkylation with iodoacetamide (22.5mM) for 30min at room temperature in the dark with agitation. The alkylation reaction was quenched with an additional 7.5mM DTT at room temperature for 30min with agitation. The pH was adjusted to 8.5 with 1M Tris pH 8.9. Lysyl endopeptidase (LysC; Wako Chemicals) was added at a 1:100 enzyme to protein ratio and protein samples were incubated overnight with agitation at room temperature. LysC digestion was quenched by addition of trifluoroacetic acid (TFA) to a final concentration of 1% and pH ~2-3 and the digests were stored at -80ºC.

Peptide samples were desalted by solid-phase extraction over 50mg Sep-Pak tC_{18} cartridges (Waters). Packing material was washed with 1mL methanol, 3 x 1mL 100% acetonitrile, 1mL 70% acetonitrile, 0.25% acetic acid, 1mL 40% acetonitrile, 0.5% acetic acid, and equilibrated with 3 x 1mL 0.1% TFA. Peptides were then loaded by gravity twice, washed with 3 x 1mL 0.1% TFA and 1mL 0.5% acetic acid. Peptides were eluted with 600µL of 40% acetonitrile, 0.5% acetic acid and 400µL 70% acetonitrile, 0.25% acetic acid, and aliquoted as follows: 40µg for high-pH reversed-phase fractionation, 200µg for Fe^{3+}-IMAC phosphopeptide enrichment, and 10µg for preliminary LC-MS/MS analysis to assess sample quality. All samples were dried by vacuum centrifugation and stored at -80ºC.

High-pH reversed-phase fractionation

Peptides were fractionated by high-pH reversed-phase fractionation on a 200µL pipette tip packed with 4 layers of SDB-XC material (Empore, 3M). The material was washed with 50µL methanol, 50µL 80% acetonitrile, 20mM ammonium formate, and 3 X 50µL 20mM ammonium formate. Peptides (40µg) were solubilized in 40µL of 5% acetonitrile, 20 mM ammonium formate, loaded onto the SDB-XC tip, and the flow-through was collected in a mass

spectrometer vial (fraction 1). Peptide fractions 2-5 were obtained by stepwise elution with 40µL of 20mM ammonium formate in 10 %, 15%, 20%, and 80% acetonitrile and collection in mass spectrometry vials. Peptide fractions were dried by vacuum centrifugation, solubilized in 3% acetonitrile, 4% formic acid, and ~1µg of each fraction was analyzed by LC-MS/MS.

Fe^{3+}-NTA IMAC phosphopeptide enrichment

Phosphopeptide enrichment was conducted by immobilized iron cation affinity chromatography in batch mode and automated in a 96-well format on a KingFisher magnetic particle processor as we described[47]. For each sample, ~200µg peptides were solubilized in 70 µL 0.1% TFA, 80 % acetonitrile and incubated with 80 µL of a 5% slurry of magnetic Fe-NTA beads (Cube Biotech) in the same solvent for 30min. Beads were washed three times with 150µL 0.1% TFA, 80% acetonitrile and phosphopeptides were eluted with 50µL 50% acetonitrile, 0.37M ammonium hydroxide. Eluates were acidified with 30µL 10% formic acid, 75% acetonitrile and filtered over two-layer C18 extraction disks (Empore, 3M) packed in 200µL pipette tip, which had been previously conditioned with 50µL100% methanol, 50µL 100% acetonitrile and 50µL 70% acetonitrile, 0.25% acetic acid. Filtered peptides were collected in a mass spectrometer vial and the peptides in the extraction disk were further eluted with 50µL 70% acetonitrile, 0.25% acetic acid and collected into the same mass spectrometry vial. Phosphopeptide-enriched samples were dried by vacuum centrifugation, solubilized in 3% acetonitrile, 4% formic acid, and one third of each sample was analyzed by LC-MS/MS.

Liquid chromatography coupled to tandem mass spectrometry

Peptide samples were analyzed by nLC-MS/MS on a nanoAcquity UPLC (Waters) coupled to an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher) with Xcalibur

(4.1.50) and Thermo Foundation 3.1 SP4 (3.1.190.0). Samples were loaded on a 100µm x 3-cm trap column packed with 3µm C18 beads (Dr. Maisch), separated on a 100µm x 30-cm capillary analytical column, packed with 1.9µm C18 beads (Dr. Maisch) and set at 50ºC, using a 90-min reversed-phase gradient of acetonitrile in 0.125% formic acid, and online analyzed by mass spectrometry using data-dependent acquisition. Each cycle consisted of 3 sec where one full MS1 scan was acquired on the orbitrap at 120,000 resolution from 300 to 1575 m/z using an AGC of 7e5 and maximum injection time of 50ms followed by MS/MS dependent scans on most intense precursor m/z ions (only considering z = 2 to 5) until exhausting the 3sec cycle time, using 1.6 m/z isolation window, HCD fragmentation at 28 normalized collision energy, and acquired at 15,000 resolution on the orbitrap with an AGC of 5e4 (peptide samples) and AGC of 1e5 (phosphopeptide samples) with a maximum injection time of 22ms. Dynamic exclusion was enabled to exclude fragmented precursors from repeated MS/MS selection for 30sec. To increase coverage, phosphopeptide samples were injected twice, and the data from the two technical replicates were combined.

Database searching, peptide quantification, phosphosite localization, and R_s calculation

MS data files for proteome samples were analyzed with MaxQuant[48] (v.1.6.7.0) to obtain peptide identifications and quantifications, using the following parameters: protein sequence database *S.cerevisiae* downloaded from SGD in July 2014, LysC enzyme specificity (cleavage C-terminal to K), maximum of 2 missed cleavages, mass tolerance of 20ppm for MS1 and 20ppm for MS2, fixed modification of carbamidomethyl on cysteines, variable modifications of oxidation on methionines and acetylation on protein N-termini. Lysine residues were only allowed to be all light or all $^{2}H_4$-Lys within the same peptide. Phosphoproteome samples were processed in MaxQuant similarly as above, with additional variable modification of

phosphorylation on serine, threonine, and tyrosine residues. All searches were combined for MaxQuant filtering set to 1% FDR at the level of peptide spectral matches and protein.

Quantification values for heavy and light peptide features were extracted from the evidence.txt file. Quantification values for features corresponding to the same peptide sequence (e.g. same peptide identified at multiple charge states or fractions) were summed up. Phosphopeptide quantification features were aggregated to the phosphopeptide isoform level by summing features corresponding to the same peptide sequence (e.g. same peptide identified at multiple charge states or replicate injections) as well as overlapping peptide sequences sharing the same combination of modifications. For each phosphopeptide isoform we required the maximum localization probability to be greater than 75% for at least one site.

R_s values were calculated as $\log_2$ ratios of the quantification value for the temperature gradient treated divided by the respective quantification for the 30°C control. Peptide R_s distributions were median normalized to 0, and the same correction value derived for each replicate was applied to normalize the corresponding phosphopeptide isoform R_s distributions. To filter out poor quality quantifications, peptides and phosphopeptide isoforms with the 5% highest R_s standard deviation across replicates were excluded from the analysis. Protein R_s values were calculated as the median of peptide R_s for that protein, requiring a minimum of two peptides per protein, and each peptide observed in at least two replicates.

To identify phosphopeptide isoforms that have different R_s than their unmodified protein counterpart, we performed a t-test comparing phosphopeptide isoform R_s values (n=6) compared to protein R_s values (n=6) and assuming unequal variance. Phosphopeptide isoforms and protein counterparts were required to be observed in at least three replicates. P-values were corrected for

multiple hypothesis testing using the Benjamini-Hochberg method[49]. All data analysis was conducted using R (v3.6.1) and RStudio (v1.2.1335), and data figures were generated in R and Adobe Illustrator CS5 (v15.0.0).

Structure visualization and bioinformatics

Protein complex annotations were extracted from the CYC2008 resource[50]. Protein structure coordinates were downloaded from PDB and visualized and manipulated with PyMOL[51]. For PUP2 interface analysis, we extracted 20S proteasome protein structure from PDB 1RYP[52]. Protein interface structures for ARO8, TPI1, and GAPDH were extracted from PDB (4JE5[53], 1NEY[54], 3PYM[55] respectively). To assess the stabilizing effect of S149 phosphorylation at the catalytic site of GAPDH, we aligned crystal structures of GAPDH with bound G3P (1NQO[56]) and inorganic phosphates (1GYP[57]) to a NAD-bound yeast GAPDH structure (3PYM[55]). Data on sequence conservation, protein interfaces, and predicted stability effects of mutations ($\Delta\Delta G_{pred}$) were obtained from the mutfunc resource[58].

Reanalysis of the Huang et al. data

Supplementary data from Huang et al.[38] was used to calculate the correlation between T_m for phosphopeptides and proteins and learn about their statistical parameters. For data reanalysis, all MS files from the study were downloaded from MassIVE data repository (dataset identifier: MSV000083786), converted to open format mzXML files with ReAdW (2015.1.0), and database searched with Comet[59] (v.2015.02 rev2) to obtain peptide and phosphopeptide identifications, with the exception of "Bulk_6_2" which failed to convert. Database search parameters were: human protein sequence database from UniProt (UP000005640), mass tolerance of 50 ppm for precursor m/z and 0.2 Da for fragment ions, trypsin enzyme specificity (cleavage Ct to K, R,

except for KP, RP), maximum of 2 missed cleavages, fixed modification of carbamidomethyl on cysteines and TMT10 (+229.1629) on lysines and peptide N-termini, and variable modifications of oxidation on methionines and acetylation on protein N-termini. Phosphorylation samples included variable modification of phosphorylation at serine, threonine, and tyrosine.

Search results were filtered to a PSM 1% FDR with Percolator[60]. Phosphosite localization was conducted with an in-house C++ implementation of Ascore[61] and sites with Ascore > 13 were considered confidently localized ($P < 0.05$). TMT10 reporter ion intensities were extracted from MS/MS scans using in-house TMT quantification software (IsobaricQuant 1.0).

We attempted to replicate the analysis by Huang et al. by following the method description provided in their manuscript. Biological and technical replicates were treated equally. For each replicate, TMT reporter ion intensities for all peptide spectral matches from the proteome files were summed to the protein level, and TMT reporter ion intensities for phosphopeptide spectral matches were summed to the phosphopeptide isoform level. In addition, we used the same strategy to aggregate to peptide-level the TMT signals for PSMs mapping to the same unmodified peptide observed in the phosphopeptide-enriched samples. We implemented the TPP package in R to fit melting curves for proteins, phosphopeptide isoforms, and unmodified peptides in the phosphorylation-enriched sample. To recapitulate the reported results, we had to conduct the fitting for all samples together (Appendix A Supplementary Discussion). Melting curves were filtered for fitting $R^2 > 0.8$. T-tests were conducted by comparing T_m values for phosphopeptide isoforms or unmodified peptides observed in the phosphopeptide-enriched samples to the unmodified protein T_m values, assuming equal variances and without multiple hypothesis correction (as implemented by Huang et al.). Of note, our

reanalysis revealed that one of the phosphoproteome technical injections for biological replicate 5 was instead a repeated MS analysis of biological replicate 4.

Data availability

The mass spectrometry proteomics data generated for this manuscript have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD016750.

Code availability

All code to reproduce analysis and data figures is available at https://gitlab.com/villenlab/dali_phospho_thermalstability.

Acknowledgements

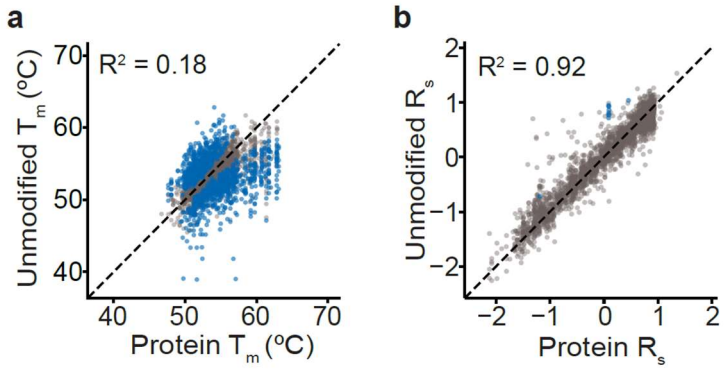
We thank members of the Villén lab for scientific discussions, in particular Bianca Ruiz, Mario Leutert, and Alex Hogrebe. We thank Ariadna Llovet Soto and Jimmy Eng for software developments on the data analysis pipeline. I.R.S. and K.N.H were supported by NIH training grant T32HG000035. A.S.V. was supported by NIH training grant T32LM012419. Most of this work was supported by NIH grant R35GM119536 to J.V. The Villén lab is additionally supported by NIH grants R01AG056359, R01NS098329, and RM1 HG010461, Human Frontiers Science Program grant RGP0034/2018, a Research Program grant from the W.M. Keck Foundation, and the University of Washington Proteome Resource UWPR95794.

Author Contributions

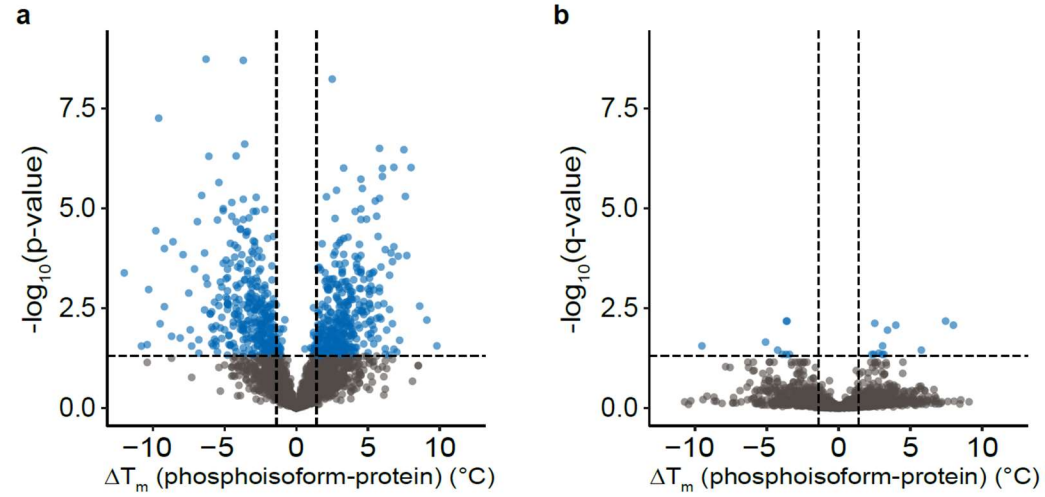
I.R.S., K.N.H., R.A.R.-M, and J.V. conceived the study and designed the experiments. I.R.S. conducted the experiments with advice from K.N.H., R.A.R.-M. and J.V., and assistance

from A.A.B. I.R.S. analyzed the data with advice from R.A.R.-M. and A.S.V. J.V. supervised the study. I.R.S. and J.V. wrote the paper and all authors edited it.

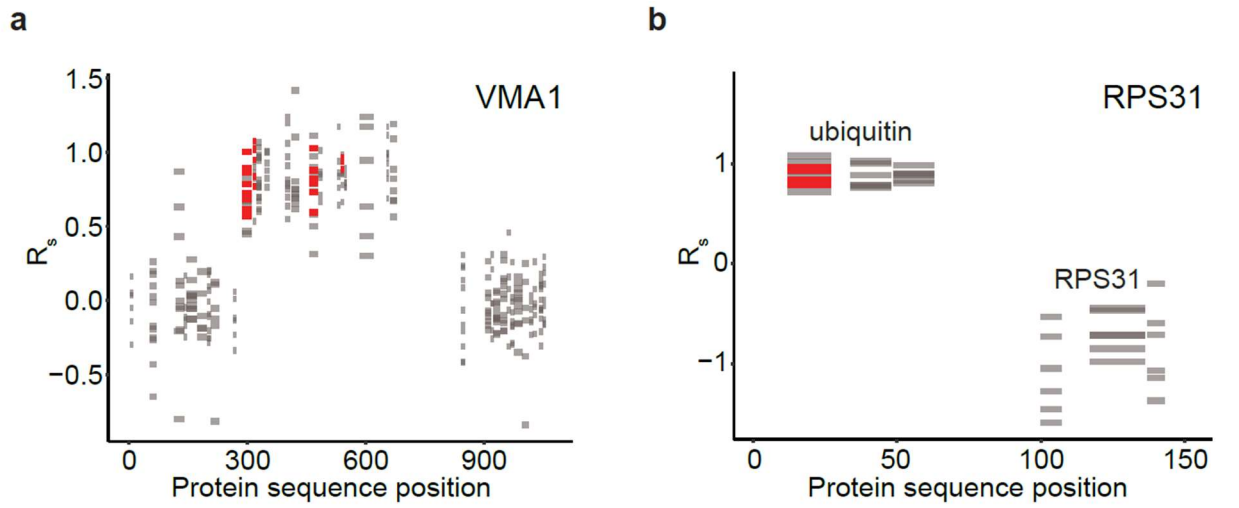
2.4 Extended Data Figures



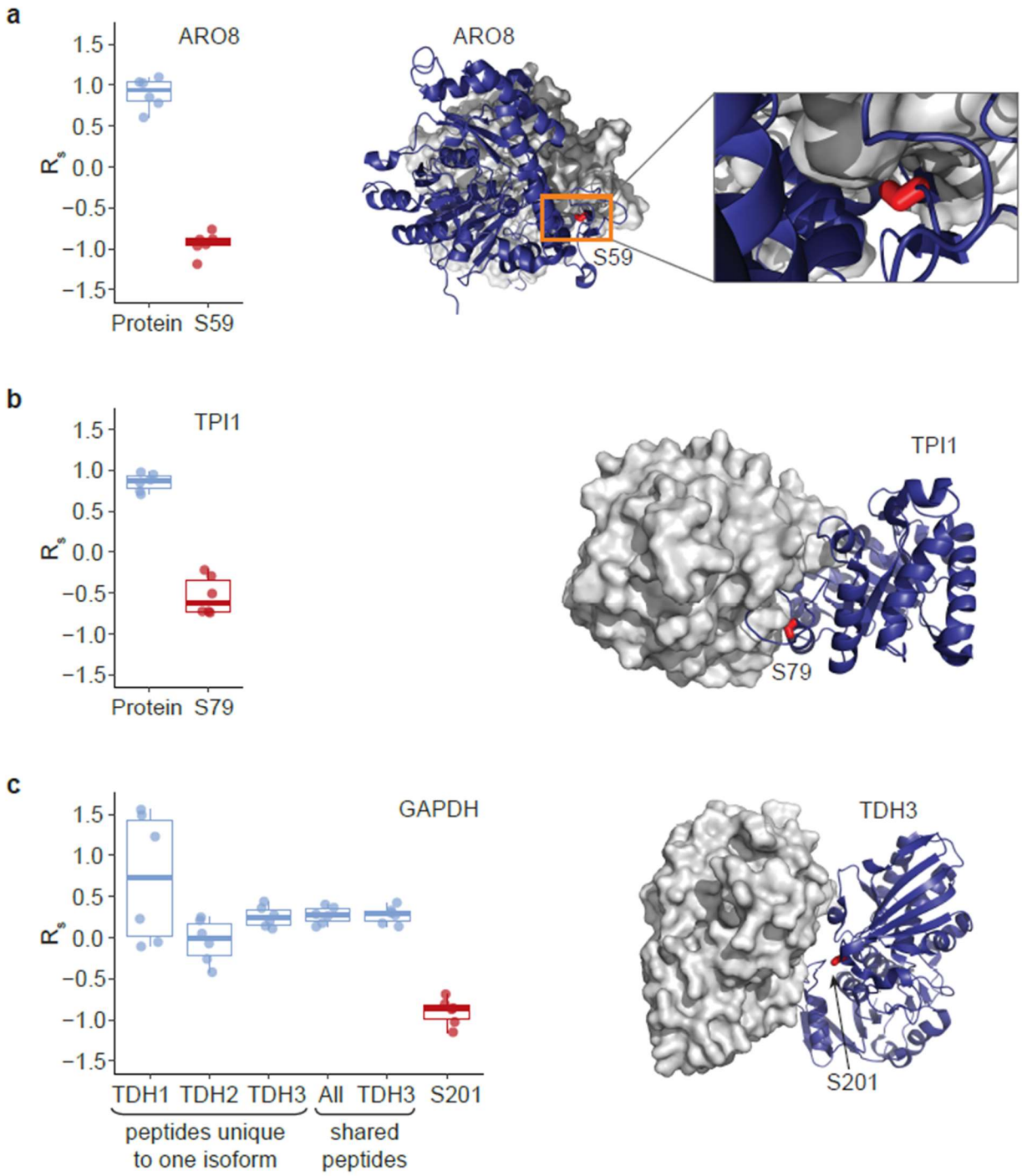
Extended Data Figure 2.1. Reproducibility and robustness of Dali compared to HTP. a, Scatter plot and Pearson correlation between the mean T_m for unmodified peptides observed in the phosphopeptide enriched samples (n=10) and the mean T_m for their corresponding proteins (n=11). Results from the Huang et al. data reanalysis conducted by us. **b,** Scatter plot and Pearson correlation as in (a) with R_s values obtained from the Dali method (n=6).



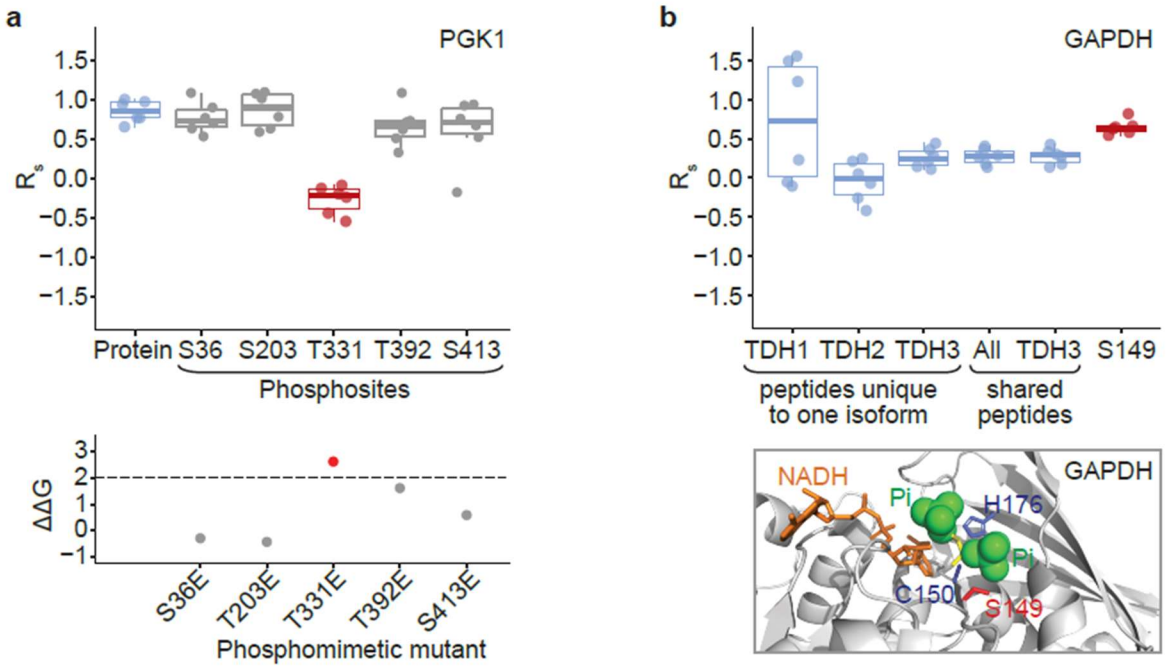
Extended Data Figure 2.2. Phosphosites that significantly alter protein thermal stability using two different statistical settings. Volcano plots showing ΔT_m for mean phosphopeptide isoform to mean protein counterpart in the x-axis, and the two-sided Student's t-test probability in the y-axis. **a,** Huang et al. implementation shows a p-value because multiple hypothesis correction was not applied. Significant phosphopeptide isoforms (blue) are defined by p-value < 0.05. **b,** Our proposed analysis consolidates data from MS reanalysis prior to statistical testing using a two-sided Welch's t-test, which is performed assuming unequal variances between phosphopeptide isoform and proteins. Benjamini-Hochberg adjustment was used to correct p-values for multiple hypothesis testing. Significant phosphopeptide isoforms (blue) are defined by q-value < 0.05.



Extended Data Figure 2.3. Examples of significant hits on proteins that undergo post-translational splicing or cleavage. a, R_s values for observed VMA1 unmodified peptides identified in phosphopeptide-enriched samples and proteome samples displayed across the length of VMA1. Spliced products from amino acid 2-283 and 738-1031 are joined to generate the V-type proton ATPase catalytic subunit A proteoform, extinguishing the 284-737 segment. Peptides derived from the proteome samples are colored in gray and significant unmodified peptides found in the phosphopeptide-enriched sample are in red. **b,** Similar plot to (a) for RPS31, which is cleaved to generate ubiquitin (1-76 amino acid segment) and 40S ribosomal protein S31 (77-152 amino acid segment) proteins.



Extended Data Figure 2.4. Examples of phosphosites that alter protein thermal stability and are located at protein interfaces. R_s boxplots for **a,** ARO8 S59, **b,** TPI1 S79, and **c,** GAPDH S201 phosphopeptide isoforms and their protein counterparts. All boxplots show results from n=6 biological replicates, and the line represents the median, the box designates the interquartile range (IQR), and the whiskers define 1.5*IQR from the box ends. ARO8 S59, TPI1 S79, and GAPDH S201 reside at dimerization interfaces as shown in the structures to the right (PDB accession: 4JE5, 1NEY, and 3PYM, respectively). Phosphomimetic mutations ARO8 S59E and TPI1 S79E are predicted to disrupt protein interfaces ($\Delta\Delta G_{pred}$ = 3.78 and $\Delta\Delta G_{pred}$ =8.04 respectively). Additionally, TPI1 S79E mutation is predicted to alter protein conformational stability ($\Delta\Delta G_{pred}$ = 2.39). $\Delta\Delta G_{pred} > 2$ is predicted to be destabilizing.



Extended Data Figure 2.5. Examples of phosphosites that alter protein thermal stability on glycolytic enzymes. a, R_s values for PGK1 and all measured PGK1 phosphopeptide isoforms, with significantly destabilizing phosphosite S331 shown in red. $\Delta\Delta G_{pred}$ for all glutamic acid phosphomimetic substitutions were obtained from mutfunc, with $\Delta\Delta G_{pred} > 2$ considered likely destabilizing. **b,** GAPDH S149 phosphopeptide is shared across all GAPDH paralogs (TDH1, TDH2, and TDH3). Boxplot shows R_s values and distributions for peptides unique to one isoform (TDH1, TDH2, TDH3), peptides shared among all GAPDH isoforms (all), all peptides for TDH3, and the S149 phosphopeptide isoform. Bottom panel shows localization of S149 on the GAPDH structure near the binding site of the enzyme substrate. Boxplots show results from 6 biological replicates, the line represents the median, the box designates the interquartile range (IQR), and the whiskers define 1.5*IQR from the box ends.

Chapter 3. KINETIC ANALYSIS OF PHOSPHORYLATION IMPACT ON EXPERIMENTALLY-DERIVED PROTEIN TURNOVER AND PROTEIN AGE-BIASED PHOSPHORYLATION

3.1 ABSTRACT

Despite being able to catalog 100,000s of phosphorylation events using mass spectrometry, we lack methods to experimentally prioritize which of these sites are functional at scale. To address this bottleneck, we identified phosphosites that have a measurable effect on protein turnover, a protein property that associates with function. Thus, we coupled dynamic SILAC labeling with phosphoproteomics to generate peptide-level readouts for relative protein turnover (R_{TO}), a nuanced protein turnover readout for application to modified proteoforms. Using phosphoprotein-specific phosphopeptides, we identify phosphorylation peptidoforms with altered R_{TO} compared to their protein, with most phosphosites demonstrating a decreased R_{TO}. Using data-supported kinetic simulations, slower R_{TO} sites were deemed difficult to interpret due to technical factors related to heavy-lysine amino acid incorporation for modified proteoforms. Interestingly, a traditional turnover model with the addition of phosphorylation cannot account for the observation of faster R_{TO} phosphorylation sites, which are observed at ~9%. These faster R_{TO} phosphorylation sites are enriched in [S/T]XX[E/D] CK2 motifs and presence in beta sheets, while depleted in [S/T]P CMGC motifs. Herein, we offer a protein age-biased phosphorylation kinetic model that serves as a possible explanation for the faster R_{TO} phosphosites that aligns with many known functional sites. We assert that a faster R_{TO} phosphosite designation can serve as a useful prioritization criteria for follow-up functional validation. Also, our molecular age-

biased phosphorylation explanation for these sites represents a novel mechanism for post-translational protein regulation that warrants further exploration.

3.2 INTRODUCTION

Protein turnover is an essential mechanism for maintaining proteostasis and proper cellular protein abundance by a balance between protein synthesis and degradation. The kinetics of this molecular balance are driven by transcriptional and translational signals for protein expression and subsequent degradation mechanisms, like the ubiquitin-proteasome system[62].

Through extensive characterization of determinants of protein turnover in *Saccharomyces cerevisiae*, Martin-Perez and Villén[29] identified that protein turnover is regulated by transcriptional, translation, and post-translation factors. Protein turnover is also influenced by subcellular context and tends to mimic the turnover of its other complex members. Additionally, protein turnover has been observed to increase in response to increased protein activity and to change in response to altered cellular state. Interestingly, sequence motifs such as degron sequences and proteins with phosphorylation-ubiquitin crosstalk tend to have faster turnover, suggesting an association between post-translational modifications and their impact on unmodified protein turnover. Given that phosphorylation has been observed to function on proteins via mediating protein-protein interactions, regulating protein activity, altering protein subcellular localization, and promoting degradation by acting as a phosphodegron[28,63], we hypothesize that many phosphosites could modulate protein turnover and apparent differences in protein turnover between a phosphorylated protein and its unmodified counterpart protein could provide a functional prioritization for phosphosites.

To date, dynamic SILAC labeling coupled with mass spectrometry has become a standard protocol for protein half-life determination at a proteome-wide scale. Using this approach, we monitor the steady state kinetics of protein synthesis and degradation following a pulse of an isotopically labeled amino acid, which can be used to calculate protein turnover[27]. Most proteomic turnover studies using this approach are conducted at the protein-level, aggregating all molecular forms of a single protein coding gene product (proteoforms)[4] to a protein representation. However, this aggregation strategy largely ignores the molecular diversity for a given protein coding gene which can scale to hundreds of proteoforms[34], thus leaving the role of post-translational regulation in the form of post-translational modifications (PTMs) and other proteoforms on protein turnover poorly understood.

Zecha et al.[31] highlighted the importance of measuring proteoform-specific protein turnover for many post-translationally modified N-termini and post-translationally cleaved proteoforms which demonstrate distinct protein turnover profiles. While protein turnover is a protein-level phenotype, proteoforms were distinguished from one another following proteome digestion enabling peptide-level protein turnover readouts that serve as molecular signatures for all protein molecules that contain that peptide. Distinct separation among peptide-level protein turnover can demarcate the one or more proteoforms. Specifically, for phosphorylation's role on protein turnover, Wu et al.[64] developed a method called DeltaSILAC, which uses a pulsed-SILAC strategy coupled with phosphoproteomics and identifies differences between unmodified protein and phosphorylation isoforms using differences in unmodified peptide's and phosphopeptide's protein turnover readouts, respectively. In the method's initial implementation in human cell lines, the authors explore phosphorylation turnover differences to their unmodified protein assuming a "phosphate transfer" independent model, which simplifies the system such

that the synthesis and degradation of the phosphorylation isoform and its unmodified protein are independent.

While phosphorylation can occur co-translationally, possibly mimicking a “phosphate transfer” independent model, often phosphorylation and dephosphorylation (via dynamic phosphate transfer by kinases and phosphatases, respectively) occurs following unmodified protein synthesis. Thus, we developed a “phosphate transfer” dependent model for phosphorylation to systematically interrogate the dependency of the phosphorylation isoform and its unmodified protein on amino acid incorporation kinetics to ascribe meaning to apparent protein turnover differences. We implemented a dynamic SILAC approach coupled to phosphoproteomics for proteome-wide comparisons of phosphorylation isoform and its protein turnover in *Saccharomyces cerevisiae*, which is a useful model system for assessing phosphate transfer kinetics due to yeast’s faster cellular doubling rates which competes with the phosphate transfer kinetics. Using experimental data informed kinetic simulations, we identify limitations in the interpretation of apparent phosphorylation protein turnover and uncover a possible protein age-biased phosphorylation model to explain faster turnover phosphosites. We identify known phosphorylation events that likely support our protein age-biased phosphorylation model, which likely has functional implications and represents a novel mechanism for post-translational protein regulation that warrants further exploration.

3.3 METHODS

Yeast strain

All yeast experiments were conducted using the Saccharomyces cerevisiae diploid strain DBY10144 (MAT a/α) provided by Maitreya Dunham (University of Washington). DBY10144 is prototrophic for lysine and is from the FY (S288C) background (parental strains FY4H and FY3G). This strain has been previously used for dynamic SILAC experiments [29,65].

Dynamic SILAC sample preparation

A 25 mL starter *Saccharomyces cerevisiae* culture was grown overnight at 30ºC in synthetic complete media composed of 6.7g/L yeast nitrogen base, 2% glucose, and 2g/L of drop-out mix that contained all amino acids except lysine. Yeast cells were then diluted in six replicates to OD_{600}=0.1 with the same media composition and grown at 30ºC . After 150 minutes cultures were diluted ~4% with the same media supplemented with heavy $^{13}C_6$,$^{15}N_2$-lysine (final concentration 0.436mM: 40 mL final volume), and incubated at 30ºC for 90 minutes (approximately one "molecular" doubling time for greatest sensitivity) before harvest. Yeast cells were harvested by addition of 100% trichloroacetic acid to 10% and centrifugation at 7,000 x *g* for 10 minutes at 4ºC. Supernatants were decanted and cell pellets were washed with 10 mL of -20ºC chilled acetone and spun at 7,000 x *g* for 10 minutes at 4ºC, snap frozen in liquid nitrogen and stored at -80ºC.

Cell lysis, protein reduction and alkylation, and protein digestion

Frozen DBY10144 pellets were resuspended in 500 uL of lysis buffer (50 mM Tris pH 8.2, phosphatase inhibitors: (75mM NaCl, 1mM sodium orthovanadate, 50 mM β-glycerophosphate, 10 mM sodium pyrophosphate, 50 mM of NaF), and 8M urea) on ice. Cells were lysed by bead beating with 0.5mm zirconia/silica beads for 4 cycles of 60 seconds of

mechanical agitation followed by 90 seconds rest on ice. Lysates were clarified by sequential centrifugation, first at 1,200 x g for 1 minute to remove the beads and then at 21,000 x g for 10 minutes at 4°C to remove cell debris. 10uL of protein supernatant was collected for a BCA assay (Pierce) to determine protein concentration. The rest of the clarified lysate was subjected to reduction with 5mM dethiothreitol (DTT) for 30 minutes at 55°C, alkylation with iodoacetamide for 30 minutes at room temperature in the dark with agitation, and quenched with an additional 5mM DTT at room temperature for 30 minutes with agitation. Reduced and alkylated samples were then diluted 1:1 (v:v) with non-urea containing pH 8.9 lysis buffer (50 mM Tris pH 8.9, 75mM NaCl, 1mM sodium orthovanadate, 50 mM β-glycerophosphate, 10 mM sodium pyrophosphate, 50 mM of NaF). Lysyl endopeptidase (LysC; Wako Chemicals in HPLC grade water) was added at a 1:100 enzyme to protein ratio and incubated overnight with agitation at room temperature. After ~14 hours, LysC digestion was quenched with trifluoroacetic acid (final concentration 1%). Acidified digest was placed at -80°C until desalting.

Peptide desalting

Roughly 1.3 mg of digested yeast lysate was cleaned up with 50mg Sep-Pak tC_{18} polymer columns (Waters). Column was activated with 1mL of methanol, and equilibrated with 3 x 1mL of 100% acetonitrile, 1mL of 70 % acetonitrile with 0.25% acetic acid, 1 mL of 40% acetonitrile with 0.5% acetic acid, and 3 x 1mL 0.1% trifluoroacetic acid (TFA). Peptides were then loaded by gravity twice and subsequently washed 3 x 1mL of 0.1% TFA and 1mL 0.5% acetic acid. Peptides were eluted with 600 uL of 40% acetonitrile in 0.5% acetic acid and 400 uL of 70% acetonitrile in 0.25% acetic acid. ~500ug was aliquoted for reversed phase basic fractionation, ~400ug was aliquoted for IMAC enrichment (2 enrichments at 200ug each), and 10 ug was

aliquoted for mass spectrometry for peptide sample quality control. All samples were dried by vacuum centrifugation and stored at -80ºC.

Reversed phase basic fractionation

Reversed phase fractionation was conducted using an Agilent 1100 offline HPLC with an Xterra 3uM C18 resin (Waters) with 20mM Ammonium formate (Solvent A) and 80% Acetonitrile in 20mM Ammonium formate as the mobile phase (Solvent B). Dried peptides were reconstituted with 3% Acetonitrile in 4% Formic Acid. ~250ug of digested peptides were loaded on column. Peptides were eluted with 3% B for two minutes and a linear gradient from 3-30% buffer B at a 700 uL/min flow rate. 12 reverse phase basic fractions were collected every two minutes on 200 uL of 70% Acetonitrile in 0.25 % Acetic Acid. All six replicates were dried by vacuum centrifugation. Dried reverse phased basic fractions were reconstituted to a final concentration of ~0.5 ug/uL in 3% Acetonitrile in 4% formic acid (shaken for 30 min at room temperature with agitation and transferred to mass spectrometry vials).

Fe^{3+}-NTA IMAC phosphopeptide enrichment

Phosphopeptide enrichment was conducted by immobilized iron cation affinity chromatography in a 96-well format on a KingFisher magnetic particle processor as we described in the Leutert et al. study [47]. For each sample, 400 µg of peptides were resolubilized in 140 µL 0.1% TFA, 80 % acetonitrile and split in two wells. 80 µL of a 5% slurry of magnetic Fe-NTA beads (Cube Biotech) in 0.1% TFA, 80 % acetonitrile was then added and incubated for 30 min. Beads were washed in the same solvent (0.1% TFA, 80 % acetonitrile) three times and then phosphopeptides were eluted off the beads with 50µL 0.37M ammonium hydroxide, 50% acetonitrile. Phosphopeptide eluates were acidified with 75% acetonitrile, 10% formic acid and filtered over two-layer C_{18} extraction disks (Empore) packed in 200µL pipette tip. Prior to

filtering, the two-layer C_{18} extraction disks had been conditioned with 50µL 100% methanol, 50µL 100% acetonitrile and 50µL 0.25% acetic acid, 70% acetonitrile. Filtered peptides were passed through the disk and collected in a mass spectrometer vial. To ensure all peptides were eluted, 50µL 0.25% acetic acid, 70% acetonitrile was added, eluted, and collected in the same mass spectrometry vial. Following vacuum centrifugation, dried phosphopeptide-enriched samples were solubilized in 4% formic acid, 3% acetonitrile, and one third of each 200µg phosphopeptide enrichment replicate sample was analyzed by LC-MS/MS.

Liquid Chromatography Mass Spectrometry of High pH reversed phase fractions and phospho-enriched samples

Peptide samples were analyzed by nLC-MS/MS using a Easy nanoLC 1200 (Thermo Fisher, Odense, Denmark) online with a Q Exactive Plus hybrid quadrupole-orbitrap mass spectrometer (Thermo Fisher, Bremen, Germany). Peptides were loaded on a 100µm x 3-cm trap column packed with 3µm C18 beads (Dr. Maisch). Using a reversed-phase gradient of acetonitrile (80% stock) in 0.125% formic acid, peptides were separated on a 100µm x 30-cm capillary analytical column packed with 1.9µm C18 beads (Dr. Maisch) heated at 50°C. Eluted peptides were analyzed by MS using data-dependent acquisition. Proteome fractions (60min run) were separated by a 39min acetonitrile gradient (11-40% of 80% ACN/0.125% formic acid for reverse phase basic fractions 1-9; 15-44% of 80% ACN/0.125% formic acid for reverse phase basic fractions 10-12) followed by a 3min gradient to 75% of 80% ACN/0.125% formic acid. Phosphoproteome samples (120min run) were separated by a 90min acetonitrile gradient (11-38% of 80% ACN/0.125% formic acid) followed by a 10min gradient to 63% of 80% ACN/0.125% formic acid. Each MS cycle consisted of a one full MS1 scan acquired on the orbitrap at 70,000 resolution from 300 to 1500 m/z using an AGC of 3e6 and maximum injection

time of 100ms followed by MS/MS dependent scans on top 20 most intense precursor m/z ions, using 1.6 m/z isolation window, HCD fragmentation at 26 normalized collision energy, and acquired at 17,500 resolution on the orbitrap with an AGC of 5e4 (peptide samples) and AGC of 1e5 (phosphopeptide samples) and a maximum injection time of 50ms. Dynamic exclusion was set 40sec to exclude precursors from repeated MS/MS. Phosphopeptide samples were injected twice and the technical replicate data were combined to a single biological replicate.

Database searching, peptide quantification, phosphosite localization, and R_{TO} calculation

MaxQuant (v.1.6.7.0)[48] database searching software was used to obtain peptide identifications and quantifications from the proteome and phosphoproteome MS data files. MS spectra were searched against a *S.cerevisiae* (SGD, download date July 2014) using LysC enzyme specificity (cleavage at lysine C-terminus, max 2 missed cleavages), and 20 ppm mass tolerance for MS1 and MS/MS. The following modifications were considered: variable modifications of acetylation on protein N-termini, oxidation on methionines, and $^{13}C_6{}^{15}N_2$-Lysine8 on lysines and fixed modification of carbamidomethyl on cysteines. Phosphoproteome samples contained the additional variable modification of phosphorylation on serine, threonine, and tyrosine residues. Search results were filtered to 1% FDR at PSM and protein level.

MS1 precursor intensities for heavy and light peptides were extracted from the evidence.txt file. MS1 intensities that belong to the same peptide sequence across different charge states and fractions were summed up. Phosphopeptide precursor intensities were aggregated to the phosphopeptide isoform level by summing features corresponding to the same peptide sequence across charge states and fractions as well as overlapping peptide sequences sharing the same combination of modifications. Phosphopeptide isoforms we filtered for a maximum localization probability > 75% for at least one site.

R_{TO} values were calculated as log_2 ratios of the precursor intensity for the heavy-lysine containing abundance divided by the respective precursor intensity for the light-lysine containing abundance. Peptide R_{TO} distributions were median normalized to 0 (median of median protein replicate R_{TO} was 0.0112). The same correction value derived for each proteome replicate was applied to normalize its corresponding phosphopeptide isoform R_{TO} distributions. Peptides and phosphopeptide isoforms that reflected the 5% highest R_{TO} standard deviation across its 6 replicates were excluded from downstream analysis, as a quality control criteria to ensure the data reflected good quantifications. Unbiased removal of peptides with high variability across replicates ensured more accurate unmodified protein R_{TO}, with equivalent filtering applied unbiasedly to phosphorylated isoforms. Unmodified protein R_{TO} values were calculated as the median of its peptide R_{TO}, requiring at least 2 peptides per protein with each peptide observed in at least two replicates. Only peptides unique to their respective protein were considered. Unmodified proteins that demonstrated an average standard deviation of peptide $R_{TO} > 0.35$ (worst 7.2%, cut-off removing distribution tail) across replicates were removed to ensure median unmodified protein R_{TO} calculation more accurately reflected its unmodified counterpart proteoform. The 7.2% of proteins excluded from this dataset were likely proteins whose median R_{TO} calculation would not reflect accurate representation of the unmodified counterpart proteoform R_{TO}.

We performed a Welch's t-test comparing phosphopeptide isoform R_{TO} (max n=6) to protein R_{TO} (max n=6), assuming unequal variance. Same t-test was conducted for phosphorylation isoforms and their unmodified counterpart peptides across 6 replicates. Phosphopeptide isoforms and unmodified protein counterparts (or unmodified counterpart peptides) must have been observed in at least two replicates. The Benjamini-Hochberg method[49]

was implemented to adjust P-values to control for multiple hypothesis testing. All data analysis was conducted using R (v3.6.1) (https://www.r-project.org/). The consolidation approach and statistical tests herein were similarly applied to address thermal stability in yeast[33].

Bioinformatic analysis of phosphorylation isoforms

All bioinformatic analysis was performed using R (v3.6.1) (https://www.r-project.org/). Sequence analysis was conducted using IceLogo for sequence (+/- 7 amino acid surrounding phospho-acceptor amino acid) comparing faster phosphorylation isoforms to the rest (not significant and slower phosphorylation isoforms) [66]. Human to yeast orthologs were obtained from www.ensembl.org/biomart/martview/. Yeast orthologs were aligned to human ortholog counterparts via pairwiseAlignment function (substitution matrix = BLOSUM62) of the BioStrings R package [67], and amino acids with matching yeast phosphorylation acceptor amino acids to its human ortholog counterpart were extracted. Functional scores based on cellular fitness from Ochoa *et al.* [68] were mapped to yeast ortholog matched phosphorylation acceptor amino acids. Fisher Exact tests were conducted via counts in R. Structural predictions and solvent accessibility predictions (burial designation based on less than 5% exposure) were generated using the JPred4 web server [69] with yeast protein FASTA sequences. SIFT scores were generated for phosphorylation acceptor amino acid to alanine mutations using the mutfunc resource [58]. Phosphodegrons were matched to Swaney *et al.* dataset [28], and relative stability (R_s) values for matching phosphorylation isoforms was extracted from Smith *et al.* [33] dataset.

Our estimated stoichiometry calculation considers the assumption that there are two forms of the protein: the unmodified protein counterpart (particularly high stoichiometry) and the phosphorylated isoform (particularly low stoichiometry). Stoichiometry calculation is most sensitive for phosphorylation isoforms with large ΔR_{TO} from unmodified protein. The

phosphorylated isoform, unmodified counterpart peptide, and the unmodified protein R_{TO} values can be used to calculated stoichiometry using the following equations:

$$(1) \quad (R_{TO}{}^{Phospho} * Stoich^{Phosp}) + (R_{TO}{}^{Unmod} * Stoich^{Unmod}) = R_{TO}{}^{Prot} * 1$$

$$(2) \quad Stoich^{Phosp} + Stoich^{Unmod} = 1$$

Thus,

$$(3) \quad Stoich^{Phosp} = \frac{(R_{TO}{}^{Prot} - R_{TO}{}^{Unmod})}{(R_{TO}{}^{Phospho} - R_{TO}{}^{Unmod})}$$

Stoichiometry general trends can be extracted visually by plotting ΔR_{TO} (phosphorylated isoform - protein) vs ΔR_{TO} (unmodified counterpart peptide - protein). Small changes in ΔR_{TO} (unmodified counterpart peptide - protein) and large changes in ΔR_{TO} (phosphorylated isoform - protein) suggest low stoichiometry phosphorylation.

Phosphorylation protein turnover kinetic simulations

Theory and standard equations in extended methods document (Appendix B Supplementary Equations for Simulations). Accompanying R markdown (simulations) and extended method contain all equations for models used herein with simulations for figures. Two protein turnover kinetic models highlighted in Figure 3.3 and Appendix B Supplementary Figure 3.3-3.4: (Model 1) traditional turnover with phosphorylation extension; (Model 2) age-biased phosphorylation model. Simulations use small time step estimations (6 seconds) iterating through respective models, proceeding in order since protein synthesis (first is unmodified: Model 1; "newer" unmodified pool: Model 2). Simulator starts first incorporating all rates pertaining to the protein molecules added to the pool (input), followed by a new heavy/light proportion calculation, and finally the new proportion of the protein molecules is used for protein molecules that are removed from the pool related to output rates. This simulation is continued to exhaustion for all protein molecule pools in the synthesis order (from translation) for each model. If output

of a protein pool is used for the input calculation for a protein pool earlier since synthesis, then those same protein molecules are removed during the input step for that protein pool in order to calculate the new protein pool heavy/light proportion. For example, in the traditional model (Model 1), the rates for dephosphorylation of the modified protein are used in the input calculation step for the unmodified protein to calculate the new heavy/light proportion. Thus, for the modified protein pool, all input rates are considered and the output rate of dephosphorylation to calculate the new modified protein heavy/light proportion. Time step estimates are repeated sequentially to simulate delta observed unmodified protein comparisons to phosphorylation isoforms via a ΔR_{TO}.

Data availability

The mass spectrometry proteomics data generated for this manuscript will be deposited to the ProteomeXchange Consortium via the PRIDE partner repository upon publication.

3.4 RESULTS

3.4.1 DYNAMIC SILAC COUPLED TO PHOSPHOPROTEOMICS ENABLES MEASUREMENT OF PHOSPHORYLATED PROTEOFORM TURNOVER PROTEOME-WIDE

To experimentally measure protein turnover, we implemented a dynamic SILAC approach[27] across six replicates of *Saccharomyces cerevisiae* (Figure 3.1a). Following cell culture expression of light lysine-containing proteins (Lys0), a 90-minute pulse of isotopically heavy lysine (Lys8) partitioned a pool of the newly synthesized proteins. After proteome digestion, peptides and enriched phosphopeptides are subjected to LC-MS/MS to measure heavy and light labeled peptide intensities. We calculated the ratio of MS intensities for heavy over

light SILAC labeled peptides and phosphopeptides, which can accurately proxy protein turnover from a single time-point[29].

Generally in a dynamic SILAC approach, ratios of heavy/light labeled peptide intensities are converted to a protein half-life ($t_{1/2}$) by fitting to a first-order kinetic model. In order to accurately calculate half-life, both the phosphorylated isoform and protein need to be synthesized and degraded independently. However, the synthesis of a phosphorylation isoform often succeeds the synthesis of its unmodified protein, rejecting this protein half-life assumption. Thus, we use relative protein turnover, R_{TO}, as a readout for protein turnover to avoid incorrect use of protein half-life nomenclature. We calculate a relative turnover, R_{TO}, as the $\log_2$ ratio of newly synthesized over the pre-existing peptide, or $\log_2$(heavy/light) labeled peptide MS intensities[70,71].

In our "bottom-up" MS approach, we incorporate the accepted "peptidoform"[10,71] nomenclature to associate a peptide's R_{TO} readout to all protein molecules that contain that peptide. For phosphorylation isoforms and unmodified proteins, phosphopeptides and unmodified peptides, respectively, are used for proteoform comparison. Herein, we compare phosphorylated proteoforms to their unmodified proteoform counterparts (all peptides not phosphorylated at its matching amino acid location) or protein (median of all non-phosphorylated peptides from its protein).

Across the experiment, we calculated R_{TO} for ~2,800 localized phosphorylation proteoforms and ~3,100 proteins (Figure 3.1b). High pairwise correlations of R_{TO} for the replicates of the proteome (average pairwise replicate $R^2 = 0.93$) and phosphoproteome (average pairwise replicate $R^2 = 0.94$) suggest that the method was highly reproducible (Figure 3.1c). Proteins that contain an observed phosphorylated proteoform R_{TO} in the dataset demonstrate a faster turnover than proteins that do not have an observed phosphorylation proteoform,

suggesting that phosphorylation could act as a biological mechanism to alter protein turnover (Figure 3.1d). This observation was observed in humans[71] and is not casual because it could be due to other factors not specific to phosphorylation. However, together with phosphorylation's known role in the ubiquitin-proteasome system, we hypothesized that phosphorylation would largely have a faster turnover than its protein. Surprisingly, the global distribution of phosphorylation proteoforms was significantly lower in R_{TO} compared to the distribution of their cognate proteins (Figure 3.1e), necessitating the comparison between phosphorylation proteoforms and their proteins.

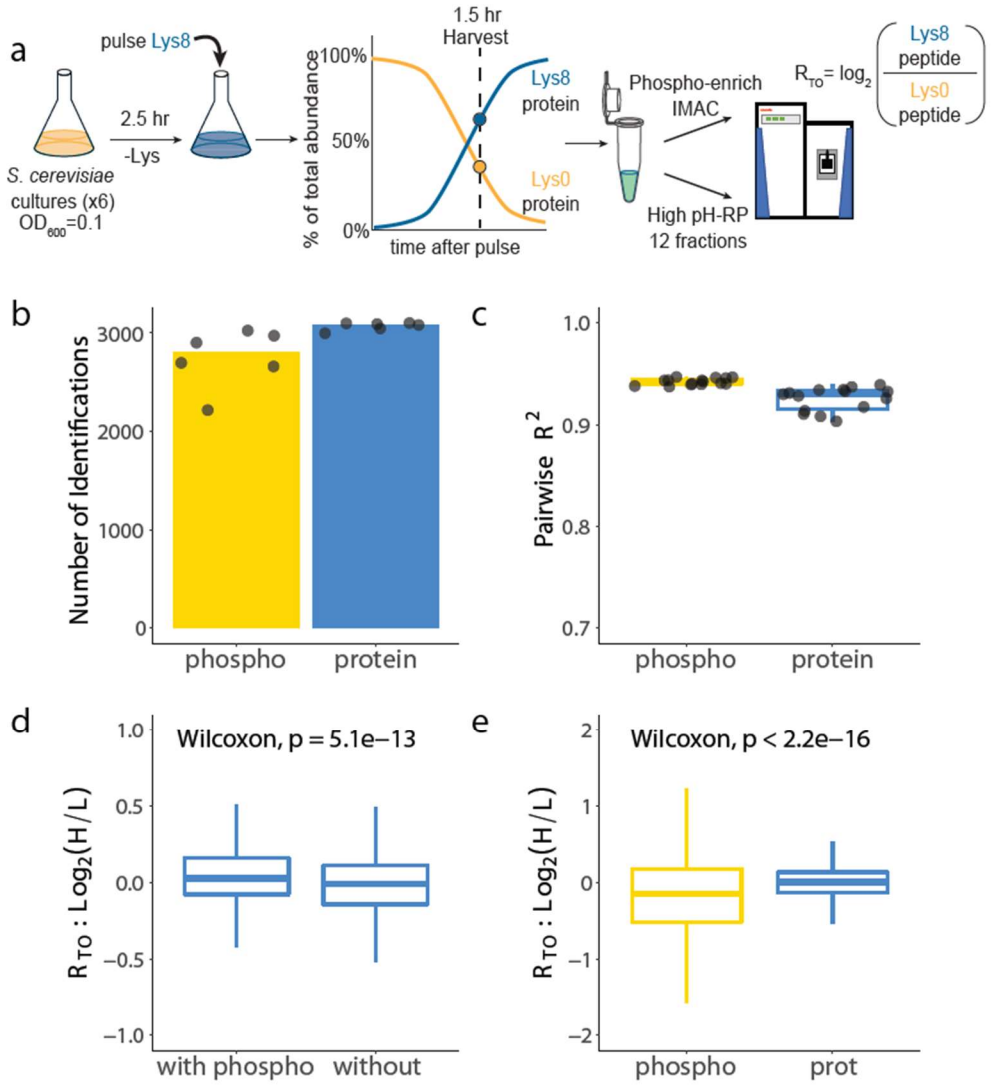


Figure 3.1: Dynamic SILAC labeling to calculate relative turnover (R_{TO}) of phosphorylation proteoforms and proteins. a) Dynamic SILAC labeling coupled with phosphoproteomics in yeast. 6 replicates of prototrophic *S. cerevisiae* cultures (seeded at OD_{600}=0.1) are grown in lysine deficient media for 2.5 hours and then pulsed with heavy lysine (Lys8). Cultures are harvested 1.5 hours after pulse and protein pool is digested with LysC. Resulting peptides were subjected to high pH-RP fractionation (12 for proteome) or IMAC-enrichment (for phosphopeptides) and analyzed by LC-MS/MS to generate peptide-level relative turnover (R_{TO}). **b)** Number of unique proteins and phosphorylation proteoforms identified across 6 replicate cultures. **c)** Boxplot depicting pairwise Pearson correlations (R^2) between replicate cultures for phosphorylation proteoforms and proteins. **d)** Boxplot comparison of R_{TO} of proteins that contain an observed phosphorylation proteoforms R_{TO} vs R_{TO} of protein without an observed phosphorylation proteoforms (wilcoxon test, proteins with phosphorylation n=1,031, proteins without phosphorylation n=2,093, P-value = 5.1e-13). **e)** Boxplot comparison of R_{TO} distributions of phosphorylation proteoforms and proteins (wilcoxon test, proteins n=3,124, phosphorylation isoforms n=3,479, P-value < 2.2e-16). Median R_{TO} across replicates was used for each phosphorylation proteoform and protein.

3.4.2 MANY PHOSPHORYLATION PROTEOFORMS HAVE APPARENT DIFFERENCES IN PROTEIN TURNOVER

To identify phosphorylation events that alter R_{TO}, phosphorylation proteoforms and counterpart proteoforms R_{TO} should be compared. However, coverage of phosphorylation proteoforms and their matching counterparts was low, about 40.2%. This is likely because peptide counterparts cannot be enriched over the complexity of the proteome background, potentially making MS detection difficult. Thus, similarly to other studies using thermal stability and protein turnover, we additionally compared phosphorylation proteoforms to their cognate proteins for differences in R_{TO}[33,38,71,72].

Protein R_{TO} readouts can serve as valuable proxies for counterpart proteoforms when the counterpart proteoform is high stoichiometry and the phosphorylation proteoform is low stoichiometry, a notion that has experimental support[73,74]. When the phosphorylation proteoform is high stoichiometry, it should mimic the R_{TO} of its protein, likely reporting the phosphorylation event as a false negative. Of note, some proteins have many modified proteoforms that can lead to a large spread in unmodified peptide R_{TO}, resulting in an inaccurate protein readout. Thus, we removed proteins with the largest standard deviations in R_{TO} of its unmodified peptide constituents (See Methods). The consolidated protein quantification acts as an averaged (median) R_{TO} of its two or more proteoforms.

To confirm the assumption that phosphorylation proteoforms in the dataset generally are low stoichiometry, a rough estimate of stoichiometry was determined from proportional R_{TO} contributions of a phosphorylated proteoform, its counterpart proteoform, and its protein. Of note, the stoichiometry estimates are more accurate for phosphorylation proteoforms with large

ΔR_{TO} differences from their protein. The stoichiometry plot demonstrates that most phosphorylation proteoforms with altered R_{TO} predominantly occur at low stoichiometry while their counterpart proteoforms generally mimic its protein R_{TO}, suggesting high stoichiometry (Appendix B Supplementary Figure 3.1a). Thus, we assert that the protein R_{TO} often reflects the counterpart's R_{TO} for the important phosphorylation proteoforms with large ΔR_{TO}.

Across the dataset, we observed that phosphorylation proteoforms tended to correlate poorly with their respective proteins ($R^2 = 0.32$, Figure 3.2a), suggesting that phosphorylation might act as a global modulator of R_{TO}. To ensure that the observed phosphorylation differences were not due to experimental bias in the phosphorylation enrichment protocol, we compared the R_{TO} of unmodified peptides observed in the phosphorylation-enriched sample to their matching unmodified peptide R_{TO} in the proteome sample. The R_{TO} of unmodified peptides in the phosphorylation-enriched (IMAC) sample strongly mimic the R_{TO} of the matching peptide in the proteome sample with a $R^2 = 0.83$, suggesting the differences are likely not technical (Figure 3.2b). This correlation is likely lower than its protein replicates correlations (Figure 3.1c) for two reasons: (1) unmodified peptide-level readouts inherently will have more variability than median-consolidated protein-level readouts and (2) unmodified peptides in the phosphorylation-enriched sample are predominantly lower abundance peptides (Appendix B Supplementary Figure 3.2a).

To further support the validity of using protein readouts for R_{TO} comparison, phosphorylation proteoform R_{TO} were compared to both their counterpart proteoform and their protein using a Welch's t-test (with FDR-corrected p-values). ΔR_{TO} phosphorylation proteoform R_{TO} differences ($R^2 = 0.91$) and significance calls from both tests were largely in agreement (Figure 3.2c). When comparing phosphorylated proteoforms to their proteins, we identified 253

phosphorylation proteoforms with faster R_{TO} and 1,172 with slower R_{TO} (Figure 3.2d). Comparatively to similar work in human cells[64,71], we observe a vast difference in the distribution of significance calls tending toward more significantly slower R_{TO} phosphorylation proteoforms in yeast. Given that human and yeast cell lines have vastly different doubling times, we set out to model the underlying kinetics of amino acid incorporation that could explain the distribution of significant calls in yeast. By applying kinetic modeling, we can also explore the putative functional context of phosphorylation proteoforms with altered R_{TO}.

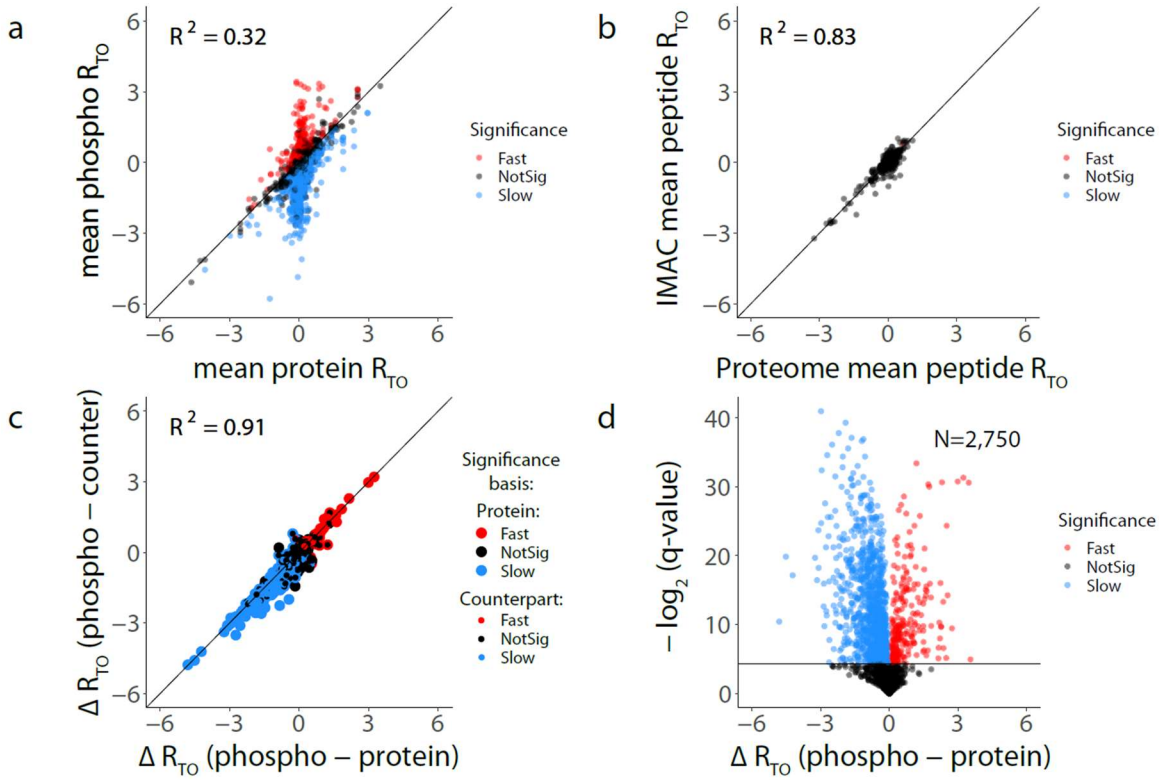


Figure 3.2: Phosphorylation proteoforms R_{TO} deviates largely from its protein. a) Scatterplot and Pearson correlation between R_{TO} of phosphorylation peptiodoforms (replicate average R_{TO} of n=6) and R_{TO} of its corresponding protein (replicate average R_{TO} of n=6). Significance calls (faster: red, slower: blue) were determined by comparing phosphorylation proteoform R_{TO} (max n=6) to R_{TO} for its protein (max n=6) via Welch's t-test. P-values were Benjamini-Hochberg corrected with significant hits called at q-value < 0.05. **b)** Scatterplot and Pearson correlation between R_{TO} of unmodified peptides observed in the phosphopeptide-enriched sample (replicate average R_{TO} of n=6) and R_{TO} of matching unmodified peptide in proteome samples (replicate average R_{TO} of n=6). Significance calls (faster: red, slower: blue) were determined by comparing unmodified peptide R_{TO} in the phosphorylation-enriched sample (max n=6) to R_{TO} for its protein in the proteome sample (max n=6) via Welch's t-test. P-values were Benjamini-Hochberg corrected with significant hits called at q-value < 0.05. **c)** Scatterplot and Pearson correlation between delta R_{TO} of (phosphorylation proteoform - its protein) (n=6) and R_{TO} of (phosphorylation proteoform - counterpart proteoform) (n=6). Large circles (protein) are based on the same significance calls as in (**a**). Smaller circles denoting significant calls (faster: red, slower: blue) determined by comparing phosphorylation petidoform R_{TO} (max n=6) to counterpart proteoform (max n=6) via Welch's t-test. P-values were Benjamini-Hochberg corrected with significant hits called at q-value < 0.05. **d)** Volcano plots showing differences in R_{TO} between phosphorylation proteoform (max n=6) and its protein (max n=6) (same significance calls as in (**a**), using Welch's t-test with P-values Benjamini-Hochberg, significant calls based on q-value < 0.05).

3.4.3 PROTEIN TURNOVER MODELING WITH PHOSPHORYLATION COMPLICATES THE ΔR_{TO} INTERPRETATION

Traditionally, heavy/light-lysine peptide abundance ratios (R_{TO}) proxy protein turnover as a snap-shot of newly-synthesized protein relative to pre-existing protein. Via steady-state kinetic equations, we can model the synthesis and degradation rates from the observed R_{TO} to calculate a protein half-life. These equations work under the constraints that proteins follow first-order kinetics, decay exponentially, and maintain a constant protein concentration in cells. Under these steady state constraints, there is no preferential degradation bias for "newer" or "older" versions of a protein. Here, we extended the traditionally used steady-state model to include phosphorylation, adding the relevant rates for the synthesis and degradation of the phosphorylated proteoform. Thus, the following rates were included: (1) the rate of phosphorylation of the unmodified protein to the phosphorylated proteoform likely via a kinase (r_{kinase}) (2) the rate of dephosphorylation of the phosphorylated proteoform to regenerate the unmodified protein likely via a phosphatase (r_{Ptase}) (3) the rate of phosphorylated proteoform loss from the system (cell) which is the combination of the rate of dilution (due to cell division: r_{dil}) and the degradation rate of the phosphorylated proteoform ($r_{deg(P)}$). To maintain steady state, we consider a constant stoichiometry of the phosphorylated proteoform to unmodified protein and an equal probability of phosphorylating heavy-("newer") and light-("older") lysine containing proteins (Figure 3.3a).

To visualize the impact of phosphorylation on R_{TO} readouts, we generated simulations using our protein turnover model which includes the kinetics of phosphorylation. Although our simulation parameters are set to specific estimated values, the general trends and relationships between rates presented herein should be consistent.

First, we wanted to ensure that our protein turnover model including phosphorylation can recapitulate known trends of unmodified protein turnover. We used the median protein R_{TO} of the proteome (R_{TO} = 0.0112) at 90 minutes, assigned a phosphorylation stoichiometry to a negligible amount, and set the doubling time to 122 minutes[65] to estimate an average contribution of protein degradation to protein turnover. Using our simulation, we estimated that the median protein has ~27% of the protein loss from the cell due to degradation. This estimate is in agreement that yeast proteome turnover is largely driven by dilution from cellular division[29,75] (Figure 3.3b).

Next, we addressed how the protein with median R_{TO} can alter its turnover due to a phosphorylation proteoform having accelerated degradation. As expected, when the degradation rate of the phosphorylated proteoform increases, the unmodified protein's turnover increases (Appendix B Supplementary Figure 3.3a). This observation agrees with findings in the literature that unmodified proteins with phosphodegron elements are more likely to have a faster unmodified protein turnover[28,29]. Additionally, we observed that the increase in unmodified protein R_{TO} scaled to the stoichiometry of the phosphorylation event. The higher the phosphorylation stoichiometry given the same phosphorylated proteoform degradation rate, the larger the increase in unmodified protein turnover.

Experimentally, we compared the R_{TO} of the phosphorylated proteoform to its protein in an attempt to capture differences in the degradation rates between them. We do so by conducting our statistical test, where we assume that the phosphorylated proteoform and its protein are sampled from independent populations. It is important to highlight that this is not the case; there is often a dependency that the generation of the phosphorylated proteoform succeeds unmodified protein synthesis. This order impacts the observed R_{TO} readout because there is a "kinetic lag" in

the R_{TO} of the phosphorylated proteoform. The "kinetic lag" is a result of the unmodified protein pool increasing in exclusively heavy protein molecules following the pulse, while the phosphorylated proteoform is synthesized by heavy and light protein molecules proportional to the ratio of heavy/light unmodified protein at that time. This relationship results in a large negative ΔR_{TO} even when the degradation rates of the unmodified protein and the phosphorylated proteoform are the same. This "kinetic lag" likely impacts the ΔR_{TO} globally for we observed a ΔR_{TO} distribution with a large skew towards a slower median ΔR_{TO} at -0.20 (Figure 3.3c).

Other factors, such as the phosphorylation proteoform degradation rate and the dephosphorylation rate, can impact the apparent ΔR_{TO} differences in our dataset. Increasing the phosphorylated proteoform's degradation rate to be much greater than the unmodified protein will make the ΔR_{TO} from the "kinetic lag" less negative (Figure 3.3d). Also, if the phosphorylation (r_{Kinase}) and dephosphorylation (r_{Ptase}) rates are acting much faster than the protein removal rates ($r_{dil} + r_{deg/(P)}$) of the phosphorylated proteoform and the unmodified protein, the unmodified protein pool will be mixing rapidly with the phosphorylated proteoform pool leading to the R_{TO} readouts mimicking each other (less negative ΔR_{TO}).

To visualize the impact of kinase/dephosphorylation activity (r_{Ptase}) and phosphorylated proteoform degradation ($r_{deg(P)}$) rates on ΔR_{TO}, we simulated a large theoretical space of dephosphorylation rate and phosphorylated proteoform degradation rate combinations. For these rate combinations, we calculated an ΔR_{TO} that would be observed after a 90 minute SILAC pulse (Figure 3.3d and Figure 3.3e, phosphorylation stoichiometry 10%; Appendix B Supplementary Figure 3.4a-b at 1.5% and 20% respectively). The line in Figure 3.3d demonstrates the rate combinations that could result in the observed median ΔR_{TO} in our experiment, -0.20. We

observed that the same ΔR_{TO} can be ambiguously explained by either a prominent increase in the phosphorylation proteoform degradation rate ($r_{deg(P)}$), dephosphorylation rate (r_{Ptase}), or a combination of both. Importantly, high dephosphorylation rates can dominate the apparent ΔR_{TO}, minimizing the contribution of the biologically-meaningful degradation rate of the phosphorylated proteoform on ΔR_{TO}.

The ambiguity in the meaning of the negative ΔR_{TO} from the combinatorial impact of "kinetic lag", phosphoprotein degradation rates, and the dephosphorylation rates on the ΔR_{TO} makes the significantly slower R_{TO} events difficult to interpret. Also, when we conduct *in silico* phospho-inhibitory substitutions from the phospho-acceptor amino acid to an alanine, there is a higher tolerance for the amino acid substitution in the significantly slower R_{TO} phosphorylation preptidoforms than its faster R_{TO} and not significant counterparts (Figure 3.3f). A similar trend was observed in human cell lines[64]. This observation further supports that significantly slower R_{TO} phosphorylation proteoforms likely do not prioritize a functional impact on protein turnover, and we assert that it is feasible that the observed ΔR_{TO} is driven by kinetics unrelated to the phosphorylation proteoform degradation rates.

To put simply, the "kinetic lag" derives a large negative R_{TO} (technical effect), and increased dephosphorylation and phosphorylation proteoform degradation rates both can increase R_{TO}. Since the significantly slower R_{TO} sites are difficult to interpret, we withdraw the notion that these sites might be related to slower phosphorylation proteoform turnover or have any biological reason for prioritization in yeast.

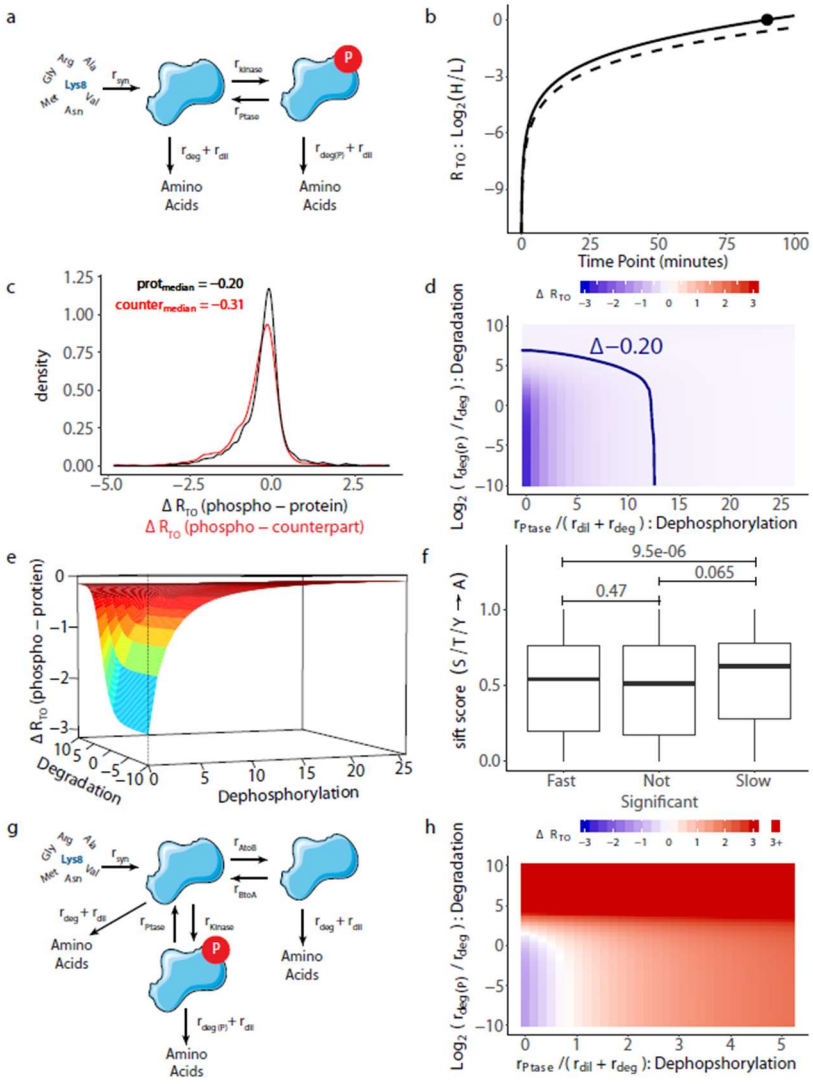


Figure 3.3: Kinetic analysis of traditional and age-dependent models for protein turnover with phosphorylation. a) Traditional model for protein turnover with the extension of phosphorylation. Following a heavy lysine pulse, unmodified protein is synthesized at rate (r_{syn}). Unmodified protein can either be removed from the cell via a combination of the rates of dilution (r_{dil}) and protein degradation (r_{deg}) or phosphorylated at the rate (r_{kinase}) resulting in the phosphorylated isoform. Phosphorylated isoform is removed from the cell by the combination of the rates of dilution (r_{dil}) and phosphorylated isoform specific degradation ($r_{deg(P)}$) or dephosphorylated back to unmodified protein at the rate (r_{Ptase}). **b)** R_{TO} readout based on simulation of the time after Lys8 pulse. Dashed line is the R_{TO} readout based on r_{dil} only impacting protein removal from the cell division. Solid line is including the proportion of unmodified protein degradation rate (r_{deg}) to match the median observed protein R_{TO} from the proteome (point; R_{TO}= 0.0112) at the harvest time (90 min). Stoichiometry of phosphorylated isoform was set to negligible amount (0.00001) and r_{dil} was set to cellular doubling (122 min^{-1}) to capture an averaged estimate of contribution of protein degradation on protein turnover. **c)** Distribution of delta R_{TO} of (phosphorylation proteoform - its protein) (black) and delta R_{TO} of (phosphorylation proteoform - counterpart proteoform) (red) across all observed phosphorylated proteoforms for each set of pairs (n=2,750 protein, n=1,098 counterpart). **d)** Simulation of ΔR_{TO} of (phosphorylation proteoform - its protein) (color: blue < 0, red > 0) for range of phosphatase rates and phosphorylation isoform degradation rates using the traditional model defined in (**a**). Phosphatase rate is relative to unmodified protein removal from the cell rates ($r_{dil} + r_{deg}$), and phosphorylation degradation rate ($r_{deg(P)}$) is depicted relative to unmodified protein degradation rate (r_{deg}). Phosphorylation stoichiometry was set to 10%, unmodified degradation rate (r_{deg}) was set to average degradation of proteome (solid line from (**b**)), and r_{dil} was set to cellular doubling. Blue line represents rate combinations that result in -0.20 ΔR_{TO}. **e)** Same data as in (**d**) with the color dimension (ΔR_{TO} (phosphorylation proteoform - its protein)) projected in the z-dimension. Simulation data depicted in a plane approaches but never crosses 0 for ΔR_{TO} (phosphorylation proteoform - its protein). **f)** SIFT score distribution for mutations of phosphorylation acceptor amino acid to alanine of different phosphorylation proteoforms categorized by their ΔR_{TO} significance calls (wilcoxon test, Fast n=239, Slow n=1,031, Not Significant n= 1,314). A lower SIFT score indicates a higher likelihood of a deleterious effect upon mutation. **g)** Protein turnover model with age-biased phosphorylation. Synthesis (r_{syn}) of intermediate unmodified protein pool A is followed by either phosphorylation isoform generation (r_{Kinase}), generation of unmodified protein pool B (r_{AtoB}), or removal from the cell ($r_{dil} + r_{deg}$). Unmodified protein pool B can be removed from the cell ($r_{dil} + r_{deg}$) or reverted to intermediate protein pool A (r_{BtoA}). Phosphorylation isoform can be removed from the cell ($r_{dil} + r_{deg(P)}$) or dephosphorylated (r_{Ptase}). **h)** Simulation of ΔR_{TO} of (phosphorylation isoform - its protein) (color: blue < 0, red > 0) for range of phosphatase rates and phosphorylation isoform degradation rates for the age-biased phosphorylation model in (**g**). Phosphatase rate (r_{Ptase}) is relative to unmodified protein removal from the cell rates ($r_{dil} + r_{deg}$), and phosphorylation degradation rate ($r_{deg(P)}$) is depicted relative to unmodified counterpart degradation rate (r_{deg}). Phosphorylation, unmodified protein pool A, and unmodified protein pool B stoichiometry was set to 10%, 20%, 70% (respectively), unmodified degradation rate (r_{deg}) was set to average degradation of proteome (solid line from (B)), r_{BtoA} = 0, and r_{dil} was set to cellular doubling.

3.4.4 AGE-BIASED PHOSPHORYLATION MODEL COULD EXPLAIN ΔR_{TO} PHOSPHORYLATION ISOFORMS

Next, we looked at the putative functional context of the observed faster R_{TO} sites. Interestingly, over the theoretical space of the traditional model, the ΔR_{TO} readouts approach but never pass 0 (Figure 3.3e), therefore suggesting faster R_{TO} phosphorylation proteoforms cannot exist given this model. The model cannot account for faster R_{TO} phosphorylation proteoforms because of the required steady state assumption that there is an equal probability for the phosphorylation event to occur on a "newer" (more heavy) or "older" (more light) unmodified protein. Therefore, at any given moment, newly generated phosphorylated isoforms cannot reflect more than the proportion of heavy and light of the unmodified protein. Even at extremely fast phosphorylation isoform degradation rates, the phosphorylation proteoform ratio of heavy/light-lysine containing proteins cannot exceed the unmodified protein ratio. Thus, our steady state model cannot explain the faster R_{TO} phosphorylation events, and an alternative explanation is required to account for these faster phosphorylation proteoforms.

One possible explanation for the presence of faster R_{TO} phosphorylation sites is a bias for newly synthesized proteins to have a higher probability of being phosphorylated compared to pre-existing protein molecules. This notion can be reflected in an alternative steady-state model in which unmodified proteins are synthesized first as an initial unmodified pool (Pool A) early in its "molecular lifetime". Then, Pool A can either be phosphorylated or transferred to a predominately "older" pool of unmodified protein (Pool B) (Figure 3.3g). Biologically, the differences between predominantly "newer" pool A and "older" pool B of the unmodified proteins can represent different "molecular fates", possibly driven by distinct subcellular localization or a distinct protein complex, etc. In this scenario, both the phosphorylated isoform

and the unmodified protein pool B experience a “kinetic lag” with respect to unmodified protein pool A. However, when the unmodified protein pool B is of predominant stoichiometry, the phosphoprotein pool will “lag” less quickly. With the addition of an increased dephosphorylation rate (r_{Ptase}) and/or an increased phosphorylation isoform degradation rate ($r_{deg(P)}$), the phosphorylation isoform can display a faster R_{TO} compared to the unmodified protein pools (Pool A and B combined).

Figure 3.3h outlines an age-biased phosphorylation scenario, where the phosphorylation isoform is 10% stoichiometry and unmodified protein pools A and B were 20% and 70% stoichiometry respectively (Figure 3.3h; Appendix B Supplementary Figure 3.4c at 1.5%/20%/98.5% and Appendix B Supplementary Figure 3.4d at 20%/20%/60% phosphorylation isoform/pool A/pool B stoichiometries respectively). The rate of protein pool B to protein pool A (r_{BtoA}) was set to 0 for simplicity and the dephosphorylation (r_{Ptase}) and phosphorylation isoform degradation rates ($r_{deg(P)}$) were varied (similarly to Figure 3.3d). Over this theoretical space, we observe a range of rate combinations in which a ΔR_{TO} can be greater than 0. Like observed in the traditional steady-state model, high dephosphorylation rates (r_{Ptase}) or fast phosphorylation isoform degradation rates ($r_{deg(P)}$) can explain the same observed ΔR_{TO}, with the added possibility of faster R_{TO}. Thus, we cannot attribute faster R_{TO} phosphorylation isoforms to increased degradation ($r_{deg(P)}$) because increased dephosphorylation rates (r_{Ptase}) could drive the readout. However, we posit that the age-bias phosphorylation and the separate pools of unmodified proteins could pertain to biological differences between the unmodified protein pool A that is selectively phosphorylated and the higher stoichiometry unmodified protein pool B.

3.4.5 PROPERTIES OF FASTER TURNOVER PHOSPHORYLATION PROTEOFORMS

Having established that our age-based model could explain faster R_{TO} phosphorylation isoforms, we next attempted to identify whether protein attributes associated with ΔR_{TO} differences of phosphorylation proteoforms. The attributes that we addressed include: phosphorylation "function", sequence properties, protein structural properties, and phosphorylation's observed impact on protein thermal stability.

For phosphorylation "function", a recent study[68] prioritized the function of phosphorylation events in humans. The authors used a machine learning approach to identify and integrate a diverse set of features to generate a functional score for each phosphosite, based on a site's role in cell fitness. We used sequence alignments to match phospho-acceptor amino acids of yeast orthologs to the functionally-scored human phosphorylation sites. The matched site's scores were used to assess whether our faster R_{TO} phosphorylation proteoforms were enriched in higher functional scores. We observed that the faster R_{TO} phosphorylation sites in yeast did not display a higher functional score than their non-significant and slower R_{TO} phosphorylation counterparts (Figure 3.4a). We are not surprised by this observation for two reasons. First, yeast and human orthologs often vary in sequence similarity, thus the two species are likely to have quite dissimilar effects on R_{TO} with a site-specific resolution. Secondly, the Ochoa et al. study highlighted that the model showed poor performance for phosphorylation sites that changed cellular localization. We hypothesized that some faster phosphorylation sites could derive their ΔR_{TO} differences from being in distinct pools due to different subcellular localizations.

Alternatively, when splitting the ΔR_{TO} distribution into thirds (faster, middle, slower ΔR_{TO} thirds), the faster phosphorylation proteoforms demonstrated a higher functional score than the middle and slower counterparts (Figure 3.4b). This observation suggests that having a faster

R_{TO} could be associated with having a higher functional score. The added phosphorylation events to the faster R_{TO} designation could have faster rates of kinase-phosphatase regulation and/or increased phosphorylation isoform degradation. These forms of protein regulation could have an association with cellular fitness and are more sensitive to being called “more functional”, despite these phosphorylation proteoforms not having a faster R_{TO} call.

For sequence analysis, we set out to look for enrichments/depletions in motifs and amino acid residues surrounding the phosphorylation acceptor site. Using a LOGO enrichment analysis, we identified that the faster R_{TO} phosphorylation proteoforms were enriched for glutamic acid and aspartic acid at the +3 position and depleted in proline at the -1 position (Figure 3.4c). A Fisher Exact test confirmed that faster phosphorylated proteoform motif profiles were indeed different (Figure 3.4d). The enrichment for a [S/T]XX[E/D] motif (consensus motif for Casein Kinase II) and a depletion in [S/T]P motif (consensus motif for CMGC family of kinases)[76] likely derives the motif profile differences between faster phosphorylation sites and the rest.

In regards to protein structural analysis, we observed that phosphorylation sites in predicted beta-sheets were likely to have a faster ΔR_{TO} than phosphorylation events in predicted coiled or alpha-helix domains (Figure 3.4e). Additionally, faster R_{TO} phosphorylation sites demonstrated an altered profile of occurring in predicted coiled, alpha-sheet, and beta-sheets compared to the rest of the phosphorylation sites (Figure 3.4f). This altered profile was likely driven by the increased proportion of faster R_{TO} phosphorylation sites in beta-sheets. For solvent accessibility, phosphorylation sites present in predicted buried regions had a faster ΔR_{TO} than exposed regions (Figure 3.4g).

Lastly, we explored the relationship between R_{TO} and another protein property associated with function, protein thermal stability. Previous work demonstrates that protein turnover and

protein thermal stability are not correlated for proteins[77]. Recently, using a method called Dali, we compared protein relative thermal stabilities (R_S) of phosphorylation isoforms to their protein counterparts proteome-wide[33]. The delta relative stability (ΔR_S) versus delta relative turnover (ΔR_{TO}) for phosphorylated proteoforms to their cognate proteins is not correlated (R^2 =0.0003, Figure 3.4h). As emphasized prior, the terms R_{TO} and protein turnover are not equivalent when applied to phosphorylation. This observation suggests that R_S and R_{TO} proxies are largely orthogonal. However, phosphorylation isoforms with significantly altered stability (increasing or decreasing in ΔR_S) demonstrated a faster ΔR_{TO} compared to not significant phosphorylated proteoforms (Figure 3.4i).

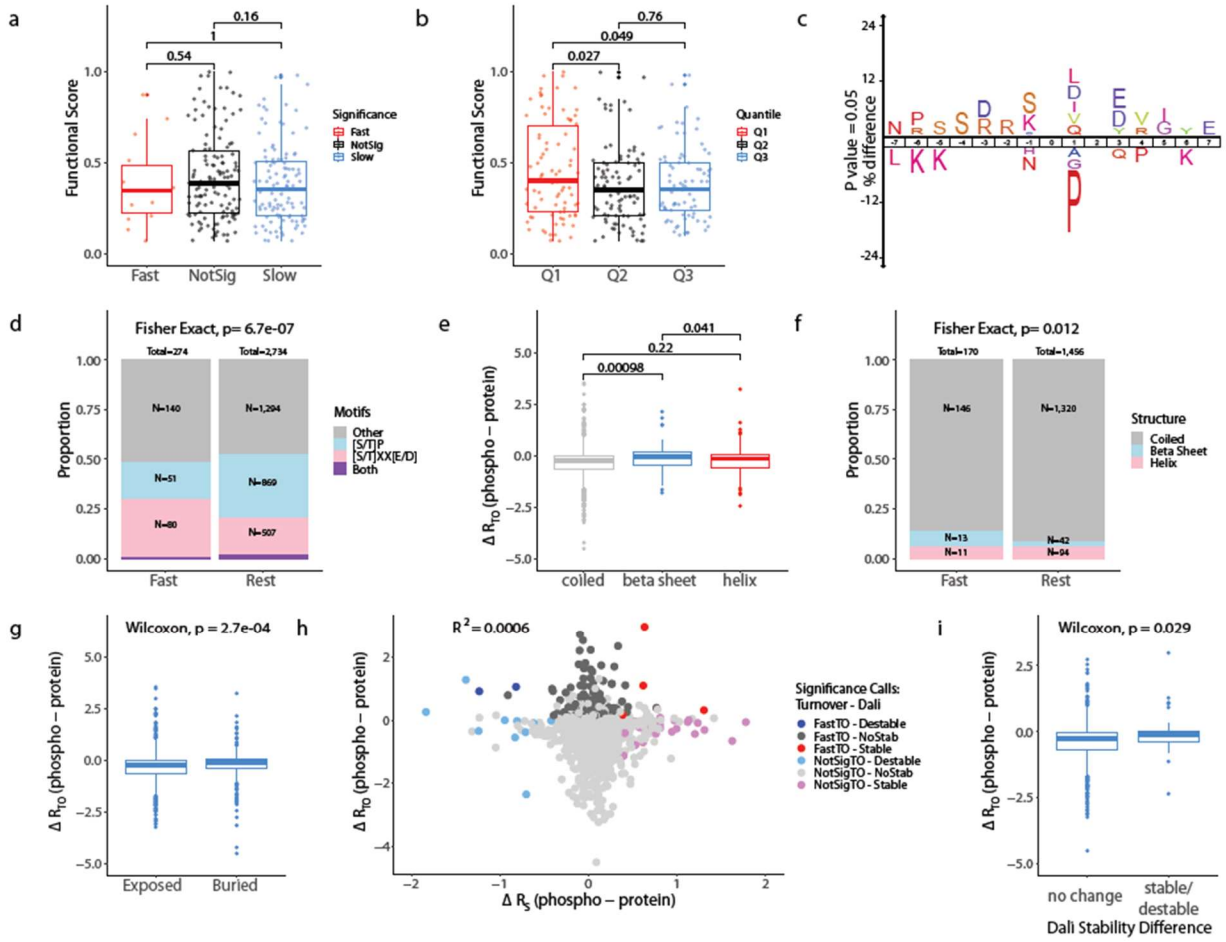


Figure 3.4: Properties of faster R_{TO} phospoisoforms. a) Functional scores from Ochoa et al.[68] annotated phosphorylation sites that successfully align to the same phospho-acceptor amino acid in yeast orthologs. Functional scores mapped from Ochoa et al. were compared for significant call classifications made from our dataset (based on protein comparisons). **b)** Same as (**a**) but phosphorylation phosphosites were categorized by phosphosites with fastest third (Q1), middle third (Q2), and slowest third (Q3) in delta R_{TO} (phosphorylated proteoform - protein). **c)** Analysis of +/- 7 amino acids flanking either side of phosphorylation acceptor site comparing faster phosphorylation proteoforms to the rest of the identified phosphorylation proteoforms. Enriched in faster on top and depleted on bottom based on percent identification with a p-value cutoff of 0.05. **d)** Fisher Exact test comparing faster and rest categorized phosphorylated proteoform motif profiles based on frequency of [S/T]P, [S/T]XX[E/D], neither or both motifs (p-value < 0.05). **e)** ΔR_{TO} comparison of phosphorylation sites based on amino acid acceptor site's secondary structure prediction (wilcoxon test). **f)** Fisher Exact test comparing faster and rest categorized phosphorylated peptiodoform's secondary structure profiles based on frequency of alpha-helix, beta-sheet, and coiled secondary structure classification (p-value < 0.05). **g)** ΔR_{TO} comparison of phosphorylated proteoforms based on amino acid acceptor site's solvent accessibility prediction (5% accessibility predictions, wilcoxon test). **h)** Scatterplot of ΔR_{TO} (this dataset) to relative stability (R_S) thermal stability delta (phospho-protein) from Smith *et al.* study[78]. Points colored by the combined significance classification from both studies. **i)** ΔR_{TO} comparison of phosphorylation proteoforms categorized as destabilizing or stabilizing vs the rest (from Smith *et al.*) that do not have an observed significant thermal stability difference between phosphorylation proteoform and its protein (wilcoxon test).

3.4.6 FASTER TURNOVER EVENTS WITH KNOWN ALTERED COMPLEXES OR SUBCELLULAR LOCALIZATION

Based on our age-biased phosphorylation hypothesis, the faster R_{TO} phosphorylation proteoforms likely reveal an underlying bias in phosphorylation that could be related to functional differences. Phosphorylation that can arise out from distinct pools of proteins could be attributed to different complex residencies and subcellular localizations that ascribe functional meaning. If the protein age-bias hypothesis is true, we still cannot ascertain whether the differences in R_{TO} are driven by phosphorylation acting as a modulator or a consequence of the two protein pools. So, we explored the literature for examples where known functional phosphorylation events that alter protein interactions or subcellular localization align with our hypothesis and could help ascribe meaning to the roles of some faster R_{TO} phosphorylation sites.

In regards to protein complexes, we found that ribosomal protein RPL12/uL11 phosphorylation at serine 38 demonstrated a faster R_{TO} compared to its counterpart and protein (Figure 3.5a). RPL12/uL11 S38 phosphorylation is an evolutionarily-conserved Cdc28 substrate[43] that is regulated during the cell cycle[44]. S38 phosphorylation is depleted in polysomes and serves a known functional role in regulating translation of a subset of mRNAs during mitosis[45]. In yeast, this site has been prioritized as functional for having an increased thermal stability compared to its protein[33]. This site's location proximal to the binding interface of elongation factor 2 (EF2) at the ribosome P-stalk, suggests a role in potentially coordinating the interaction with EF2[45]. This phosphosite faster R_{TO} adds further support that the phosphorylation proteoform is in a distinct pool of ribosomes from its unmodified counterpart. The site's increased thermal stability likely suggests that the phosphoprotein's faster turnover, while possibly age-biased, is likely not due to a faster phosphoprotein degradation rate, since protein

complexes often preserve stability and slow protein turnover[29]. In agreement with its regulation during the cell cycle, the faster R_{TO} is likely due to being regulated by high rates of Cdc28 phosphorylation and dephosphorylation by its phosphatase for tuning a cell cycle specific time resolution.

Additionally, distinct pools can be derived from discrete subcellular localizations. For example, according to a study by Tsang *et al.*[79], superoxide dismutase 1 (SOD1) translocates to the nucleus from the cytoplasm. Upon elevated levels of endogenous and exogenous reactive oxygen species, general ROS signaling mediates Mec1 phosphorylation and effector Dun1 kinase phosphorylation. Dun1 upon phosphorylation interacts with SOD1 and phosphorylates SOD1 at serines 60 and 99. Phosphorylated serines S60 and S99 result in SOD1 translocation to the nucleus which binds to promoters and regulates expression of oxidative resistance and repair genes to maintain genomic stability[79]. Although we did not observe S60 phosphorylation in our dataset, phosphorylated SOD1 S99 demonstrates a faster ΔR_{TO} suggesting that this site might be phosphorylated in an protein age-biased manner (Appendix B Supplementary Figure 3.5a). Based on the literature support, we likely observe the functional S99 phosphorylation isoform as faster turnover due to its altered nuclear localization residency of the "newly synthesized" phosphorylated isoform upon response to natural levels of endogenous reactive oxygen species in the cell.

Phosphorylation has also been observed to participate in crosstalk with ubiquitination. Swaney *et al.*[28] identified novel phosphodegrons, phosphorylation sites that function in a *cis*-regulatory manner to promote the subsequent ubiquitination and proteasome-dependent degradation. Swaney *et al.* identified phospho-ubiquitin crosstalk by identifying phosphorylation-ubiquitination co-modification pairs that demonstrate correlated accumulation

upon proteasome inhibition[28]. We found that two known phosphodegron sites Psd1 S253 (Figure 3.5b) and Cdc48 S519 (Appendix B Supplementary Figure 3.5b) have faster ΔR_{TO}. Based on our hypothesis, faster turnover phosphodegrons could suggest that their proteins (Psd1 and Cdc48) exist in two unmodified protein pools with the more newly synthesized proteins being phosphorylated. The observed phosphosite faster R_{TO} in our dataset could likely be explained by an increased degradation rate of the phosphorylated isoform compared to its predominantly "older" unmodified counterpart protein pool. The observation of a phosphodegron that does not have faster turnover does not present an inconsistency because even slower ΔR_{TO} sites observed in our dataset could have significantly faster phosphoprotein degradation rate (Figure 3.3d and Figure 3.3h). However, the overlap of faster turnover sites with known phosphodegrons could suggest a unique biological phenomenon for these phosphorylation events likely promoting unmodified protein degradation of a predominantly more newly synthesized unmodified protein pool. Of note, Cdc48 S519 phosphorylation is one of the two known functional phosphorylation sites, along with T674 phosphorylation, which coordinates complexing with ubiquitinated G1 cyclin Cln3, increasing Cln3's stability and releasing Cln3 from the endoplasmic reticulum to allow for Cln3's accumulation in the nucleus[80]. Through dual phospho-mimetics and phospho-inhibitory Cdc48 mutants to S519E T674E and S519A T674A respectively, phosphorylation at these sites was determined to stimulate a positive role in G1-cyclin activity for cell cycle entry[80]. Collectively, the functional role of the Cdc48 S519 phosphodegron in the cell cycle could be derived from specific phosphorylation of the predominately newly synthesized pool of unmodified Cdc48.

Lastly, the dual phosphorylated glycerol-3-phosphate dehydrogenase 1 (Gpd1) S24 S27 demonstrates the strongest support for our phosphorylation-age biased hypothesis. This dual

phosphorylation site was observed to have a significantly faster ΔR_{TO} (Figure 3.5c). In the literature, Gpd1 has been studied for its role in glycerol synthesis and stress response, particularly its role in osmotic stress[81]. Gpd1 has been observed to be distributed across the nucleus, cytosol, and the peroxisome during unstressed conditions in yeast[82]. There has been a direct link between phosphorylation regulation and Gpd1 localization and activity. The dephosphorylated Gpd1 increases the catalytic activity of Gpd1 in the cell[83] and the dual phospho-inhibitory mutant (dephosphorylated mimic) has an impaired import into the peroxisome[82]. Alternatively, the S24 S27 dual phosphorylated Gpd1 is linked to peroxisomal import[82] and is the less catalytically active form of the protein[83].

Upon hyperosmotic stress in the Jung *et al.* study[82], there is an increase in the accumulation of Gpd1 in the cytosol and the nucleus, and a depletion in the peroxisome. From an experiment halting translation coupled with hyperosmotic stress, the cytosolic accumulation of Gpd1 upon hyperosmotic stress was determined not to derive from export of the peroxisomal Gpd1 (phosphorylated), but from the newly synthesized Gpd1 protein which is not being phosphorylated, thus not being sent to the peroxisome[82]. This highlights two main points that align with our age-biased hypothesis. Firstly, the determination of the "molecular fates" of the unmodified Gpd1 to accumulate in the cytosol or be phosphorylated at S24 and S27 and transported to the peroxisome occurs early in the Gpd1's molecular "age" (following synthesis) based on the metabolic state of the cell. Secondly, upon "early" phosphorylation and subsequent transport to the peroxisome, the phosphorylated protein pool in the peroxisome is in a distinct pool/subcellular compartment from the cytosolic Gpd1. These distinct pools delineate their difference in Gpd1 function. The unphosphorylated, active Gpd1 in the cytoplasm serves the functional role generating increased glycerol production under osmotic stress[83]. The role of the

dual phosphorylated Gpd1 in the peroxisome is less clear but has been hypothesized that the fast process of Gpd1 peroxisomal import can rapidly attenuate the steady-state dosage of active Gpd1 in the cytosol and serve as a "stress-relief valve"[84]. Interestingly, the phosphorylated Gpd1 in the peroxisome does not transport to the cytosol upon hyper-osmotic stress, suggesting that these two pools actually define their determinate "molecular fates". Thus, the observed faster turnover of S24 S27 dually phosphorylated GPD1 contains literature support for being phosphorylated with a molecular age-bias that separates the proteoforms to different pools delineating the functional differences between them.

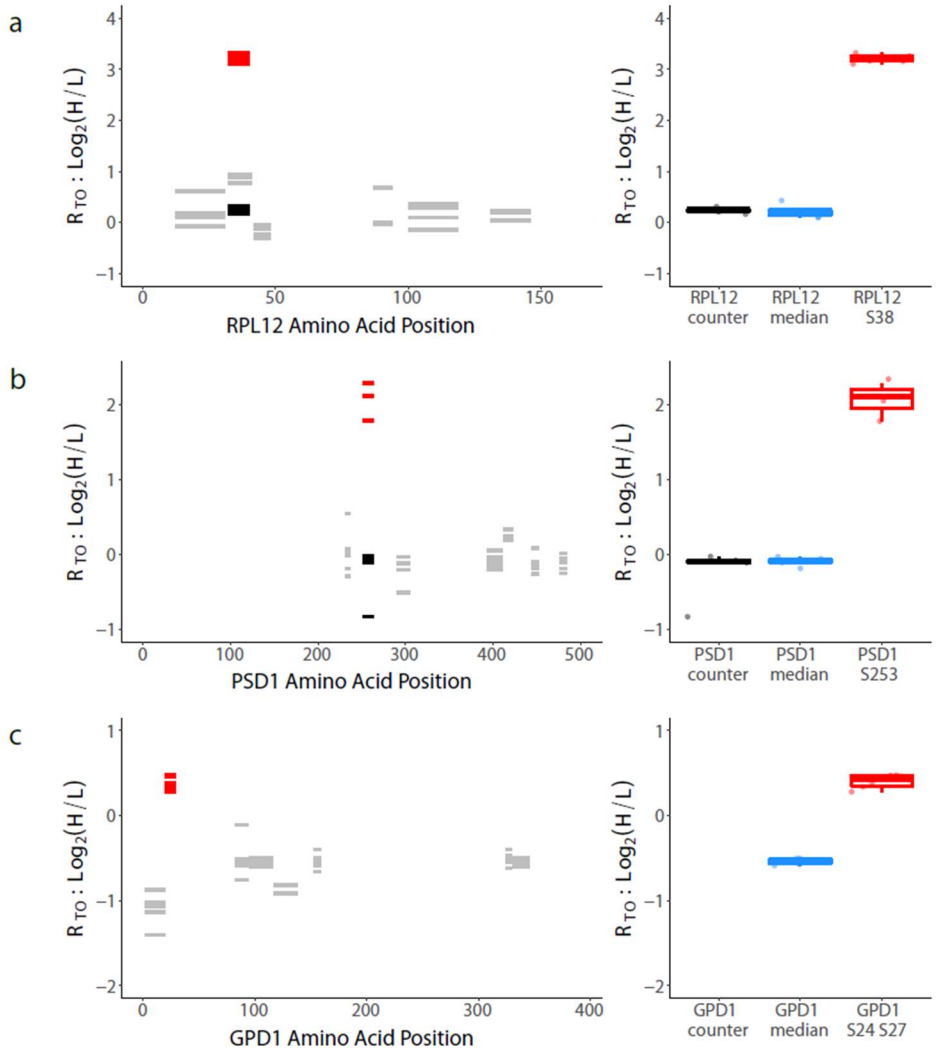


Figure 3.5: Known functional phosphorylation isoforms have faster R_{TO}. a) (*left*) Replicate R_{TO} values for observed RPL12 unmodified peptides identified in the proteome sample (grey), unmodified counterpart proteoform (black), and RPL12 phospho-serine 38 phosphorylation proteoform (red) displayed across the length of RPL12. (*right*) Boxplot of replicate R_{TO} for unmodified counterpart proteoform (black), protein (blue), and phosphorylated proteoform S38 (red). **b-c)** Same as (**a**) for PSD1 and its faster turnover phosphorylation proteoform S253 (**b**), and GPD1 and its faster turnover dual phosphorylation proteoform S24 S27 (**c**).

3.5 Discussion

With advances in mass spectrometry, the field of phosphoproteomics has achieved an unprecedented feat of cataloging 100,000s of phosphorylation sites across many organisms. However, there is an unmet need to identify which of these phosphorylation sites are functional and ultimately annotating what these functions are. Current experimental methods lack the

ability to assess function at high-throughput. In order to address this bottleneck, we coupled dynamic SILAC labeling with phosphoproteomics to measure phosphorylation proteoform turnover and protein turnover proteome-wide. Protein turnover has been observed to be sensitive to protein functions; many such functions are known to be modulated by phosphorylation. We performed proteoform comparisons between phosphorylated proteoforms and their proteins to identify phosphorylation events that alter relative protein turnover (R_{TO}).

Our kinetic analysis of experimentally-derived protein turnover uncovered that a dynamic SILAC labeling approach when applied to phosphorylation is impacted by the dependency between the phosphorylated proteoform and its unmodified counterpart. The synthesis of the unmodified protein likely precedes the synthesis of a phosphorylated isoform. This dependency can result in a large "kinetic lag" in the incorporation of the heavy lysine amino acid for phosphorylated proteoforms following the pulse, which could predominate the signal for "slower" R_{TO} phosphorylation proteoforms. Also, dynamic exchange of phosphate addition and removal from proteins could also obfuscate the biologically-meaningful phosphorylation isoform degradation rate on ΔR_{TO}. Thus, we withdrew the notion that the slower R_{TO} sites are biologically meaningful because they could derive from technical origins.

The proportion of slower phosphorylation proteoforms in yeast was much greater than observed in humans[64,71]. This does not invalidate or contradict their findings when we consider the kinetic differences in cellular doubling time between yeast and humans. Alternatively, this difference does align with notions presented here that the kinetics of the amino acid incorporation can play a large role in determining a phosphorylated proteoform's apparent turnover. In yeast, the impact of the "kinetic lag" is more prominent, and ample time is required for the amino acid to incorporate into phosphorylation proteoforms for its R_{TO} to "catch up" with

its protein R_{TO}. In humans, the impact of the "kinetic lag" is minimized due to its longer doubling time (decreased r_{dil}) and duration of heavy amino acid pulse.

Generally, human cell lines will have ΔR_{TO} measurements that should better reflect degradation rate differences between phosphorylation proteoform and its protein compared to yeast. However, the rates of phosphorylation and dephosphorylation, instead of the phosphorylation proteoform degradation rate, could drive the ΔR_{TO} differences observed in both yeast and human cells. This possibility could suggest the presence of false positives among the population of slower R_{TO} phosphorylation proteoforms in the other studies. Further experiments will be necessary to decouple the contributions of degradation rates from phosphorylation addition/removal rates to better understand the basis of ΔR_{TO} differences. Important to note, slower R_{TO} phosphorylation proteoforms could be driven by a decreased phosphorylation proteoform degradation rate. Thus, slower R_{TO} phosphorylation proteoforms with follow-up validation are accurate support for degradation differences attributed to the phosphorylation event.

We identified ~9% of phosphorylation proteoforms have faster R_{TO} than its protein. We found that our traditional kinetic model (Figure 3.3a) could not explain the presence of faster R_{TO} phosphorylation proteoforms. The synthesis dependency of the phosphorylated proteoform succeeding its unmodified protein underlies this constraint. Herein, we presented a possible explanation for "faster" R_{TO} phosphorylated proteoforms being derived from molecular age-biased phosphorylation. In this scenario, a pool of unmodified proteins that are predominantly "newer" in molecular age (since protein synthesis) are more likely to be phosphorylated than the predominantly "older" unmodified protein pool. We assert that biased phosphorylation towards a predominantly "newer" unmodified protein pool could be a consequence of the predominantly

"newer" and "older" unmodified proteins being in distinct protein pools, i.e. in different protein interactions or subcellular residencies.

The notion that unmodified proteins can undergo molecular age-biased effects has been observed in mouse and human cell lines[85]. Non-exponentially decaying (NED) proteins, which account for ~10% of the proteome, have two-phase degradation dynamics. NED proteins degrade more quickly early in their molecular lifetime, while a portion of the synthesized protein which finds a protein interaction partner demonstrates a secondary phase of slower degradation[85]. While the interplay of phosphorylation with this age-biased phenomenon was not explored, we argue this observation supports that unmodified proteins can form two distinct pools that are age-biased. Also, the separation of the pools could delineate the functional differences between them. Age-biased protein degradation and phosphorylation are early discoveries in the large web of molecular age-biased protein biology and its role in post-translational protein regulation, which warrants further exploration. While the link between our age-biased phosphorylation hypothesis and the functional role of a phosphosite is not feasible with our data, we can provide a narrowed scope of biological roles that these faster R_{TO} phosphosites may perform in the cell. Thus, faster R_{TO} phosphorylation proteoforms could serve as a useful prioritization criteria for individual site functional validation.

While our approach is advantageous for its high-throughput capacity to interrogate phosphorylation proteome-wide, our dynamic SILAC method does have limitations.

(1) Sometimes the assumption that the phosphorylated proteoform is low stoichiometry and the unmodified protein counterpart is high stoichiometry is incorrect. This likely results in false negatives, where a high stoichiometry phosphosites will likely match its protein R_{TO}, but have an altered R_{TO} to its counterpart proteoform. Also, the presence of false positives is possible

when a low stoichiometry phosphorylation proteoform does not have a faster R_{TO} to its low stoichiometry counterpart proteoform, but does for its protein. For example, Ste20 protein R_{TO} signal is dominated by a high stoichiometry modified proteoform over the same peptide sequence, contributing to a false positive assignment for Ste20 S562 phosphorylation proteoform (Appendix B Supplementary Figure 3.6a). This observation further highlights that the protein turnover readouts can largely be driven by high stoichiometry modified proteoforms, which is largely ignored in protein-level turnover studies. The most appropriate comparison would be measuring the R_{TO} of intact phosphorylated and unmodified counterpart proteoforms via top-down proteomics, however this approach is currently not feasible and not reliable at the same throughput.

(2) The use of data dependent acquisition (DDA) MS for R_{TO} readouts results in stochastic sampling of highly abundant phosphopeptides and unmodified peptides. Improved data completeness, increased coverage of proteoform counterparts, and increased quantification accuracy could be improved by the use of data independent acquisition (DIA) MS strategies[86].

(3) We suggest protein age-biased phosphorylation as a possible explanation for faster turnover sites, however alternative possibilities exist. For instance, faster R_{TO} phosphorylation proteoforms can arise from cellular age-biased protein dynamics. This could present as bias toward altered protein synthesis/degradation and/or elevated levels of phosphorylation in "younger" cells. Yeast asymmetrically divide, potentially introducing differences between mother and daughter cells. However, little is known about phosphorylation differences among them. Additionally, yeast cell culture should be dominated by younger cell populations, likely diminishing possible differences in batch culture. While cellular age-biased phosphorylation

would be an exciting conclusion for its aging implications, we currently lack the methods to address this in yeast.

(4) Single time point protein turnover calculations can be less accurate. Alternatively, the comparison between the two proteoforms are relative and internally controlled. Collecting samples at earlier time points might result in increased sensitivity for some faster R_{TO} sites, in which the phosphorylation isoform might be exclusively heavy lysine-containing by 90 minutes. However, lysine label mixing and a greater impact from "kinetic lag" could hamper the applicability of early time points.

Despite these limitations, our method offers many advantages to the community. Our approach enables high-throughput comparison of R_{TO} across thousands of phosphorylation proteoforms to their protein in a single experiment. The dynamic SILAC labeling approach can easily be extended to enrich other post-translational modifications (ubiquitination, methylation, acetylation, etc.)[71]. Conducting this experiment in yeast serves valuable contrasts to similar work performed in human cell lines. We can leverage these differences in amino acid incorporation kinetics between human and yeast and our kinetic models to better interpret the meaning behind phosphorylation proteoform R_{TO} values. Additionally, significantly faster R_{TO} phosphosites can serve as a useful prioritization criteria for follow-up functional validation. Lastly, our molecular age-biased phosphorylation interpretation could represent a novel role of phosphorylation in post-translational protein regulation that warrants further exploration.

Contributions: Experiments, simulations, and data analysis were performed by Ian Smith. Technical guidance was given by Miguel Martin-Perez and Ricard Rodriguez. Anthony Barente

assisted in the generation of the kinetic models. Judit Villén, Miguel Martin-Perez, and Ian Smith designed the experiments.

Chapter 4. IDENTIFICATION OF SARS-COV-2 NSP5 HOST PROTEASE SUBSTRATES BY PROTEIN TURNOVER AND THERMAL STABILITY

Author contributions: This project was a joint collaborative effort between Kyle Hess, Mario Leutert, and Ian Smith. Mario Leutert collected the experimental data for the Dynamic SILAC labeling experiment and Kyle Hess generated the Thermal Proteome Profiling (TPP) data. Ian Smith analyzed the Dynamic SILAC data and wrote up the Dynamic SILAC labeling results. Kyle Hess analyzed the TPP data and wrote up its corresponding results. Kyle Hess and Ian Smith contributed equally to the manuscript writing and figure generation. Co-first authorship will be given at the time of publication with the following order: Kyle Hess, Ian Smith, Mario Leutert.

4.1 ABSTRACT

The SARS-CoV-2 main protease, NSP5, is essential for viral propagation and cleaves both viral and host proteins with specificity towards the putative motif, LQ|[AS]. Here, we separately overexpressed GFP, catalytically inactive NSP5, and wildtype NSP5 in HEK293T cells. Following overexpression, proteomes were subjected to dynamic SILAC labeling to measure protein turnover and thermal proteome profiling (TPP) to measure protein thermal stability proteome-wide. Using these functional readouts, we identified hundreds of proteins with altered protein turnover or thermal stability exclusively in the presence of catalytically active NSP5. Proteins with the LQ|[AS] motif that demonstrated an altered protein turnover tended to have faster turnover in the presence of NSP5, supporting that NSP5 cleaved substrates generally are destabilized and have increased degradation upon cleavage. Using protein-level or peptide-

level readouts, we were able to identify candidate NSP5 substrates, many of which aligned with known NSP5 substrates and proteins containing the NSP5 motif. Peptide-level NSP5 cleavage detection could implicate functional differences in protein turnover and protein thermal stability among the substrate's cleaved protein products. In combination with N-terminomics, our protein thermal stability and protein turnover assays can unbiasedly catalog NSP5's protease substrates. Also, the methods presented herein enable unprecedented functional insight to the consequences of a protein cleavage event on the protein and its cleaved products.

4.2 Introduction

The positive-sense, single-stranded RNA virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), responsible for the disease COVID-19 continues to be a major focus of extensive research efforts to understand molecular mechanisms that underlie viral infection and to identify avenues for therapeutic intervention. Upon SARS-CoV-2 infection, the host undergoes dramatic cellular and subcellular restructuring leading to system-wide molecular changes to the transcriptome[87], proteome[87–91], ubiquitome[87], phosphoproteome[87,90,91], protein interactome[87,92,93], translatome[88], and proteome thermal stability[89]. Despite vaccines, SARS-CoV-2 infection can still evade our immune system and propagate to others, creating a need for effective antivirals to combat active infection. One of the leading candidates for therapeutic targets is SARS-CoV-2's main protease NSP5 (M^{pro}; chymotrypsin-like protease: $3CL^{pro}$), which is essential for viral replication.

SARS-CoV-2 genome contains ORF1 which is translated into two polyproteins that must be cleaved by NSP5 and a papin-like protease NSP3 (PL^{pro}) to generate the functional protein units for assembly of the essential replication complex. Due to the essential functions of NSP5 and its cleaved substrates for viral replication, inhibiting NSP5 function with covalent small

molecules[94,95] has been explored to prevent viral propagation. NSP5, whose sequence and function is conserved across coronaviruses[96,97], demonstrates catalytic specificity for substrates containing the putative LQ|[AS] motif[98].

Many proteomics methods have been employed to identify NSP5 protein interactions and protease substrates to both viral and host proteins. Initial work with AP-MS identified host interacting proteins, HDAC2 for NSP5 and TRMT1 and GPX for a catalytically dead NSP5 (C145A)[92]. To improve sensitivity of NSP5 interactors, N-terminomics MS approaches, in the context of viral infection[99] or cell lysates with dosed recombinant NSP5[98], have enabled the discovery of 100's of viral and host neo-N-terminus peptides. In a N-terminomics approach like TAILS[100,101], protease cleavage events generate neo-N-terminus peptides with newly accessible amine moieties that are chemically labeled. Compared to a condition lacking NSP5, increased abundance of these neo-N-terminus peptides in the presence of NSP5 indicates a high confidence protease substrate[98]. Despite the discovery of 100's of NSP5 cleavage events, extensive follow-up experiments are required to determine the impact of the NSP5 cleavage on the substrate's protein function. Also, NSP5 protein cleavages could generate neo-N-terminus peptides that are unable to be detected by MS resulting in missed cleavage substrates.

Alternatively, orthogonal proteomics techniques using protein turnover[31] and thermal stability[33] have been leveraged to identify cleaved proteoforms that occur naturally in human and yeast cells. These methods leverage differences in peptide-level readouts of protein turnover or thermal stability surrounding a breakpoint to identify protein cleavage events. An advantage of these approaches is that they do not require identification of the cleaved neo-N-terminus peptide, and the protein cleavages are detected over a small region of the protease substrate. Unique to these methods, one can measure differences in protein turnover or thermal stability of the

resultant cleaved polypeptide products, indicating changes in degradation or stability due to a protein cleavage event.

Herein, we assayed protein thermal stability and protein turnover proteome-wide in HEK293T cells overexpressing GFP, catalytically inactive NSP5 C145A variant, or wildtype NSP5. We observed global proteome changes in protein turnover and thermal stability which correlated with active NSP5 protease activity. We identified many known NSP5 host substrates with altered protein turnover and thermal stability at the protein-level or across known LQ|[AS] breakpoints at the peptide-level when wildtype NSP5 is present. Generally, proteins that contained the putative NSP5 motif had accelerated protein turnover and altered thermal stability, suggesting NSP5 cleavage likely increases protein degradation of its substrates. In tandem with neo-N-terminomics approaches, protein turnover and thermal stability approaches could complement and validate NSP5 protease substrates, while offering mechanistic insight into functional changes of its protease substrates upon cleavage.

4.3 METHODS

Transfection, overexpression, and dynamic SILAC labeling in HEK293T cells

HEK293T cells were seeded in 6-well plates at 0.3 x 10^6 cells per well. Following growth for about 1.5 days, transfection of HEK293T cells were carried out using the polyjet transfection reagent according to the manufacturer's recommendation. This includes: (1) media exchange 30 minutes prior to transfection, (2) DNA constructs for GFP, NSP5 C145A, and wildtype NSP5 were mixed 2:3 with transfection reagent in serum, antibiotics-free media, (3) and 10 minutes was allowed for transfection complexing prior to addition with cell lines. Transfection media was exchanged 16 hours post-transfection. At 24 hours post-transfection, HEK293T cells were washed three times with PBS and once with media containing $^{13}C_6{}^{15}N_2$-Lysine-8. Then, HEK293T cells were exchanged for media containing $^{13}C_6{}^{15}N_2$-Lysine-8 and grown for an additional 15 hours (39 hours post-transfection). Transfection efficiency was evaluated by microscopy on GFP control, which was determined to be ~80%. HEK293T cells were harvested by detaching in PBS and washed three times with PBS by centrifugation. HEK293T cell pellets were snap frozen and stored at -80ºC.

Protein turnover sample preparation

Frozen cell pellets were resuspended in a lysis buffer composed of 8 M urea, 150 mM NaCl, and 100 mM HEPES, pH 8.2. Cells were lysed by 3 cycles of 30s tip sonication on ice. Lysate protein concentration was measured by BCA assay. 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 10mM DTT for 15 min at room temperature. Lysates were processed on a KingFisher Flex

(Thermo Scientific) and digested with LysC (Wako Pure Chemicals Industries) using the R2-P1 protocol as described in the Leutert et al.[47] study.

Peptides were desalted and fractionated on StageTips[102] by basic reverse-phase using a stepwise gradient of increasing acetonitrile (5%, 10%, 15%, 20%, and 80%; designated as RPB1 through RPB5 respectively) in 0.1% NH_4OH.

Thermal Proteome Profiling (TPP) in crude cell extracts

Two replicates of pelleted HEK293T cells cultured in the same conditions discussed above, with the exception of the Lys8 pulse, were resuspended in 800 µl of native lysis buffer (1x PBS, 2 mM $MgCl_2$, 0.25x protease inhibitor) and lysed by four cycles of freezing in liquid nitrogen for 1 minute, followed by thawing at 35ºC for 1 min. At this point, protein concentration was checked by BCA. Lysates were snap frozen and stored at -80ºC until ready for further processing (all performed on the same day). Samples were diluted to 3.5 mg/mL with an additional native lysis buffer. These lysates were then aliquoted, 80 µl into PCR tubes on ice (12 PCR tubes for each replicate and each cell extract). PCR tubes were incubated on a thermal cycler in two phases: first, a 5-min incubation at 37ºC; second, a 5-min incubation at 10 different temperatures (37.0ºC, 39.5ºC, 42.4ºC, 46.3ºC, 50.1ºC, 53.8ºC, 57.6ºC, 61.5ºC, 64.4ºC, 67.0ºC) for 5 min. The 11th and 12th PCR tube was treated at 37.0ºC. After temperature treatment, lysates were incubated at room temperature for 5 min, followed by the addition of 10 µl of 10x soluble protein extraction buffer (1x PBS, 2 mM $MgCl_2$, 0.25x protease inhibitor, 8% NP40) to each of the 10 temperature-gradient samples and the 12th sample. The 11th PCR tube received 10 µl of 10x SDS extraction buffer (1x PBS, 2 mM $MgCl_2$, 0.25x protease inhibitor, 10% SDS). A final of 10 µl of 10x Benzonase solution (Millipore) was then added to each of the 11 PCR tubes (final concentration 25 U/mL) and left shaking for 1 hour at 4ºC. Samples were then

centrifuged at 4˚C for 1 hour at 17,100 x g. After centrifugation, 75 µl from the first 11 samples were mixed 1:1 with 2x denaturing buffer (9M urea, 50 mM HEPES pH 8.2, 100 mM NaCl, 10 mM DTT) and incubated at 55˚C for 30 minutes. The 12th sample was taken through BCA to get an estimate of protein concentration before digestion. All samples were then incubated in the dark with 15 mM iodoacetamide for 30 min to alkylate cysteines and the reaction was quenched with 5 mM DTT for 15 min at RT.

TPP sample sp3 sample clean-up and digestion

For each sample, 50 µg of reduced and alkylated protein lysate per channel were cleaned up using a modified SP3 protocol[47] and a robotic magnetic bead processor KingFisher™ Flex (Thermo Scientific). Briefly, a 1:1 mix of carboxylated paramagnetic beads (Sera-Mag SpeedBeads, CAT# 09-981-121, 09-981-123) at a concentration of 10 µg/µl was conditioned in water. A lysate-ethanol-bead mixture was incubated with 5 µg of beads per µg of protein (for a final of 0.25 µg/µl protein, 75% ethanol v/v) and washed 4 times with 200 µl of 80% ethanol. On bead digestion and elution was carried out in 125 µl of 50 mM HEPES buffer pH 8.2 using 1 µg LysC at 37˚C for 4h. A second elution step was carried out in 75 µl of 50 mM HEPES, pH 8.2. After digestion and elution, both eluates were combined (final volume 200 µl of 50 mM HEPES, pH 8.2). Residual beads were removed by centrifugation at 4˚C, 17,100 x g for 10 min and supernatant transferred to new PCR tubes, and subsequently dried down on a speedvac.

TPP sample TMT labeling and peptide desalting

Dried down peptides were resuspended in 50 µl of 30% ACN solution, vortexed, and left shaking at room temperature for 5 min. We then labeled 15 µl of the above resuspension (15 µg of peptides) with 60 µg of TMT11plex Isobaric Label Reagent (ThermoFisher Scientific) for 1 hour at room temperature. The reaction was quenched with 2 µl of 5% hydroxylamine for 15 min

and channels pooled together prior to acidification to pH 3 with 10% TFA (to final concentration around 1-2%). Acidified peptides were briefly placed on a speedvac to remove residual ACN before desalting further using Sep-Pak tC18 polymer columns (Waters). Sep-Pak tC18 cartridges were equilibrated with sequential additions of 100% acetonitrile (ACN), 75% ACN with 0.5% acetic acid (AA), 50% ACN with 0.5% AA, and 0.1% TFA. Peptide samples were loaded onto the column and washed three times with 0.1% TFA and 0.5% AA. Peptide samples were eluted by sequential additions of 500 µl of 50% ACN with 0.5% AA, and 500 µl of 75% ACN with 0.5% AA. Eluates for each sample were separated into an aliquot of 10 µg (for single-shot injection analysis to assess labeling efficiency), and 200 µg for downstreams fractionation.

Offline peptide fractionation for TPP samples

Peptides were fractionated using a pentafluorophenyl (PFP) reverse-phase fractionation[103], using a Waters XSelect HSS PFP 2.5 µm 2.1 x 150 mm column and HPLC and fraction collector. Approximately 200 µg of TMT-labeled peptides were resuspended in 100 µl of buffer A (3% acetonitrile in 0.1% TFA) and separated with buffer B (95% acetonitrile in 0.1% TFA) along a 90 minute gradient (0–3 min: 3–10%, 3-63 min: 10–32%, 63–73 min: 32–55%, 73-74 min: 55%-95%, 74–79 min: 95%, 79–80 min: 95%-3%, 80-90 min: 3%) at 300 nl min^{-1}. There were 48 fractions collected horizontally between 12 minutes and 60 minutes which were combined vertically to 12 fractions. Fractions were dried by vacuum centrifugation and stored at -20˚C until LC-MS analysis. Fractions were solubilized in 5% acetonitrile, 5% formic acid, and 500 ng of each fraction was analyzed by LC-MS/MS.

Mass spectrometry analysis of TPP samples

Lyophilized TMT-labeled peptides were resuspended in 5% ACN, 5% formic acid and subjected to liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Peptide samples were loaded onto a 100 μm ID x 3 cm precolumn packed with Reprosil C18 1.9 μm, 120Å particles (Dr. Maisch). Peptides were eluted over a 100 μm ID x 30 cm analytical column packed with the same material housed in a column heater set to 50°C and separated by gradient elution of 10 to 32% ACN in 0.15% FA over 60 min at 400 nl/min delivered by an Easy1200 nLC system (Thermo Scientific). Peptides were online analyzed on an Orbitrap Eclipse mass spectrometer (Thermo Scientific). Mass spectra were collected using a data dependent SPS-MS3 acquisition method[104]. For each cycle a full MS scan (400-1400 m/z, resolution 120,000, AGC target 4e5, max injection time 50 ms, charge states 2-6, dynamic exclusion 30s), followed by MS/MS scans on the most intense precursor peaks using CID fragmentation and acquisition in the linear ion trap (isolation width of 0.7 Da, normalized collision energy 36, rapid, AGC target 1e4, max injection time 50 ms), each followed by an MS/MS/MS scan from coisolating and co-fragmenting the 10 most intense MS/MS fragments, using HCD fragmentation and acquisition in the Orbitrap for reporter ion quantification (isolation width of 0.7 Da, normalized collision energy 55, resolution of 50,000, 1e5 AGC, max injection time 120 ms).

Mass spectrometry analysis of dynamic SILAC samples

Lyophilized peptides from the dynamic SILAC fractionated samples were subjected to LC-MS/MS with a Easy1000 nLC system (Thermo Scientific) coupled to a QExactive hybrid mass spectrometer (Thermo Scientific), using columns of the same composition as above. Peptides were separated over a 94 minute gradient of 80% ACN, 0.125% Formic acid (Solvent B). The LC method started at 4% Solvent B for two minutes then continued with a gradient of

Solvent B ranging from 8-25% for RPB1 to 18-38% for RPB5. For a 120 minute MS run, the MS duty cycle consisted of an MS1 followed by 20 data dependent MS/MS scans of top most abundant precursors (dynamic exclusion set to 60 seconds). MS1 scans were performed at 70,000 resolution, 3e6 AGC, and max injection time of 100 ms over a 300-1,500 m/z range. Most abundant precursors were isolated with a 2 m/z isolation window, fragmented at 26 NCE with HCD, and subjected to MS/MS in the Orbitrap at 17,500 resolution, 5e4 AGC, and 54 ms maximum injection time.

Mass spectrometry data analysis of TPP experiment

Raw files were converted to the mzXML format and MS/MS spectra were searched against a target/decoy protein sequence database using Comet[59,105] (version 2019.01) to make peptide-spectral matches. The database consisted of the Uniprot human canonical proteome (Proteome ID: UP000005640, download date 8/11/2020) with GFP, NSP5, and NSP5 C145A amino acid sequences appended.

Mass tolerance search parameters were adjusted to acquisition instruments following recommendations by Comet source website, i.e. 20 ppm precursor mass tolerance (Orbitrap), 0.02 Da fragment tolerance for MS/MS acquired on an orbitrap mass analyzer and 1.0005Da tolerance with 0.4 Da offset for MS/MS acquired on a linear ion trap mass analyzer. LysC was selected as the digestive enzyme with a maximum of 2 missed cleavages, constant carbamidomethylation modification of cysteines (+57.0215 Da) and variable modifications of methionine oxidation (+15.9949 Da). Variable modifications were also used to search for the incorporation of non canonical amino acids. TMT-labeled samples were searched with constant modification (+229.1629 Da) on lysines and peptide N-termini. Search results were filtered with Percolator[106] to 1% false discovery rate at the PSM level. Peptide abundance was determined

using in-house quantification software to extract MS1 intensity or TMT reporter ion intensities. Protein groups were assembled using ProteinProphet[107]. TMT reporter ion intensities were corrected for isotopic interference.

Selection of peptides for melting curve fitting

For fitting melting curves, we only considered PSMs with reporter ion intensity greater than 0 in channel 126 and channel 127N (SDS and NP40 channels) and where at least 5 of the top 10 most intense fragment ions in the MS2 belonged to the assigned peptide. After filtering, TMT reporter ion intensities were transformed into relative fold-changes by normalizing each channel intensity to the channel containing the 30˚C control (channel 126). PSMs were consolidated into unique peptides by taking the median fold-change across all PSMs and charge states for each unique peptide.

Peptide level melting curve normalization

To account for differences in the amount of material labeled in each channel, we applied a normalization approach implemented in the Miettinen et al. study[108]. Briefly, we selected a set of proteins with relative fold-changes between 0.5 and 2 across the entire temperature range, and with a minimum of 5 unique peptides. We defined this set of proteins as our “non-melting” proteins and used this protein set for sample loading normalization across the entire dataset. Specifically, relative fold-changes for each protein were calculated by taking the median fold-change from all peptides assigned to that protein. We then calculate correction factors so that the median relative fold-change for each replicate and each temperature were equal to 1. These correction factors were then applied across the entire dataset.

Fitting melting curves and identifying proteins with significant changes in thermal stability

For statistical analysis, peptides were additionally filtered for being observed in both replicates across all three experimental conditions. Protein melting curves were calculated by taking the median fold-change across the temperature range. Proteins were required to be quantified using at least two unique peptides. To identify changes in protein thermal stability, we fit melting curves and applied non-parametric analysis of response curves (NPARC)[109] in a pairwise manner (i.e. GFP vs NSP5 (WT); GFP vs NSP5 (C145A); NSP5 (WT) vs NSP5 (C145A). The F-distribution was estimated empirically for each comparison to calculate p-values, which were corrected for multiple comparisons using the Benjamini-Hochberg method. Principal Component Analysis (PCA) was conducted in R using the ggbiplot R package. Lastly, we also used NPARC to fit melting curves to individual peptides, and extracted melting temperature and area under the melting curve (AUC).

Database searching, protein turnover calculation, statistical testing, and bioinformatic analysis of dynamic SILAC samples

For protein turnover, MS raw files from NSP5, NSP5 C145A, and GFP overexpression, fractionated HEK293T samples were database searched using MaxQuant[48] (v.1.6.14.0) for peptide identification and quantification. The data was searched against the Uniprot human canonical proteome (Proteome ID: UP000005640, download date 8/11/2020) with GFP, NSP5, and NSP5 C145A amino acid sequences appended. The following parameters were used in the database search: LysC enzyme specificity (cleavage C-terminal to lysine except when followed by proline) with maximum two missed cleavages, 20 ppm MS1 and MS2 mass tolerance, fixed carbamidomethyl modification on cysteines, and variable modifications for oxidation on

methionine, $^{13}C_6{}^{15}N_2$-Lysine-8 on lysine, and acetylation at protein N-termini. Peptide spectral matches and proteins were filtered globally at a 1% FDR.

Heavy and light peptide features for the fractionated samples were extracted from the evidence.txt file. Following PSM filtering requiring both heavy and light features, PFP fractions were median normalized by PSM total intensities (heavy + light), and heavy and light intensities were corrected by the fraction's total intensity normalization factor in order to preserve the observed heavy/light intensity ratio. PSM quantifications across fractions were consolidated to the peptide level for each sample using a weighted average heavy/light ratio to calculate a relative protein turnover ratio ($R_{TO}=\log_2(\Sigma(\text{heavy})/\Sigma(\text{light}))$[71]. Peptides were required to be observed in at least two of the four replicate conditions for generation of protein replicate correlations and identifications across samples (Appendix C Supplementary Figures 1a and 2).

For statistical analysis, peptides were additionally filtered for being observed in all three experimental conditions in at least two replicates. Protein R_{TO} was calculated by taking the median of its peptides R_{TO} per sample with at least two peptides per protein. For ANOVA and limma analysis, proteins needed to be observed in at least three replicates across all three conditions (n=3-4). Limma was used to compare each combination of conditions with n=3-4 replicates per condition. ANOVA and limma[110] (R package limma) statistical tests were conducted using R (v.3.6.1) in the RStudio environment (v.1.4.1103) to calculate p-values, which were corrected for multiple comparisons using the Benjamini-Hochberg method. Principal Component Analysis (PCA) was conducted in R using the ggbiplot R package, which required the additional filter that proteins be found in all replicates. Perseus[111] was used to generate dendrograms and perform hierarchical clustering for the ANOVA significant hits Z-scored R_{TO} values. Enriched gene ontology terms for the ANOVA significant protein clusters were

determined using all ANOVA tested proteins as background. Protein structures for the ribosome and U4/U6 spliceosome were downloaded from the Protein Data Bank and visualized using open-source PyMOL.

4.4 Results

4.4.1 Dynamic SILAC assay to explore NSP5 protease activity impact on the protein turnover across the HEK293T proteome

Herein, we explore using protein thermal stability and protein turnover assays to identify proteome-wide changes due to NSP5 protease activity and to identify NSP5 protease substrates in HEK293T. First, we applied a dynamic SILAC labeling approach to HEK293T cells overexpressing GFP, catalytically inactive NSP5 C145A, or wildtype NSP5 protease (each N=4). In our implementation of dynamic SILAC, we monitor the steady state incorporation of a pulsed isotopically heavy lysine amino acid into newly-synthesized proteins to proxy protein turnover. Upon harvest, pre-existing (light Lys0 containing) and newly-synthesized (heavy Lys8 containing) proteins are digested into peptides and analyzed by MS/MS (Figure 4.1a). We calculated a protein turnover proxy, or R_{TO}, at the peptide-level via the $\log_2$(heavy/light) SILAC peptide MS intensities. Peptide-level R_{TO} readouts can be surmised to a protein-level R_{TO} by taking the median R_{TO} of its constituent peptides.

To sensitively assay the functional role of NSP5 protease activity, we overexpressed GFP and the NSP5 C145A protein variant as controls. GFP overexpression controls for overexpression toxicity, while the catalytically inactive NSP5 C145A variant controls for NSP5-specific protein interactions unrelated to protease activity. When compared to these controls, wildtype NSP5 overexpression should uniquely proxy protein turnover changes driven by NSP5's protease activity. If a change in protein turnover (R_{TO}) tracks to the wildtype NSP5

condition, one could surmise that NSP5 protease is either directly cleaving the protein substrate altering its turnover or is indirectly acting to modulate the protein's turnover.

Alternatively, we can confidently identify NSP5 protease substrates in the HEK293T proteome using a different perspective. Since the dynamic SILAC protein turnover assay occurs at the protein-level, peptide-level R_{TO} readouts should reflect the protein turnover of all protein molecules that contain that unique peptide. Upon a protein cleavage, resultant protein products should contain peptides specific to each cleaved product reflecting their respective turnovers. We could identify a protein cleavage event when the cleaved protein products have different turnovers. This would be reflected by each product's peptide R_{TO} readouts tracking to their respective product but deviating between products at the location of the protein cleavage. Together, we leverage both our protein-level and peptide-level perspectives to identify global proteome changes due to NSP5 protease activity and to identify high confidence NSP5 host protein substrates proteome-wide.

We observed that despite all constructs being under the same promoter, they varied in their R_{TO}. NSP5 demonstrated a substantially slower R_{TO} than GFP and NSP5 C145A (Figure 4.1b). Thus, given the same promoter and likely similar synthesis rates, one could attribute the R_{TO} differences among constructs to be related to differences in degradation. Of note, all overexpressed proteins (NSP5, NSP5 C145A, and GFP) demonstrated extremely fast turnover compared to the HEK293T proteome (all > 98^{th} percentile). Not surprisingly, this suggests the promoter system likely enables exceedingly faster synthesis rates compared to what is naturally observed across the HEK293T during the 39 hour overexpression.

Across the proteome, we captured ~4,000 unique human proteins across all replicates and all overexpression conditions (Appendix C Supplementary Figure 4.1a). Also, the protein-level

replicate correlations of R_{TO} were highly reproducible, with Pearson Correlations of R=0.89-0.93 for all replicates and protein expression conditions (Appendix C Supplementary Figure 4.2). Next, we addressed the similarity of the proteome's protein turnover responses across the different overexpressed proteins by performing a principal component analysis (PCA) for GFP, NSP5 C145A, and wildtype NSP5 replicates (Figure 4.1c). The replicates for the same overexpressed protein clustered closely together, and replicates of different overexpressed proteins clearly separated. Particularly, the variance in principal component 1 (PC1) which explained 22% of the total variance clearly separated NSP5 wildtype replicates from GFP and NSP5 C145A replicates, highlighting that protease activity likely contributes unique differences in protein turnover across the proteome.

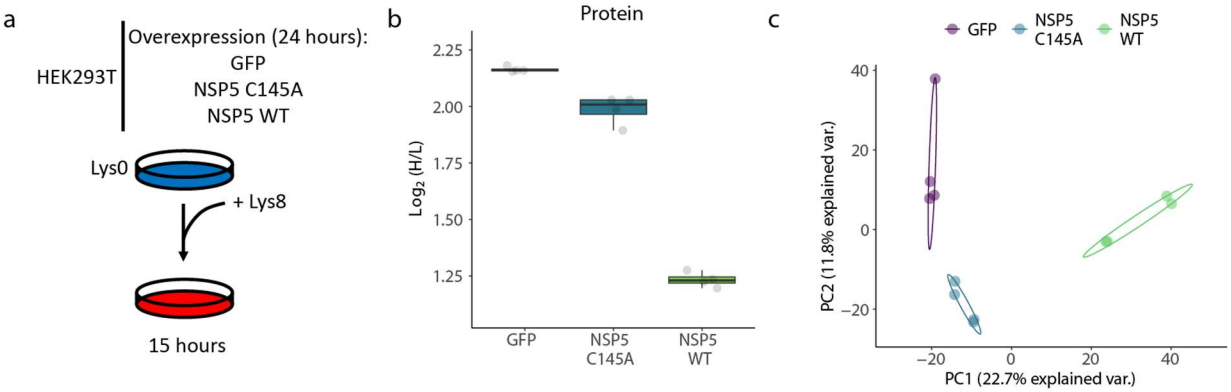


Figure 4.1: Dynamic SILAC to measure HEK293T protein turnover. a) HEK293T cells overexpressing GFP, NSP5 C145A, and NSP5 wildtype proteins for 24 hours were pulsed with media containing heavy lysine (Lys8) to incorporate into newly-synthesized protein for 15 hours. Harvest cell's proteins were digested into peptides and analyzed by MS/MS. Protein turnover was calculated at the peptide-level as $R_{TO}=\log_2$(heavy/light) peptide intensities (new/pre-existing). **b)** R_{TO} protein-level readouts across replicates represented as a boxplot and with replicate R_{TO} values as jittered points. **c)** Principal Component Analysis (PCA) was performed for replicates (points) across the different protein expression conditions (purple:GFP ; blue:NSP5 C145A ; green:NSP5 WT).

4.4.2 NSP5 PROTEASE ACTIVITY MODULATES HOST PROTEIN TURNOVER

Using the dynamic SILAC data, we explored whether NSP5 activity modulated the protein turnover globally in HEK293T cells. No method to date has been able to explicitly explore the functional impacts of the critical NSP5 protease on the host proteome. Using a

Limma statistical analysis, we conducted pairwise comparisons across all our HEK293T overexpressing strains (Figure 4.2a, Appendix C Supplementary Figure 4.3). First, we observed very few proteins with significantly altered protein turnover when comparing our expression control NSP5 to our catalytically inactive NSP5 C145A, suggesting the NSP5 interaction events likely have limited impact on protein turnover. Alternatively, we observed greater than 100 proteins with altered protein turnover when comparing NSP5 wildtype and GFP conditions, suggesting NSP5 activity plays a role in modulating protein turnover across the proteome. When we compared protease active NSP5 wildtype to inactive NSP5 C145A conditions, we observed less significantly slower turnover proteins compared to NSP5 wildtype vs. GFP. However, the prominent number of significantly faster turnover proteins suggests that NSP5 protease activity likely accelerates protein turnover across the proteome.

We postulate that proteins with faster protein turnover in the presence of active NSP5 could be attributed to the protein being a NSP5 protease substrate. Our rationale behind this hypothesis is that a NSP5 protease cleavage event likely will destabilize the resultant protein cleaved products, rendering them non-functional. To compensate for the destabilized protein fragments, the cell likely accelerates the degradation of the cleaved protein products inducing a faster turnover compared to the full length protein.

Next we set out to obtain a holistic view of the turnover changes across all the overexpression conditions. To this end, we performed an analysis of variance (ANOVA) to prioritize proteins with significant protein turnover changes across all overexpression conditions. Proteins with significantly altered protein turnover (Benjamini-Hochberg adjusted p-value $<$ 0.05) were visualized by plotting ΔR_{TO} of (NSP5 C145A - GFP) against ΔR_{TO} of (NSP5 wildtype - GFP) (Figure 4.2b). In alignment with our Limma analysis, very few proteins with

altered ΔR_{TO} lie on the diagonal (linear line with slope=1) and deviate from the origin, suggesting that few proteins have altered R_{TO} from the GFP condition. Most of the differences in ΔR_{TO} are spread across the x-axis and not the y-axis, suggesting that most of the proteins with significantly altered R_{TO} arise from NSP5 protease activity.

From hierarchical clustering of the ANOVA significant results, we found two prominent clusters with changes in R_{TO} (Cluster 2 : slower R_{TO}; Cluster 4 : faster R_{TO}) specific to the active wiltype NSP5 protease (Appendix C Supplementary Figure 4.4a). The slower R_{TO} proteins in Cluster 2 were enriched in the gene ontology (GO) term: aminoacyl-tRNA ligase activity, while the faster R_{TO} proteins in Cluster 4 in were enriched in GO terms, such as apoptosis, vesicle, cell death, and cell platelet degranulation (Appendix C Supplementary Figure 4.4b). The GO terms enriched for the faster R_{TO} proteins of Cluster 4 GO enrichments in terms related to cell death implicate a proteome response to compensate for the likely toxic, proteostasis stress induced by NSP5 protease activity.

Given the known putative NSP5 protease substrate motif (LQ|[AS][92,98,99]), we highlighted whether or not the ANOVA significant proteins contained at least one NSP5 motif across its sequence. In theory, NSP5 motif-containing proteins should be more likely to be protease substrates of NSP5. When mapped to our ANOVA significant calls, we observed proteins with the LQ|[AS] motif predominantly demonstrated a faster R_{TO} for NSP5 wildtype to GFP and little/no change in R_{TO} for NSP5 C145A to GFP (Figure 4.2b). This supports our hypothesis that NSP5 protease cleavage of protein substrates could accelerate their protein turnover, likely due to destabilization and increased degradation of its protein cleaved products. When we categorized the ANOVA significant proteins as either containing or not containing the putative NSP5 motif, we observed that proteins with the motif presented a faster R_{TO} exclusively when

the active NSP5 protease was expressed (Figure 4.2c-e). When comparing proteins with significantly altered R_{TO} by ANOVA to proteins without a changing R_{TO}, we observed an increased propensity to have one or more NSP5 motifs across the length of the protein (Figure 4.2f). Collectively, the protein turnover at the protein-level revealed that NSP5 protease activity can modulate proteome-wide protein turnover (R_{TO}) changes. Also, many faster R_{TO} proteins during NSP5 overexpression contain the putative NSP5 motif, potentially implicating them as NSP5 protease substrates.

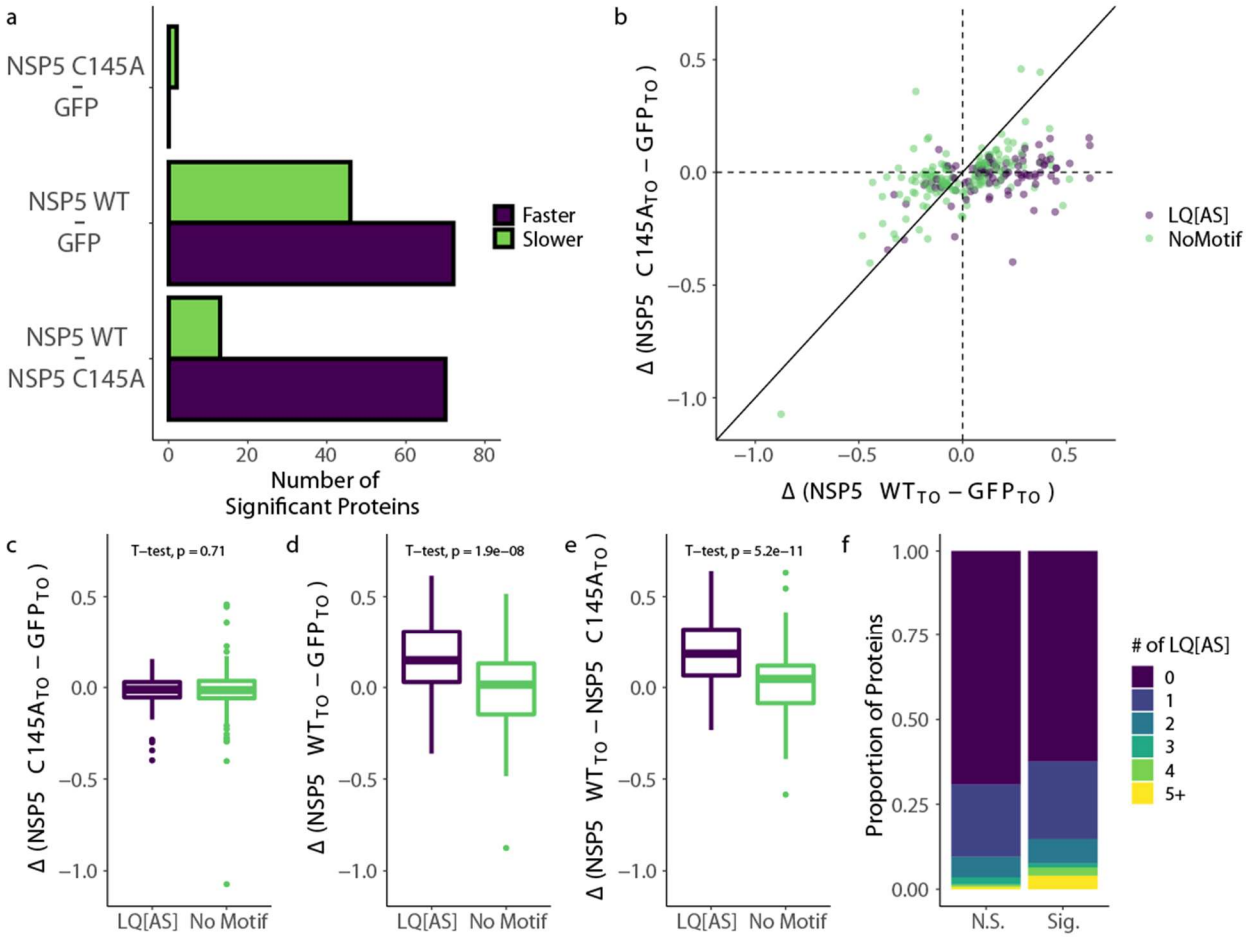


Figure 4.2: NSP5 protease activity modulated changes in protein R_{TO} and the faster R_{TO} proteins associated with the NSP5 motif a) Limma-based statistical analysis for pairwise comparisons of overexpression conditions (NSP5 C145A vs. GFP; NSP5 WT vs. GFP; NSP5 WT vs. NSP5 C145A) for N=4 replicates. Bar plot of significantly faster (purple) and slower (green) R_{TO} proteins for above comparisons (Benjamini-Hochberg adjusted p-values <0.05). **b)** ANOVA significant calls (Benjamini-Hochberg adjusted p-values <0.05) across all three conditions presented in a scatter plot. Each point defines the ΔR_{TO} based on the difference between the median replicate R_{TO} per condition (x-axis : ΔR_{TO} (wildtype NSP5 - GFP) ; y-axis : ΔR_{TO} (NSP5 C145A - GFP)). Color of points designate whether it contains (purple) or does not contain (green) at least one LQ|[AS] motif in its protein sequence. Dotted lines designate no difference in ΔR_{TO} across both axes. Solid line on the diagonal has slope of 0 with no intercept to depict ΔR_{TO} driven by GFP. **c)** Boxplot of ANOVA significant protein calls partitioned by containing at least one LQ|[AS] motif in its protein sequence (purple) or not (green) and the ΔR_{TO} between NSP5 C145A and GFP (based on the difference between the median replicate R_{TO} of each condition). Student's t-test conducted with presented p-values. All boxplots are for n=4 biological replicates (line = median, box = interquartile range (IQR), and whiskers = 1.5*IQR from box ends). **d)** Same as in (**c**) for the ΔR_{TO} between NSP5 wildtype and GFP. **e)** Same as in (**c**) for the ΔR_{TO} between NSP5 wildtype and NSP5 C145A. **f)** Barplot of the proportion of proteins colored by the number of instances (0-5) the LQ|[AS] motif appears in the protein's sequence, categorized by the ANOVA significance call or not across the proteome analysis.

4.4.3 Crude Thermal Proteome Profiling to explore the effects of NSP5 activity on protein thermal stability across the HEK293T proteome

We next turned to looking at the effects of NSP5 catalytic activity on protein thermal stability. To do this, we grew two biological replicates of HEK293T in identical conditions as turnover (without isotopically heavy lysine) and applied crude thermal proteome profiling. Briefly, cells were lysed in non-denaturing buffer and equal volumes of cell extracts were distributed across 11 PCR tubes, ten for temperature treatment and one for SDS total proteome extraction. After temperature treatment, soluble proteins were extracted by first incubated lysates with non-denaturing detergent and benzonase, followed by aggregate removal by centrifugation. The non-denatured protein fraction from each channel were removed, reduced, alkylated, LysC digested, and labeled with TMT11plex. Samples were fractionated offline and data were acquired using synchronous precursor selection with MS3 quantification on an Orbitrap Eclipse.

Across the entire sample set, we quantified 386,031 PSMs for 55,228 unique peptides that map to 9,232 proteins, with 265,082 PSMs, 21,886 unique peptides, and 3637 proteins quantified in all samples and all replicates with a minimum of 2 unique peptides. The corresponding melting curves for these overlapping peptides and proteins were highly reproducible and spanned a wide range of apparent thermal stability (Supplemental Figure 4.5), with 2570 proteins showing partial or full precipitation (i.e. $T_m \leq 67^{\circ}C$) in at least one condition within the temperature range tested here and 471 showing no observable precipitation (i.e. $T_m > 67^{\circ}C$) in any sample.

We first assessed the thermal stability of overexpressed proteins in our sample. Both NSP5 (C145A) and NSP5 wildtype fully precipitated within the temperature range, with melting temperatures around 51.63˚C and 49.73 ˚C, respectively (Figure 4.3a). Interestingly, the

increased stability of catalytically-dead NSP5 lines up with prior observations that point mutants decreasing enzymatic activity result in a concomitant increase in stability[112]. Conversely, despite performing these experiments in the context of the cellular milieu, which can cause some shift in thermal stability, GFP did not precipitate within the temperature range. Soluble and folded GFP is known to be incredibly stable at high temperature (T_m > 67˚C)[113], thus suggesting that GFP is resistant to precipitation at these temperatures even in a complex lysate background. Taken together, these data highlight a concordance between thermal stability as measured by TPP with what is already known or can be inferred about the overexpressed proteins.

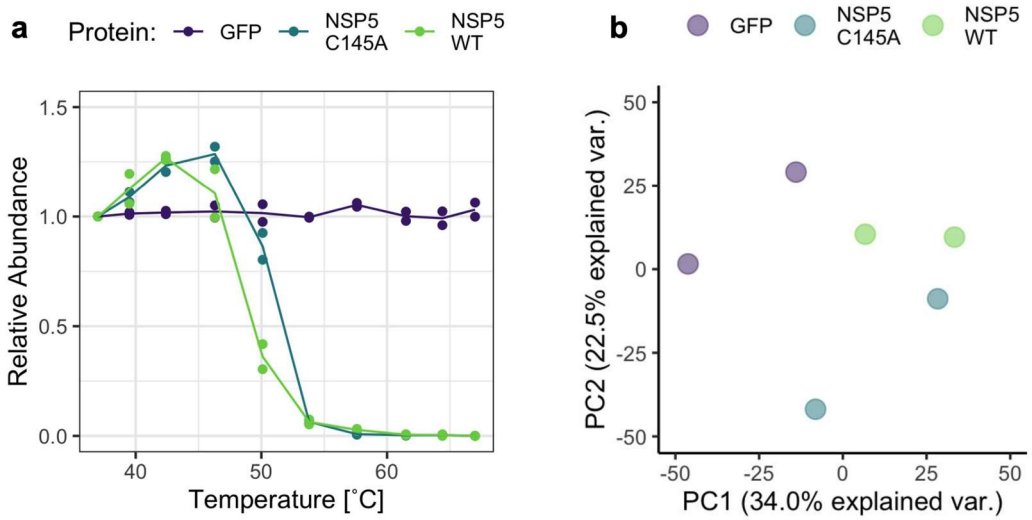


Figure 4.3: Crude Thermal Proteome Profiling to measure NSP5-dependent changes in protein stability. a) Thermal stability of the three overexpressed proteins GFP, catalytically-dead NSP5 (NSP5 C145A), and wildtype NSP5 (NSP5 WT). **b)** Principal Component Analysis (PCA) was performed for replicates (points) using all available melting curves across the proteome quantified in the different protein expression conditions (purple:GFP ; blue:NSP5 C145A ; green:NSP5 WT).

4.4.4 NSP5 PROTEASE ACTIVITY ALTERS HOST PROTEIN THERMAL STABILITY

Next, we determined the effect of NSP5 activity on proteome thermal stability. We did this by performing a principle component analysis, where we observed broad separation of samples that clustered based on which enzyme was overexpressed (Figure 4.3b). Then, to establish which proteins are driving these sample-specific changes, we compared changes in

protein-level melting curves across all three overexpression conditions using non-parametric analysis of response curves (NPARC)[109]. Specifically, we focused on the 2,570 proteins that show partial or full precipitation within the temperature window in at least one of the conditions, requiring at least 2 unique peptides per protein. We then applied NPARC across all pairwise combinations (GFP vs NSP5 (C145A); GFP vs NSP5 (WT); NSP5 (C145A) vs NSP5 (WT)).

In total, 139 unique proteins (~5.5%) showed a significant change in protein thermal stability in at least one comparison, 35 of which showed a change in two or more comparisons. As expected, we saw the most proteins with a significant change when comparing catalytically-dead NSP5 with wildtype NSP5 overexpressing cells (26 destabilized in wildtype NSP5 compared to 52 stabilized in wildtype NSP5) (Figure 4.4a), potentially highlighting a subset of proteins whose change in stability is a direct result of proteolytic cleavage or the result of proteolytic cleavage of a binding partner. Surprisingly, the second most number of changes were observed when comparing catalytically-dead NSP5 with GFP (57 destabilized by NSP5 C145A compared to 6 that were stabilized). In both of these comparison, the group of significantly altered proteins were enriched for the NSP5 motif (i.e. LQ|[AS]) (Figure 4.4b), suggesting that the change in protein stability may be associated with either wild type NSP5's proteolytic targeting of proteins or catalytically-dead NSP5's ability to dock and destabilize target proteins.

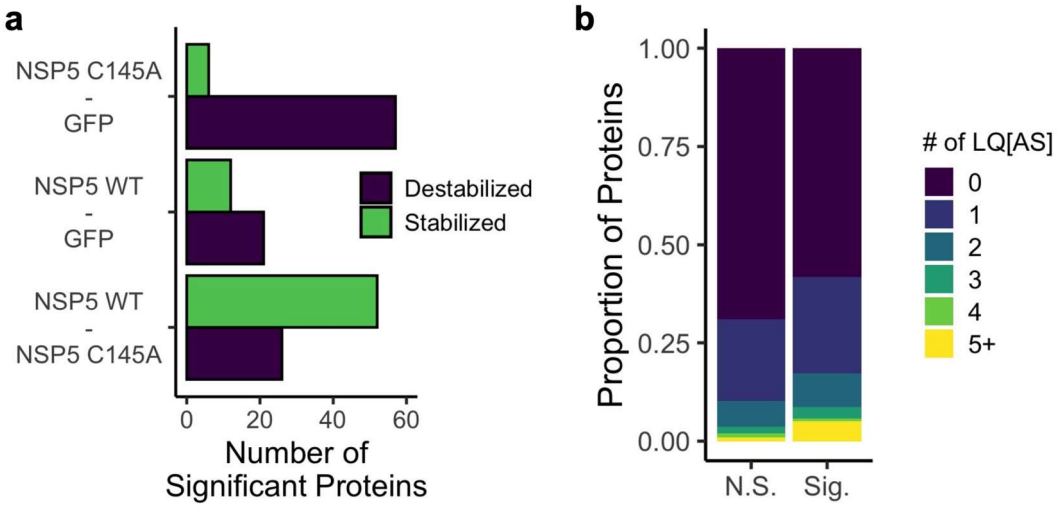


Figure 4.4: Proteins with altered stability due to NSP5 overexpression contain the NSP5 motif. a) Non-parametric analysis of response curves for comparing overexpression conditions (NSP5 C145A vs. GFP; NSP5 WT vs. GFP; NSP5 WT vs. NSP5 C145A) for N=2 replicates. Bar plot of significantly destabilized (purple) and stabilized (green) proteins for above comparisons (Benjamini-Hochberg adjusted p-values <0.05). **b)** Barplot of the proportion of proteins colored by the number of instances (0-5) the LQ|[AS] motif appears in the protein's sequence, categorized by whether that protein was significant in at least one comparison.

4.4.5 Protein-level changes in protein turnover and thermal stability overlap with known proteolytic substrates

We anticipate that proteins with altered protein turnover and thermal stability from the NSP5 protease could be NSP5 protease substrates. Thus, we expected to observe substantial overlap between known NSP5 substrates and proteins with altered turnover and thermal stability. In this section, we will explore the overlap from a protein-level perspective. From this perspective, we anticipate capturing a subset of NSP5 substrates where the cleavage event causes both protein cleaved products to assume a similar functional change to each other but different from the full-length protein. These substrates potentially could prioriterize a loss-of-function phenotype for the whole-length protein upon NSP5 cleavage. For example, accelerated protein turnover for the NSP5 cleaved products can indicate protein destabilization and increased degradation of the non-functional cleaved proteins. Fortunately, several recent papers have established a list of host proteins that are targeted for proteolysis by NSP5 in both *in vitro* and *in*

vivo conditions, some of which have an accompanying location of the cleavage site[98,99,114]. Thus, we leveraged the collective curated list for comparison.

First, we asked whether any of these known targets of NSP5 align with proteins that have altered protein thermal stability during active NSP5 overexpression. From a protein-level perspective, we quantified changes in stability for 113 proteins out of a possible 244 previously reported substrates of NSP5. Out of this subset, we found a significant change in thermal stability for 16 substrate proteins. In general, proteins that are known substrates of NSP5 were three times more likely to have a significant change in protein thermal stability in one of the conditions compared to the rest of the proteome, illustrating a correlation between known targets and changes in thermal stability (Figure 4.5a).

As an example, we highlight NSP5-dependent changes in thermal stability for the mitochondrial protein ornithine aminotransferase (OAT). OAT is a known substrate of NSP5 that has a noncanonical cleavage site near its N-terminus. When mapping peptide-level stability measurements and plotting the protein-level melting curves, we saw a consistent shift in thermal stability in the cells overexpressing wildtype NSP5 (Figure 4.5b-c). The change in thermal stability was similar across the different peptides quantified within OAT (Figure 4.6b), resulting in a reproducible stability change at the protein-level (Figure 4.6c). Interestingly, this change in thermal stability coincides with a potential change in subcellular localization of OAT caused by NSP5 cleavage. The N-terminal cleavage site identified separates OAT's mitochondrial import sequence from the rest of the protein, potentially causing mislocalization of OAT during SARS-CoV-2 infection. This mislocalization could underlie some of the observed changes in thermal stability, perhaps due to different cellular environments in the mitochondrial matrix compared to the cytosol.

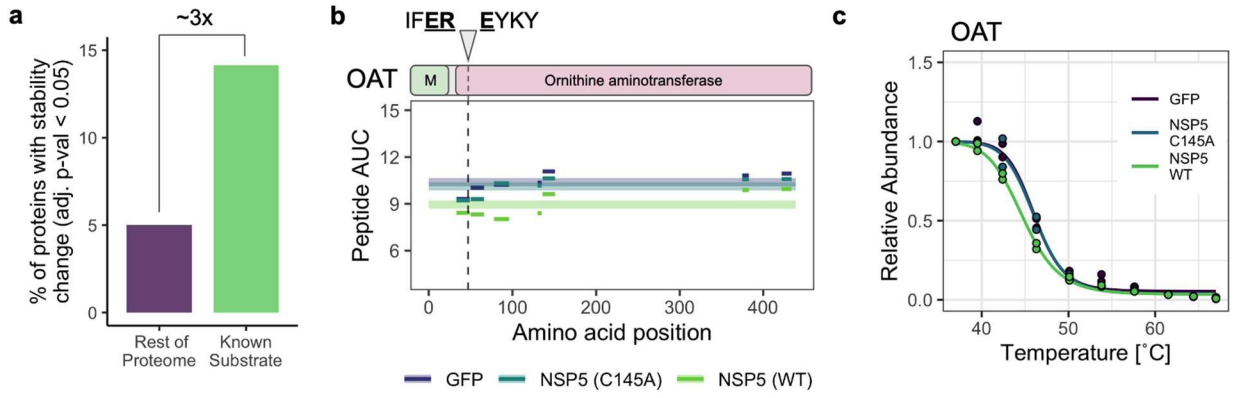


Figure 4.5: Protein-level thermal stability changes of known NSP5 substrates. (a) Barplot displaying the percent of the proteome and percent of known NSP5 substrates showing a significant change in thermal stability. **(b)** Peptide-level analysis of stability for peptides quantified for the known NSP5 substrates ornithine aminotransferase (OAT). Each color and bar represent median estimates of stability (n=2) for unique peptides quantified in different conditions (purple:GFP; blue:NSP5 C145A; green:NSP5 WT). The lighter bar in the background is the median stability value for peptides. The protein sequence track (grey) are shown relative to the NSP5 cleavage site and its motif annotated (underline text). **(c)** Protein-level melting curves for OAT across all three expression conditions.

We next explored the overlap of known NSP5 substrates with proteins that demonstrated significant changes in protein turnover during NSP5 protease overexpression. We quantified protein turnover of 157 proteins out of a possible 244 previously reported substrates of NSP5. We observed a significantly altered protein turnover for 45 of the known substrate proteins (or 29%). 84% of the 45 had a positive ΔR_{TO} (NSP5 WT - NSP5 C145A), which likely suggests a faster R_{TO} due to NSP5 protease activity. Importantly, we found a four-fold enrichment of observing a known substrate in proteins with significantly altered protein turnover (18.1%) compared to proteins with a non-significant designation (4.6%). This is in agreement with the notion presented previously that proteins with significantly altered protein turnover could derive its change from being a NSP5 protease substrate.

As observed with protein thermal stability, known NSP5 substrates can have profound changes in protein turnover that can be observed at the protein-level, which in turn should translate at the peptide-level. For instance, the NSP5 substrate EIF4G1, which has a known protease cleavage site near its N-terminus, demonstrated one of the most prominent faster protein

turnover changes during active NSP5 protease overexpression. The R_{TO} change was observed to the same extent for all its peptides across the length of the protein (Figure 4.6a), resulting in a reproducible turnover change at the protein-level (Figure 4.6b). EIF4G1 has been observed in multiple studies to bind viral RNA prominently after 24 hours of SARS-CoV-2 infection[115–117]. Blabeau *et al.* followed-up on many of the viral RNA human binding proteins (RBP), including EIF4G1, for their function role during SARS-CoV-2 infection[117]. In this experiment, siRNA-mediated knockdown of EIF4G1 and other RBPs was performed in A549-ACE2 cells, and viral titres and viral replication was then assessed following SARS-CoV-2 infection. For EIF4G1 knockdown cells, there was little/no observed effect on viral replication upon infection, however viral titres were over 200% greater than infected cells with non-targeting siRNA. Taken together with the increased R_{TO} of EIF4G1 from NSP5 overexpression, EIF4G1 could be targeted for NSP5 cleavage and its accelerated degradation during SARS-CoV-2 infection could serve as a mechanism to increase viral titres. Future experiments will need to be conducted to validate whether EIF4G1 is a NSP5 substrate during active SARS-CoV-2 infection and whether its protein cleavage plays a role in modulating viral titres.

We also observed increased protein turnover for the polypyrimidine tract binding protein (PTBP1) for all its peptides (Figure 4.6c) and at the protein-level (Figure 4.6d). PTBP1 was validated to be cleaved *in vitro* by NSP5 at position 152 for all 3 PTBP1 isoforms and position 352 for PTBP1 isoforms 2 and 3. As follow-up, Pablos *et al.* confirmed that PTBP1 cleavage also occurred during SARS-CoV-2 infection in Vero E6 cells, supported by a decrease in the abundance of full length PTBP1 48 hours post-infection[98]. Given PTBP1's NSP5 site between the RMR1 and RMR2 domain, the cleavage was suggested to extinguish PTBP1's N-terminal nuclear localization signal and as a result would cause a loss in PTBP1 transit to the nucleus.

Indeed, the NSP5 cleavage events on PTBP1 dramatically increased the ratio of PTBP1 cytoplasm/nucleus residency during SARS-CoV-2 infection, which was suggested as a possible viral strategy to repress host cell translation[98]. Given PTBP1's functional changes during infection, PTBP1's faster protein turnover could be explained by either protein destabilization by the cleavage event promoting degradation or an accelerated turnover as a consequence of its predominant cytoplasmic residency. Taken together, these protein turnover and thermal stability data highlight a correlation between changes in protein turnover and thermal stability, NSP5 activity, and known target substrates.

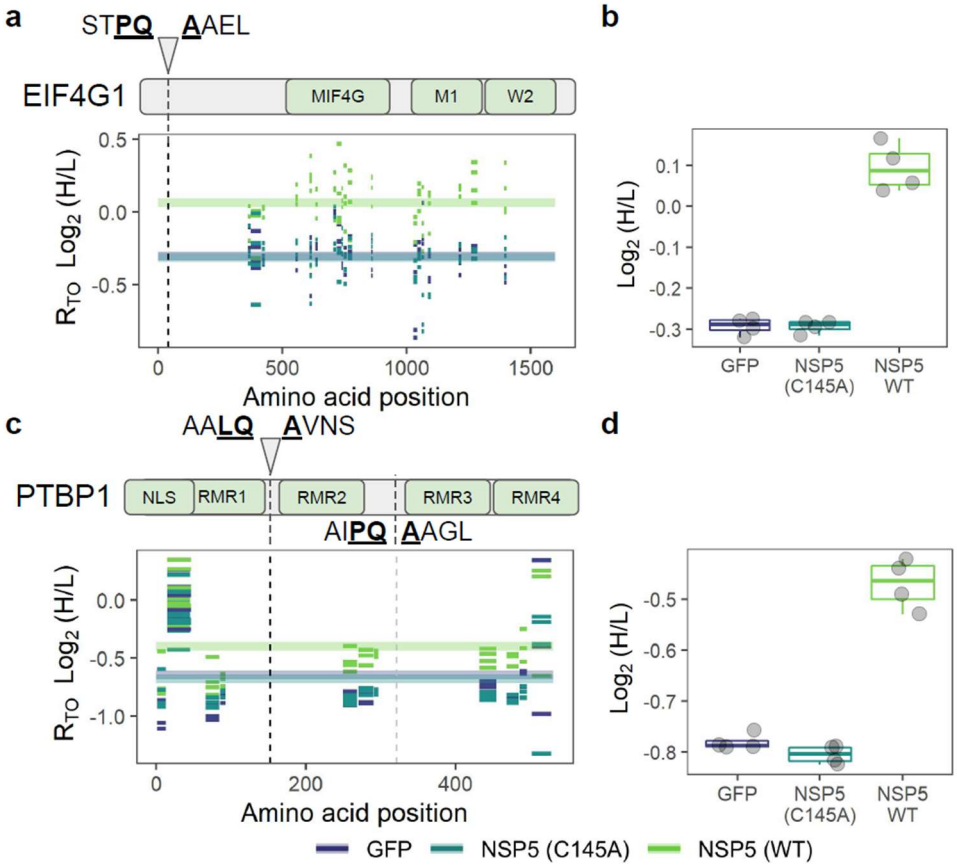


Figure 4.6: Protein-level faster R_{TO} in known NSP5 substrates. a) Peptide-level analysis of turnover for peptides quantified for the known NSP5 substrate EIF4G1.The vertical dashed lines represent the location of the known NSP5 cleavage sites. Each color and bar represent estimates of turnover for unique peptides quantified in different conditions (purple:GFP; blue:NSP5 C145A; green:NSP5 WT). The lighter bar in the background is the mean turnover value for all peptides of the protein. The protein sequence track (grey) and relevant protein domains (green) are shown relative to the NSP5 cleavage site and its motif annotated (underline text). **b)** Protein-level analysis of turnover for sample replicates of EIF4G1. Data is presented as a boxplot with the same sample color designations as in (**a**) with replicate protein turnover readouts jittered as points. **c)** Same as (**a**) for protein PTBP1. **d)** Same as (**b**) for PTBP1.

4.4.6 Integrating peptide-level readouts for turnover and thermal stability uncover known and potentially novel substrates of NSP5

Our analyses above indicate that NSP5-dependent cleavage of host proteins can alter thermal stability or turnover, observable by a global shift of peptides derived from the protein that is independent of the location of the cleavage site. We next asked whether we could infer the precise location of cleavage sites based on peptide-level differences in turnover or stability that differ based on which part of the protein they are derived from, i.e., from the N-terminal or C-terminal side of the cleavage site. To assess this possibility, we first focused on validated substrates where we identified enough peptides on either side of the proposed cleavage site that would allow us to pinpoint the cleavage site, potentially with high confidence.

To illustrate this phenomena, we first focus on two proteins in particular that showed cleavage site-dependent changes in stability and turnover: SEPTIN9 and MAGE-D2 (Figure 4.7). Both of these proteins are high-confidence substrates of NSP5 in the HEK293T proteome[98] and contain only a single cleavage site, which is the common LQS motif. We assessed the cleavage site dependency of these proteins by mapping peptides and their corresponding turnover and stability values back to each protein's primary amino acid sequence. To our excitement, both of these proteins contain a shift in peptide-level behaviors that coincide with both the location of the cleavage sites and the catalytic activity of NSP5. For example, in SEPTIN9, peptides derived from the N-terminal side of the cleavage site at position 221 are more stable and have a slower turnover, specifically in the cells overexpressing wildtype NSP5, compared to the peptides derived from the C-terminal side of the cleavage site. We see a similar trend in MAGE-D2,

where peptides derived from the N-terminal side of the cleavage site at 264 are more stable and have a slower turnover compared to the peptides on the C-terminal side of the cleavage site.

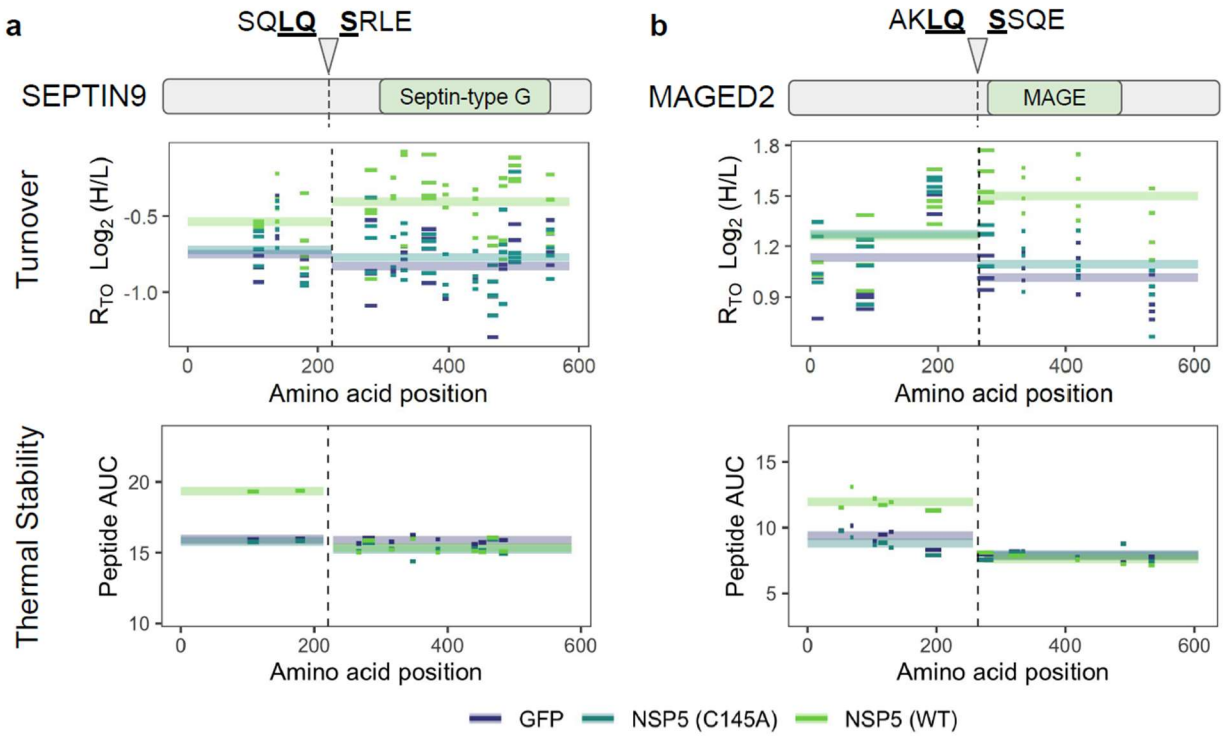


Figure 4.7: Peptide-level readouts for turnover and stability locate the specific cleavage sites for known NSP5 substrates. (**a and b)** Peptide-level analysis of turnover and stability for peptides quantified for the known NSP5 substrates **(a)** SEPTIN9 and **(b)** MAGE-D2. The vertical dashed lines represent the location of the known NSP5 cleavage sites. Plots in the first row are for turnover. Plots in the second row are for stability. Each color and bar represent estimates of stability or turnover for unique peptides quantified in different conditions (purple:GFP; blue:NSP5 C145A; green:NSP5 WT). The lighter bar in the background is the median turnover or stability value for peptides falling on either side of the known cut site. The protein sequence track (grey) and relevant protein domains (green) are shown relative to the NSP5 cleavage site and its motif annotated (underline text).

A key requirement of this cleavage site-specific behavior is that at least one of the products of proteolysis has to be distinctly different in its turnover or thermal stability compared to the original full-length protein. As a result, both dynamic SILAC and TPP offer unique perspectives and opportunities to capture these cleavage events, given that both thermal stability and turnover, while intrinsically related to protein structure and fold, will inherently be orthogonal and sensitive to different subsets of cleaved proteins.

For instance, we observed a number of known NSP5 substrates that demonstrated altered protein turnover due to NSP5 with minimal observed changes in protein thermal stability. Similarly to SEPTIN9 and MAGE-D2, we observe a breakpoint in the peptide-level R_{TO} values that coincide with the known cleavage site at position 452 only in the active NSP5 overexpression condition. The C-terminus of the protein demonstrated an accelerated protein turnover, which is likely attributed to an altered function of that portion of the protein to the full length protein and the N-terminal cleavage product. Pablos *et al.*[98] further validated that NSP5 has moderate-high specificity for the EIF4G2 LQG substrate with an apparent $k_{cat}/K_m>64$. Of note, the absence of the cleavage event in the thermal stability data does not contradict the turnover phenotype, for both N-terminal and C-terminal cleavage products could share the same thermal stability as their full length protein. Also, we captured the known NSP5 cleavage event at position 34 in the SAICAR synthetase portion of the multi-functional protein ADE2 (PAICS). This was indicated by the increased peptide-level R_{TO} readouts N-terminal to position 34 during NSP5 activity. This PAICS NSP5 cleavage event has been observed and validated to occur during active SARS-CoV-2 infection in A549-Ace2 cells[99]. A siRNA knockdown of PAICS did not significantly reduce viral titres, however it did significantly reduce plaque-forming units 10-fold in a plaque assay. While it is unclear whether PAICS cleavage is beneficial for SARS-CoV-2 during infection, we can support the observation that the small peptide N-terminus released upon NSP5 cleavage is likely behaving functional different than its full length protein and its C-terminal cleavage product.

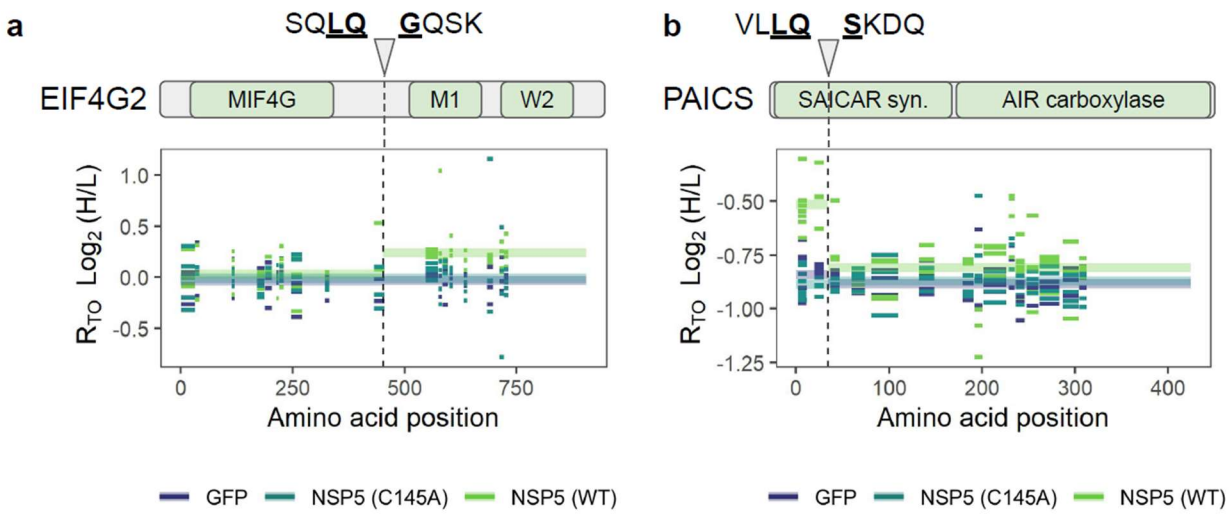


Figure 4.8: Known NSP5 substrates, EIF4G2 and PAICS, have altered protein turnover for their cleaved products. a and b) Peptide-level analysis of protein turnover for peptides quantified for the known NSP5 substrates **(a)** EIF4G2 and **(b)** Multiprotein ADE2 or PAICS. The vertical dashed lines represent the location of the known NSP5 cleavage sites. Each color and bar represent estimates of protein turnover for unique peptides quantified in different conditions (purple:GFP; blue:NSP5 C145A; green:NSP5 WT). The lighter bar in the background is the median turnover or stability value for peptides falling on either side of the known cut site.

Given that this cleavage site-dependent change in turnover and stability is most specific to cells overexpressing wildtype NSP5, we reasoned that novel substrates of NSP5 could be identified by looking for these region-specific and NSP5-dependent differences in peptide stability and turnover. To identify additional proteins with this behavior, we looked across the HEK293T proteome and mapped peptide turnover and stability values back to protein sequence. We then implemented a novel statistical workflow to identify proteins with suspected sequence-specific clustering of significant changes in peptide-level turnover or stability. Below, we discuss a few examples of proteins that emerge from this analysis that represent potentially-novel substrates of NSP5.

From our analysis of the TPP samples, two proteins emerge as potential substrates with possible biological relevance: the ribosomal subunit, RPL4 (Figure 4.9) and the U4/U6-U5 tri-snRNP pre-spliceosome associated protein, PRPF3 (Figure 4.10). For RPL4, there is one potential NSP5 cut site, the LQA motif at position 362. In our TPP assay, we capture sufficient

coverage of peptides across the entire length of RPL4 in all three overexpression conditions to detect a potential change. When mapping peptides and their stability values back to primary sequence, we observe a marked shift in thermal stability for peptides derived from the C-terminus that is specific to wildtype NSP5 (Figure 4.8a). Interestingly, while this C-terminal region is not captured in crystal structures of the ribosome, the region spanning the LQA motif is crystalized (Figure 4.8b). This region is at the very edge of the ribosome, solvent accessible, and therefore likely accessible wildtype NSP5. Interestingly, RPL4 overexpression has been shown to increase the efficiency of viral translation for viruses requiring frameshifts[118], such as SARS-CoV-2, suggesting the proteolytic cleavage of this part of RPL4 may have direct consequences on the efficiency of viral replication and virion production.

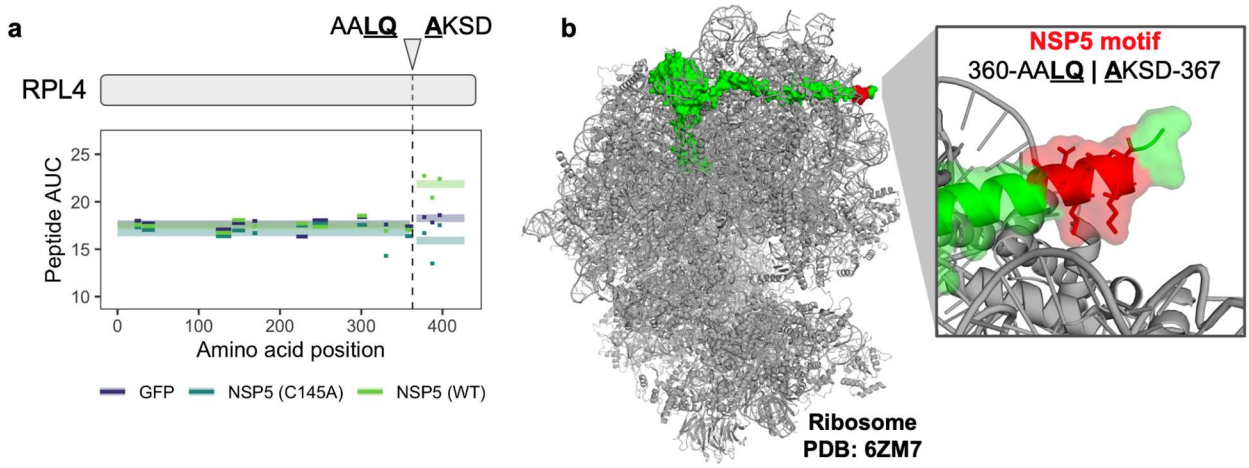


Figure 4.9: The ribosomal subunit RPL4 is a potential substrate of NSP5. a) Peptide-level analysis of stability for peptides derived from RPL4. The vertical dashed line represents the location of the proposed NSP5 cleavage site. Each color and bar represent median estimates of protein stability (n=2) for unique peptides quantified in different conditions (purple:GFP; blue:NSP5 C145A; green:NSP5 WT). The lighter bar in the background is the median turnover or stability value for peptides falling on either side of the proposed cut site. **b)** Placement of RPL4 within the human ribosome as determined by electron microscopy (PDB file 6ZM7). RPL4 is highlighted in green with a surface representation. Colored in gray is the rest of the ribosome. Highlighted in red is the proposed NSP5 motif.

The second protein whose cleavage may have relevant biological consequences is the spliceosomal-associated protein PRPF3 (Figure 4.10). SARS-CoV-2 has already been shown to suppress global mRNA splicing through targeting U1/U2 RNAs by NSP16[119]. Additionally,

knocking down PRPF3 has been shown to increase splicing defects in neuronal cells[120]. In our analysis, we find differences in peptide-level stability measurements between the N-terminal region of PRPF3, before the cut site at position 211. Interestingly, this separates the PRPF3 domain which binds the U4/U6-U5 tri-snRNP complex from the low complexity N-terminal region. While this region has not been implicated in viral replication, this cleavage event could alter the association of PRP3 with the U4/U6-U5 tri-snRNP complex, potentially disrupting the formation of functional spliceosomes.

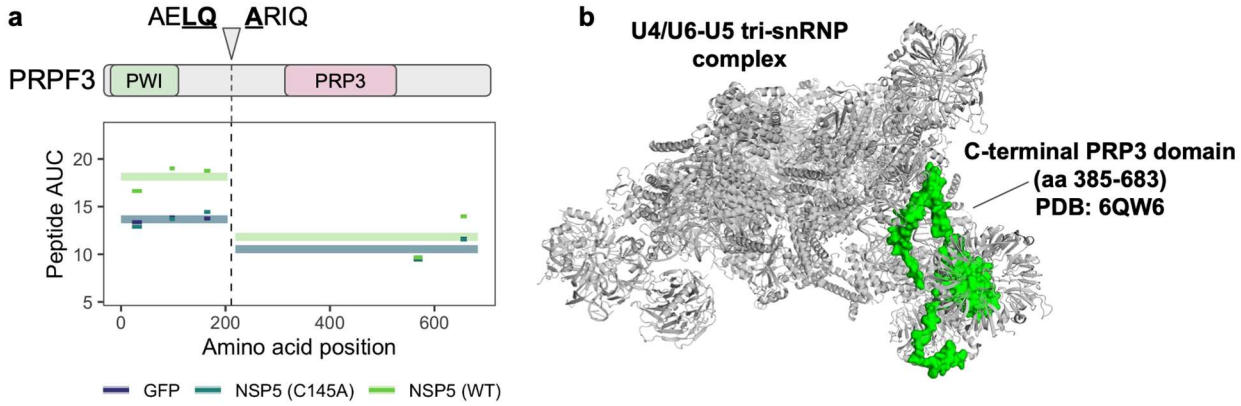


Figure 4.10: The pre-mRNA splicing protein PRPF3 is a potential substrate of NSP5. a) Peptide-level analysis of stability for peptides derived from PRPF3. The vertical dashed line represents the location of the proposed NSP5 cleavage site. Each color and bar represent median estimates of protein stability (n=2) for unique peptides quantified in different conditions (purple:GFP; blue:NSP5 C145A; green:NSP5 WT). The lighter bar in the background is the median turnover or stability value for peptides falling on either side of the proposed cut site. **b)** Placement of PRPF3 within the human U4/U6 spliceosome (PDB file 6ZM7). PRPF3 (truncased to positions 385-683) is highlighted in green with a surface representation. Colored in gray is the rest of the U4/U6-spliceosome. The proposed cleavage site is not shown but the model highlights what likely drives the stability of the C-terminal product.

4.5 Discussion

The SARS-CoV-2 main viral protease, NSP5, has been extensively studied for its essential role cleaving a large viral polyprotein into its many functional protein units, a process essential for viral propagation. Beyond targeting viral proteins, N-terminomics studies[98,99] have revealed that NSP5 can target over 100 human proteins as protease substrates, many relevant during infection. However, less is known about functional implications of NSP5 protease activity globally on the host proteome and whether host NSP5 substrates are functionally impacted by the protein cleavage event.

Thus, we coupled protein overexpression in HEK293T with protein turnover and protein thermal stability assays to identify NSP5 host substrates and proteome-wide host functional changes due to protease activity. By comparing HEK293T cells with overexpression of wildtype NSP5 to conditions with GFP and catalytically inactive NSP5 C145A overexpression, we captured over a 100 proteins with altered protein turnover and/or thermal stability attributed specifically to the NSP5's protease activity. Our boost in sensitivity to gain functional insight derives from combing the overexpression conditions, enabling the isolation of NSP5's protease activity contributions from its other functional roles. Thus, making this study, the first of its kind assaying functional changes induced by NSP5 protease activity at a proteome-wide scale.

For protein turnover, most proteins with an altered R_{TO} that contain the putative NSP5 LQ|[AS] motif were associated with faster turnover when NSP5 activity was present. Also, proteins with a significantly altered thermal stability in the presence of NSP5 were enriched for the LQ|[AS] motif. Importantly, proteins with altered protein turnover and thermal stability were enriched in known NSP5 substrates. Thus, we argue proteins that (1) contain the putative

LQ|[AS] motif and (2) have a faster R_{TO} or altered thermal stability due to NSP5 activity are likely candidate NSP5 protease substrates.

As a novel extension to these functional approaches, we were able to identify NSP5 host protease substrates by leveraging our protein turnover and protein thermal stability readouts at the peptide-level. To assign a protein cleavage, we can identify a breakpoint across the length of the protein where the peptide-level functional readouts differed between its two or more NSP5-cleaved products. Most high confident breakpoints for protein turnover or thermal stability aligned with known NSP5 substrates or were on proteins that contained a LQ|[AS] sequence in the vicinity of the breakpoint. Altered protein properties among a NSP5 substrate's cleaved products could implicate loss-of-function or non-canonical functions for the protein or its products, which could be important mechanistically during SARS-CoV-2 infection.

Despite our method's advantages, our protein turnover and thermal stability assays do have limitations for NSP5 protease substrate analysis. First, while many known NSP5 cleavages align with protein-level changes in turnover and thermal stability during NSP5 overexpression, we can not easily distinguish whether the altered property is due to a direct NSP5 cleavage of a substrate or to some other indirect functional origin. However, we highlight that presence of an NSP5 motif and certain directional changes in the functional properties could help prioritize between the two scenarios. Second, absence of an altered thermal stability or turnover does not negate that a cleavage event is occurring due to NSP5 protease. This could be due to biological and technical reasons. For instance, some proteins are not amenable to our thermal stability and protein turnover assays (technical). We rely on differences in the functional properties to identify a relevant NSP5-driven effect, however some proteins do not melt over the TPP temperature range or turnover too slowly or too quickly to capture a reliable turnover readout for. These

proteins will likely result in false negatives or missed proteins in our analysis. Also, some proteins that are cleaved and are amenable to the assays might not have a change in thermal stability or turnover, thus lacking functional support from our assays for the cleavage (biological). Third, making reliable peptide-level breakpoint calls is difficult without robust models that can leverage the fold-change of the effect against the noise in the peptide-level readouts. Thus, developments must be made for a statistical framework to reproducibly call breakpoints and give confidence in our NSP5 cleavage calls.

Often we contrast our methods to the field standard approach to study protease substrates, N-termnomics. We find that N-terminomics studies likely have better sensitivity for cataloging NSP5 substrates, while our assays have better functional insight into the impact of NSP5 cleavage on its substrates. Thus, we believe these different perspectives can complement each other. For example, we assert NSP5 substrates that (1) have a known neo-N-terminus peptide in previous studies and (2) altered protein turnover and/or thermal stability at the protein or peptide-level are more likely to be strong candidates for functional validation in the context of SARS-CoV-2 infection. This is supported by the overlap in N-terminomic studies and our study for the protein PTBP1, whose functional impact upon cleavage has been extensively characterized during infection. Despite potentially having less sensitivity to call substrates than N-terminomics at scale, we present two novel NSP5 substrates that have functional evidence of a NSP5 breakpoints, which warrant follow-up validation during infection. Therefore, our methods could pick up NSP5 substrates whose neo-n-termini peptide was unable to be detected by MS in other studies.

Collectively, we argue that a functional change to an NSP5 substrate could provide a better prioritization criteria for follow-up validation than extensively cataloging all NSP5

cleavage sites. However, combining such approaches could be an exceedingly powerful strategy for characterizing protease substrates for the future.

Chapter 5. DEVELOPING A PEPTIDE BARCODE METHOD TO ASSESS STABILITY OF THOUSANDS OF TPMT PROTEIN VARIANTS

5.1 ABSTRACT

With advances in genomic sequencing, we have been able to catalog millions of missense variants, however for a vast majority of variants, the functional impact on the protein and in the cell remains elusive. Deep mutational scanning (DMS) methods have accelerated our ability to annotate variant functions at scale, however experiments generally require tuning a protein function to a cellular phenotype and often do not encompass all the functions of a protein. Thus, we piloted a peptide barcoding approach which should enable parallel detection and quantification of thousands protein variants in a pooled format by mass spectrometry. Peptide barcodes, which are tagged to their representative protein variants, accurately reflected thiopurine methyltransferase (TPMT) wildtype and unstable variant protein abundances by MS. These MS phenotypes matched known phenotypes from a DMS abundance assay. Also, we successfully enriched peptide barcodes over the proteome background, suggesting robust detection of many protein variants in a single experiment. A majority of peptide barcodes when tagged to TPMT did not alter TPMT stability. We can fine tune our selection of peptide barcodes to develop an optimal inert peptide barcode library to ensure accurate representation of its tagged proteins. This work benchmarks a modular platform to assay 10's of functional molecular phenotypes on a single peptide-barcoded variant library, accelerating the functional inference of missense variation on protein function.

5.2 INTRODUCTION

Due to modern day genome and exome sequencing, over 5 million missense mutations in protein-coding genes have been cataloged across the human population[121], however only 2% of them have clinical annotation[122]. Importantly, more than half of these observed missense mutations are categorized as "Variants of Uncertain Significance" (VUS), many existing in disease causing genes and have little to no functional information known about them. To characterize functional VUS, clinicians require reoccurrence of these mutations in the population coupled with extensive phenotypic data to annotate functional effects[123]. This standard approach of annotation requires extensive population-level sequencing and limits our functional characterization to common variants.

To improve annotation of missense variation, Deep Mutational Scanning (DMS) was developed to experimentally assess the functional impact of 10,000's of protein variants for a gene in a single experiment[124]. This method leverages inexpensive DNA synthesis to generate a library of protein variants with a single variant expressed per cell. Cell populations are then subjected to a functional selection to separate protein variants for a particular function. DNA sequencing of pre- and post-selection cells enable an enrichment metric to be calculated for each variant. Importantly, DMS data has successfully resolved many VUS[125]. However, DMS experiments are often labor intensive and indirect, requiring a specific protein's function to be linked to a cellular phenotype. Also, DMS selections generally do not translate to all proteins and all functions of a protein.

Alternatively, functional annotation of protein variants could be improved by identifying and functionally assaying variants at the protein-level, thus decoupling the need for a protein function and cellular phenotype linkage. By probing molecular phenotypes, a single library of

protein variants can be subjected to 10's of generalizable biochemical selections. This approach would greatly expand the functional landscape assayed per protein-coding gene and diversify the set of the protein-coding genes accessible to functional interpretation.

To date, only mass spectrometry (MS) would have the capacity to identify and assay 1,000's of variants at the protein-level in a single experiment. However, identification and quantification of missense proteoforms is difficult due to technical limitations of traditional "bottom-up" mass spectrometry-based proteomics (where proteins are digested to peptides). For instance, proteoform-specific peptides likely have different ionization efficiencies ("detectabilities") making abundance comparisons between them ambiguous. Additionally, redundant peptides between proteoforms aggregate during MS detection, making decoupling proteoform-specific abundances difficult. "Bottom-up"-based proteoform detection also suffers a dynamic range problem because proteoform peptides will likely be at least an order of magnitude lower abundance compared to their wildtype counterparts. In attempt to reduce the wildtype background and deconvolute proteoforms, the field has explored whole protein sequencing, or "top-down" proteomics, however this approach is limited by low identifications rates making functional annotation at scale of 100,000s not feasible.

To address this limitation, a method called NestLink tagged a library of binder proteins each with a unique peptide barcode to encode protein-level detection by MS[126]. In this study, peptide barcodes were linked to protein library members using deep sequencing. Following a protein interaction selection assay, peptide barcodes were analyzed by MS in order to measure their corresponding tagged proteins' binding affinity and kinetics. This "proof-of-principle" study suggests feasibility of using peptide barcodes to detect and reflect functions of their tagged protein variants at the protein-level by MS.

Here, we pilot and advance upon the peptide barcoding strategy to be extended for the functional annotation of a DMS library of protein variants by MS. First, we encode a wildtype thiopurine methyltransferase (TPMT) and the less stability TPMT*3A haplotype by tagging each with the three same peptide barcodes. To ensure our MS approach using peptide barcodes is consistent with DMS approaches, we assess, in parallel, a protein abundance molecular phenotype by MS and a cellular phenotype-based abundance assay called VAMP-Seq[32] for the peptide barcoded TPMT proteins. Then, using wildtype TPMT, we scale to 23,000, MS-compatible, non-human peptide barcodes and apply VAMP-Seq to determine the proportion of peptide barcodes that are biologically inert. Lastly, we employ a novel enrichment module to enable high purity preparation of peptide barcodes for MS analysis. This pilot study provides a foundation for optimal peptide barcode design and MS analysis to functionally annotate 1,000's of protein variants across many molecular phenotypes and many different proteins.

5.3 METHODS

Cell culture

All chemicals and cell culture reagents were obtained from Sigma, Thermo Fisher, or New England Biosciences. HEK293T cells were cultured in Delbecco's modified Eagle's media (DMEM) with 10% FBS, 100 U/mL penicillin, and 0.1 mg/mL streptomycin. TetOn induction was performed with doxycycline (Sigma-Aldrich) at 2 μg/uL. Detachment of cells from plates was performed with Trypsin-EDTA (0.25%).

Peptide barcoding TPMT in VAMP-Seq construct

Peptide barcode modules were DNA oligonucleotides synthesized via a gBlock (IDT) or Agilent array (Agilent). Barcode modules contained a human codon optimized sequence for a lysine amino acid, HA tag, +/- TEV site, and a peptide barcode (in that order). The pilot

experiment did not contain a TEV site and had three peptide barcodes cloned separately: (1) IGDYLGIK, (2) HVLTSLGEK, and (3) IGDYVGIK. The peptide barcoded library contained mutagenized yeast peptides totaling 23,000 peptide barcodes and contained the TEV cut site cloned in a pooled format. The peptide barcode module for the pilot contained HA-tag with an arginine (YPYDVPDYAR) and for the library contained an HA-tag and TEV cut-site (YPYDVPDYA-ENLYFQG). Following BsrGI endonuclease digestion, peptide barcode modules were cloned into VAMP-Seq expression vectors[32] as a linker between GFP and TPMT using Gibson assembly[127].

VAMP-Seq expression vectors were recombined using a previously described protocol in Matreyek et al.[128] into TetOn TetBxb1BFP landing pad HEK293T cells (Clone 4). Successfully transfected and integrated clones were selected by forward/side scatter, mCherry signal, and depleted in BFP signal by FACSAria III (BD Biosciences). Successful integration ranged from 1-5% (Appendix D Supplementary Figure 5.1). Successfully recombined clones were sorted into 6cm dishes and expanded for future VAMP-Seq or proteomics assays at roughly 5 million cells per assay.

VAMP-Seq FACS assay of TPMT peptide barcoded cells

VAMP-Seq assay was carried out as previously described[32]. Successfully recombined clones are $mCherry^+$ and BFP^-, and highly stable peptide barcoded TPMT expressing cells are $mCherry^+$ vs. GFP^+. Unstable peptide barcodes are $mCherry^+$ vs. GFP^-. Using FACS, cells were gated for successful recombination, high mCherry expression control, and varying bins of GFP signal based on stability.

Mass spectrometry assay preparation of TPMT peptide barcoded cells

~ 5-20 million doxycycline induced, TPMT-expressing HEK293T cells (15cm plate) were washed 3 times with PBS. Cell pellets were then snap frozen and stored at -80ºC. CEll pellets were then reconstituted with 500 µL of a lysis buffer (8M Urea, 50 mM Tris pH 8.2, 75 mM NaCl, 1mM orthovanadate, 50 mM β-glycerophosphate, 10 mM sodium pyrophosphate) on ice. Cells were then ultrasonicated for 3 rounds of 15 seconds with 15 seconds recovery on ice in between. Lysate was clarified by centrifugation at 21,000 x g for 10 minutes at 4ºC. Supernatant was extracted and concentration was determined by BCA assay. Pilot experiment peptide barcoded proteomes were normalized for analysis by input material, aiming for them to be equal based on BCA values. Proteins were reduced with 5 mM DTT for 30 minutes with agitation at room temperature (RT), alkylated with IAA at 15 mM final concentration in the dark with agitation at RT, and quenched with 5 mM final concentration of DTT with agitation at RT. For the pilot experiment, lysates were diluted 1:1 with 50 mM Tris pH 8.9 and digested with LysC (1:100 (w:w) LysC:lysate) at RT overnight. For the library of 23,000 peptide barcodes, lysates were diluted 1:5 with 50mM Tris pH 8.2 and digested with trypsin (1:200 (w:w) trypsin:lysate) at 37ºC overnight. Both digestion was halted with the addition of TFA at 1% final concentration.

Peptide mixture was clarified by centrifugation at 21,000 x g for 10 minutes at 4ºC. Peptide mixtures were further desalted using 200 mg Sep-Pak tC_{18} cartridges (Waters) with the same protocol as described previously[33]. Desalted peptide mixtures were aliquoted to known peptide amounts and dried by speedvac.

Peptide barcode enrichment

A 1:1 slurry Anti-HA beads (Sigma) were added to filter columns (3M) and conditioned three times with PBS. 2-3 mg/mL at ~400 µL of peptides was reconstituted with PBS and added to HA beads and incubated for two hours at 4°C. HA beads were washed two-three times with PBS.

For the pilot experiment, beads were washed once more with sterile water, incubated peptides with 100 µL of 0.1% TFA, and then eluted. Eluted peptides were then neutralized with 10 µL of 1M Tris pH 8.9 and digested with 500 ng of trypsin for three hours at 37°C with agitation. Digestion was quenched with 10 µL of 10% TFA (pH < 3).

For the 23,000 peptide barcoded library, 10 µL of TEV stock solution (from Ricard) was combined with 40 µL of 50mM Tris pH 8.2 (1mM DTT and 0.5 mM EDTA). 50 µL TEV solution was added to beads for on-bead TEV digestion for one hour at 30°C. Cleaved peptide barcodes were then eluted off bead and an additional 50 µL of PBS was added and eluted. Enriched peptide barcoded samples were acidified with TFA to final concentration 0.5% (< pH 3).

For both the pilot and 23,000 peptide barcode library experiments, peptide barcodes were desalted using StageTips as previously described[102]. Cleaned up enriched peptide barcode mixture was placed on speedvac and resultant dried peptides were ready for MS/MS analysis.

Mass spectrometry analysis of TPMT peptide barcoded cells

Roughly 1 µg of peptide barcode mixtures (in 5% ACN and 4% formic acid) were subjected to nLC-MS/MS on Easy-nLC 1000 (Thermo Scientific) in-line with a QExactive (Thermo Scientific) hybrid mass spectrometer or an Easy-nLC II (Thermo Scientific) in-line with a Orbitrap Velos Pro (Thermo Scientific). All samples were loaded on a 100µm x 30-cm trap

column with 3 μm C18 beads (Dr. Maisch) then loaded on a column packed with 1.9 μm C18 beads (Dr. Maisch) at 100μm x 30-cm. Peptides were separated by a 45 or 94 minute gradient of acetonitrile (either 80% or 100%), 0.125% formic acid and injected into the MS.

The pilot experiment for peptide barcode 3 (IGDYVGIK) was analyzed on the Orbitrap Velos Pro using a 60 minute run. The duty cycle included an MS1 scan on the Orbitrap (60,000 resolution, scan range 300-1,500 m/z) and MS/MS on the top5 most abundant precursors and added MS/MS scans targeting peptide barcode 2 ($HVLTSLGEK^{2+}$: 492.28 m/z), peptide barcode 3 ($IGDYVGIK^{2+}$: 432.74 m/z), and the HA peptide ($YPYDVPDYAR^{2+}$: 629.79 m/z). All MS/MS were carried out with the following parameters: precursors were isolated with 2 m/z windows, fragmented with CID at 35 NCE, and analyzed by MS/MS with an activation Q of 0.250 and activation time of 10ms. All MS and MS/MS scans were processed and analyzed using Skyline[129].

The 23,000 peptide barcoded TPMT sample was analyzed on the QExactive using a 120 minute run. The duty cycle included an MS1 survey scan (70,000 resolution, scan range 300-800 m/z, 3e6 AGC, 100 ms max injection time) and MS/MS on top20 most abundant precursors in the Orbitrap (2.0 m/z isolations, HCD fragmentation at 26 NCE, 17,500 resolution, 50 ms max injection time, and 5e4 AGC). Spectra were searched against a Uniprot human proteome (Proteome ID: UP000005640, download date 5/26/20218) with peptide barcode sequences appended using Comet[59,105]. Peptide-spectral matching was performed under the following parameters: trypsin specificity (KR|P; max 2), 50 ppm precursor tolerance, 0.02 Da fragment tolerance, variable modifications of oxidation on methionine and acetylation on protein N-terminus, and constant modification of carbamidomethylation on cysteines. PSMs were filtered

at 1% PSM FDR using Percolator[60], and precursor signals were quantified using ThunderQuant (in-house software).

Data analysis and figure generation

All data analysis and data visualization was implemented in R (version 3.6.1) using the Rstudio framework (version 1.4.1103).

5.4 RESULTS

5.4.1 RATIONALE FOR PEPTIDE BARCODE DESIGN

To generate a multiplexed, modular platform to identify and characterize protein variants from a missense library, we implement a peptide barcoding strategy. Similar to the design developed by Egloff et al.[126], peptide barcodes can be cloned terminal to each protein variant in a missense protein library as a unique molecular identifier to represent its tagged variant. Peptide barcodes can then be associated to their protein variants in a pooled format by deep sequencing. To assess a functional proxy of the protein variant, abundance readouts of peptide barcodes by MS pre- and post-molecular phenotyping selection are used to calculate an enrichment score to stratify functional differences between tagged protein variants (Figure 5.1a).

For an optimal peptide barcode design, we will construct a peptide barcode library using DNA array-based synthesis, considering the following criteria (Figure 5.1a). Since equimolar equivalents of peptides can vary in MS abundance, we must identify a set of peptides that are readily observable with uniform MS “detectability”. Additionally, considering a human cell system, peptide barcodes need to be non-human peptides. With access to the swath of MS-identified yeast peptides, we can pilot yeast peptides that are non-redundant with humans to

serve as a starting collection of observable peptide barcodes. We can expect single amino acid substitutions on the C-terminal end of the peptide barcode will expand the library of peptides with similar detectability, while having diverse diagnostic y-ions for unique peptide barcode identification.

Since mass spectrometers utilize a single detector, peptides are separated by liquid chromatography to allow the detection of many peptides by spacing them out over time. We can leverage the known retention time of the yeast peptides to spread the peptide barcodes uniformly over the elution profile. Additionally, we know *a priori* all the peptide barcodes that we need to detect, allowing us to optimize the MS method to capture all peptide barcodes in a targeted manner. Restricting the barcodes to a limited mass range will ensure reproducible identification and quantification. Based on the above-mentioned criteria, we can generate 10,000's of unique peptide barcodes to clone and represent their protein variants in a missense library.

With the molecular selections applied at the protein-level, a vast range of biochemical selections which encapsulate broad functional contexts of proteins can accelerate characterization of these variants. The modular suite of molecular selections available for phenotyping include the following: protein turnover, nucleic acid, small molecular, and protein binding, protein thermal stability, protein solubility, protein activity, and post-translational modifications (Figure 5.1b). These selections can be applied orthogonally in parallel and should extend to all protein variant libraries across the proteome. These molecular selections applied at the protein-level are encoded and readout at the peptide-level by MS using the peptide barcodes. An optimal peptide barcode library can be tagged to all variant libraries and analyzed using a single, targeted MS method to ensure detection of all peptide barcodes in a single MS run.

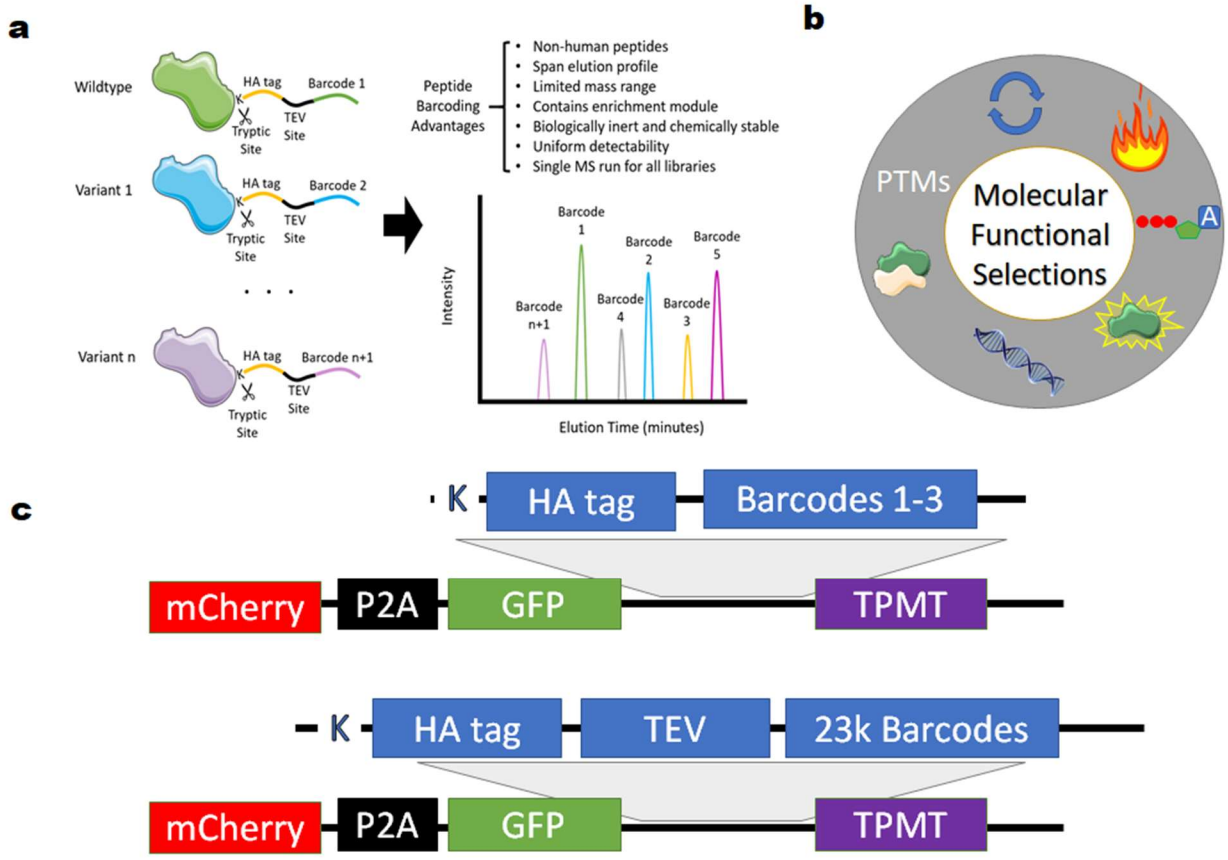


Figure 5.1: Peptide Barcoding Technology. a) Each variant is tagged with a unique peptide barcode. The enrichment module contains HA tag that can be affinity purified after tryptic cleavage, and peptide barcodes can be enriched by TEV protease cleavage off-bead at TEV site. Peptide barcodes are designed to have uniform MS detectability, good chromatographic separation, and lack interference with protein structure and function. **b)** Suite of modular protein-level molecular functional selections that can be applied to protein variant libraries (clockwise): post-translational modifications (PTMs), protein turnover, protein thermal stability, small molecule binding (like ATP), protein activity, nucleic acid binding, and protein interactions/complex formation. **c)** VAMP-Seq construct contains mCherry (red) sequence followed by the ribosomal shuffling sequence, P2A (black), and then GFP (green). Barcode module as linker between GFP and TPMT (purple). Pilot experiment contained a tryptic site (K) followed by a HA tag (blue) and Barcodes 1 through 3 (blue). The bottom construct includes a TEV cleavage site (blue) between HA tag and 23,000 peptide barcodes.

5.4.2 Peptide Barcode Pilot Study with Thiopurine Methyltransferase

We piloted the peptide barcoding approach by tagging the protein thiopurine methyltransferase (TPMT) and a known unstable mutant haplotype TPMT*3A each with the same three peptide barcodes. Peptide barcodes 1 (IGDYLGIK) and 2 (HVLTSLGEK) were

distinct peptide sequences while peptide barcode 3 (IGDYVGIK) contained a single amino acid substitution at the -4 C-terminal of peptide barcode 1. To ensure that the assay results from the MS peptide barcode approach are consistent with known biological phenotypes, we cloned the peptide barcoded TPMT proteoforms into the GPS vector[130] to assess the matching abundance molecular phenotype using the DMS-based assay VAMP-Seq[32] in parallel. Our GPS vector contains Bxb recombinase sites for single copy integration into a "landing pad" HEK293T cell line[128], which is essential for cellular phenotype inference in VAMP-Seq. The P2A sequence between mCherry and GFP tagged-TPMT controls 1-to-1 expression of the resulting polypeptides (Figure 5.1c). In this pilot study, we inserted the peptide barcode sequence between GFP and the TPMT proteoform accompanied with a tryptic cut site and HA-tag for enrichment (Figure 5.1c *top*).

Demonstrated by Matreyek et al.[32], wildtype TPMT and haplotype TPMT*3A should demonstrate high and intermediate protein abundance respectively in the VAMP-Seq assay. VAMP-Seq determines protein variant stability by their steady state abundance in live cells. With mCherry to normalize as an expression control, GFP absorbance in single cells is measured by fluorescence-assisted cell sorting (FACS) to proxy the abundance of its tagged TPMT proteoform. Cells with highly abundant TPMT proteoforms will have high GFP and mCherry absorbance signals.

By FACS, we sorted 25,000 single cells expressing TPMT and unstable TPMT*3A tagged with the three peptide barcodes to ensure our peptide barcodes did not alter the known VAMP-Seq abundance readouts. For TPMT proteoforms tagged with peptide barcodes 1 and 3, TPMT and TPMT*3A GFP signal recapitulated their known high and intermediate abundance

(Figure 5.2). This observation suggests peptide barcodes 1 and 3, which differ in only a single amino acid, likely do not alter TPMT and TPMT*3A stability and abundance. Identification and quantification of these peptide barcodes by MS would likely reflect an accurate functional abundance molecular phenotype of their tagged TPMT proteoforms and could be viable candidates for the final set of inert peptide barcodes. Alternatively, peptide barcode 2 demonstrated extremely low GFP abundance for both TPMT and TPMT*3A proteoforms, suggesting peptide barcode 2 likely alters TPMT proteoform stability and thus is not biologically inert.

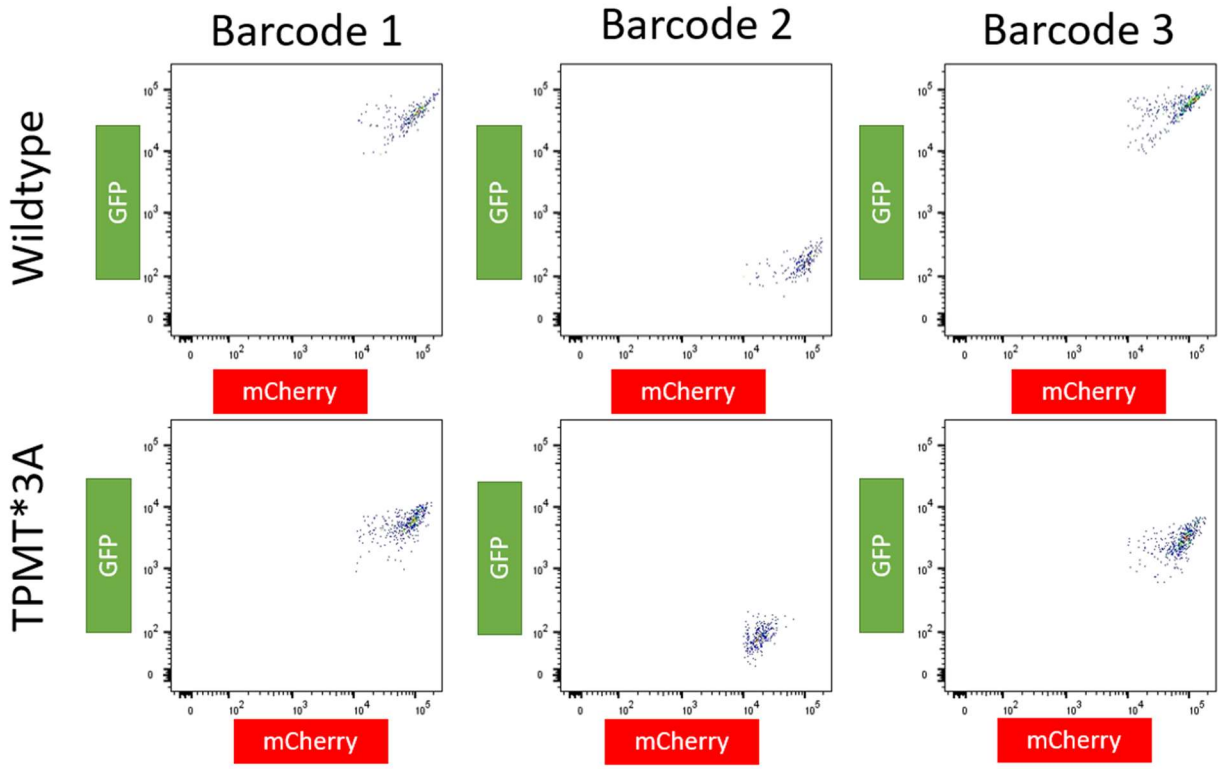


Figure 5.2: VAMP-Seq validation of peptide barcoded TPMT protein's abundance. a) Wildtype and TPMT*3A haplotype proteins are tagged with GFP and a peptide barcode. Peptide barcodes: (1) IGDYLGIK, (2) HVLTSLGEK, and (3) IGDYVGIK. Using a VAMP-Seq approach, 25,000 HEK293T landing pad single cells were sorted by FACS for GFP (stability/abundance) and mCherry (expression control) for protein barcode combinations.

Since tagged TPMT proteoforms with peptide barcodes 1 and 3 recapitulated known abundance readouts by VAMP-Seq (Figure 5.3a), we set out to determine if we can capture the same abundance readout of the proteoforms at the protein-level by measuring peptide barcode abundances with MS. Thus, we expanded HEK293T cells expressing TPMT and TPMT*3A both tagged with peptide barcode 3 for abundance comparisons. To assay the abundance molecular phenotype by MS, the same peptide barcode must be compared between TPMT proteoforms. Cellular proteome input abundance determined by BCA was used to normalize separate cultures and MS preparations between TPMT and TPMT*3A proteoforms. Following trypsin digestion and HA tag enrichment, peptide barcode-enriched samples were subjected to targeted parallel reaction monitoring (PRM) mass spectrometry for the IGDYVGIK peptide barcode. By measuring MS/MS fragment intensity chromatograms, we observed ~10X higher abundance of the peptide barcode for the wildtype TPMT proteoform compared to its TPMT*3A haplotype. The abundance differences by MS successfully reflect the abundance differences observed by VAMP-Seq (Figure 5.3b). However, the peptide barcode for TPMT*3A seems to be approaching the limit of quantification, suggesting optimization by scaling input amount for MS injection will be necessary to capture lower abundance variants and spread the dynamic range of variant abundances for analysis.

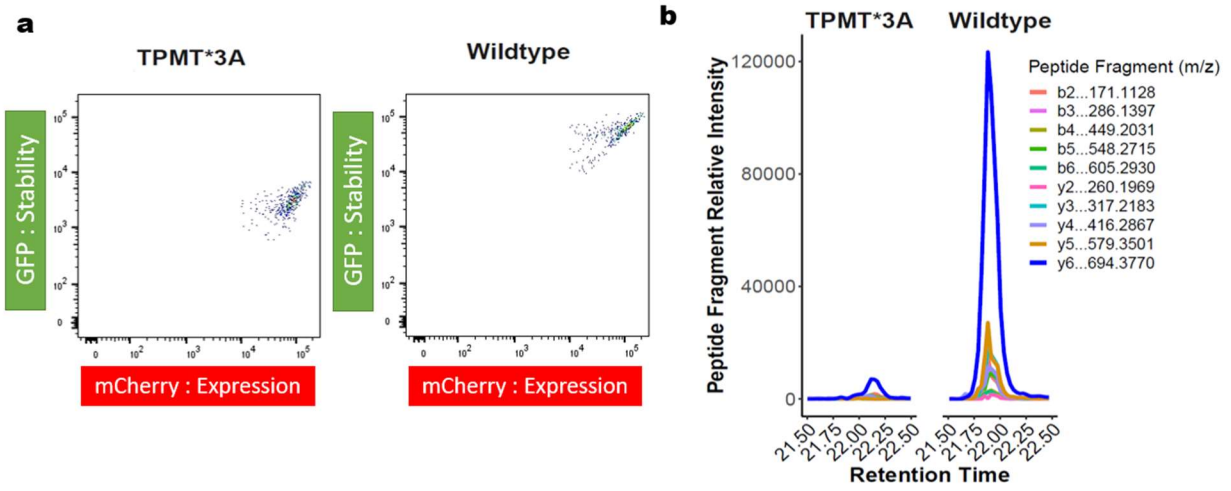


Figure 5.3: Peptide barcode abundance reflects stability difference observed in VAMP-Seq. a) Using Barcode 3 (IGDYVGIK) tagged TPMT and TPMT*3A overexpressed landing pad cell line, FACS sorting results of 25,000 cells as in the right column of Figure 5.2. **b)** Using the peptide barcode IGDYVGIK abundance as a proxy (normalized by cellular input), MS/MS fragment ion intensity of IGDYVGIK peptide barcoded TPMT and TPMT*3A protein over the liquid chromatography elution time.

5.4.3 ASSAYING 10,000'S OF PEPTIDE BARCODES FOR IMPACT ON PROTEIN ABUNDANCE

Essential for the feasibility of a peptide barcode approach, the addition of the peptide barcodes cannot disrupt the protein stability and function being assayed. Peptide barcodes that are not inert can confound the function inference in the assay, making the readout an effect of the barcode and not the selection. Also, the peptide barcode can inaccurately amplify or suppress the effect of the missense variant on the molecular phenotype, if not inert. For instance, the pilot study demonstrated that peptide barcode 2 when tagged to both TPMT and TPMT*3A proteoforms produced an inaccurate, low abundance phenotype. Thus, we set out to explore 10,000's of peptide barcodes tagged to the wildtype TPMT proteoform to assess the prevalence of a tagged peptide barcode altering TPMT protein stability and abundance.

With the goal of generating an optimal set of inert peptide barcodes, we identified 1,642 non-human (yeast) peptide sequences that have uniform MS detectability (350-550 m/z range)

and spread across the elution profile. Next, we *in silico* generated deterministic missense mutations at the C-terminal -3 amino acid position along the sequence to increase our peptide barcode library size to 23,000, while maintaining uniform "detectability". By DNA array-based synthesis, we generated 23,000 oligos that code the peptide product: HA tag, TEV cut site, and each of the 23,000 peptide barcodes (Figure 5.1c *bottom*).

Similarly to the pilot, we cloned the library of 23,000 peptide barcodes into the GPS vector as a linker, C-terminal to GFP and N-terminal to TPMT. The construct was transfected in HEK293T "landing pad cells" for single copy, single cell integration. Across two transfections and two technical sorts, ~82% of the transfected cells contained a peptide barcode that was inert when tagged to the wiltype TPMT proteoform, suggesting most peptide barcodes do not impact TPMT protein stability (Figure 5.4a, Appendix D Supplementary Figure 5.2). DNA sequencing of sorted cell populations will determine which peptide barcodes are destabilizing TPMT and should be removed from the final peptide barcode library.

Due to single copy expression, the sorted cells gated for stable barcodes were expanded to 20 million cells for more feasible MS detection. Peptide barcodes were enriched by the following: (1) digest with Lys-C to decouple linker from TPMT and GFP, (2) pulldown barcode module with HA antibody-coated beads, and (3) digest with TEV protease to release peptide barcodes off bead for MS analysis. We captured ~2,200 distinct barcodes across the four replicates. Unfortunately, due to an error in construct design, a large fraction (~45%) of peptide barcodes were not amenable to the MS preparation because of the peptide barcode C-terminal amino acid was arginine and thus could not be detached from TPMT upon Lys-C digestion. Only ~20% of the observable (lysine C-terminal) peptide barcode library was captured by MS, likely

due to library bottlenecking during transfection and cell expansion. Despite low coverage, the HA-TEV enrichment technique successfully enriched the peptide barcodes to an abundance 3.5-fold greater than the mammalian proteome background (Figure 5.4b).

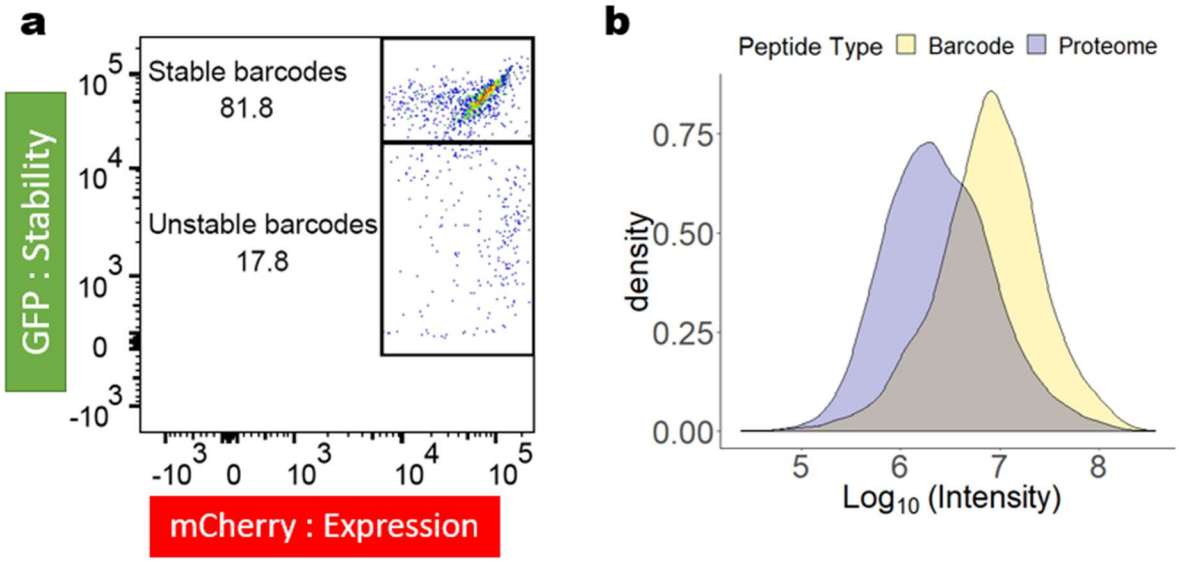


Figure 5.4: Peptide barcodes generally do not alter stability and can be enriched. a) 25,000 HEK293T landing pad cells expressing wildtype TPMT tagged with a library of 23,000 peptide barcodes sorted using VAMP-Seq FACS assay for stability/abundance (GFP) and expression control (mCherry). **b)** Density plot of MS1 precursor intensity of identified peptide barcodes (purple) and HEK293T proteome's peptides (yellow) from "Stable" gated peptide barcoded TPMT landing pad cells.

5.5 Discussion

Herein, we pilot a peptide barcoding strategy to enable detection and biological characterization of TPMT mutational proteoforms directly by MS. By comparing our molecular-based peptide barcode assay to the cellular-based VAMP-Seq assay, we demonstrate initial feasibility that peptide barcodes can accurately reflect known abundance molecular phenotypes by measuring protein molecules directly. However, we observed that peptide barcodes can alter TPMT stability which can influence the accuracy of inference to a functional abundance readout. Thus, we tagged wildtype TPMT with 23,000 unique peptide barcodes. We measured that ~80% of the cells expressing a peptide barcoded TPMT did not have altered GFP abundance, likely suggesting these peptide barcodes are inert. Despite likely bottlenecking the library during the sort and cell expansion, we were able to demonstrate that our peptide barcode enrichment module resulted in highly pure peptide barcode MS samples. From these pilot experiments, we have laid the foundation for using peptide barcoding to detect and assay functions of protein variants.

Despite the progress, our current implementation comes with many limitations and locations for improvement. The largest limitation in this pilot was that the current protocol only resulted in ~2,200 peptide barcodes detected out of a library of 23,000. A large fraction lost was related to having peptide barcodes with C-terminal arginine which was not amenable to our current MS preparation. This could be improved by only synthesizing peptide barcodes with C-terminal lysine.

Also, we lost many peptide barcodes from bottlenecking our library during transfection, FACS, and cellular expansion steps. Since the transfection efficiency was below 5%, we were

only able to sort a limited number of cells, resulting in on average only a few cells per peptide barcoded TPMT. Thus, we likely lost representation of many peptide barcodes by chance, and unequal barcode frequencies among the cell population was likely present from the transfection and sorting steps. Additionally, many cells did not survive the transfection and sort protocol further causing bottlenecking of the library. Of note, to capture enough peptide barcode signal, we expanded 10,000's of cells from the sort to ~20 million, likely bottlenecking the peptide barcode library further. VAMP-Seq requires single cell integration, however a peptide barcode MS approach does not share this requirement. By shifting to a multiple copy per cell system, we can greatly accelerate workflow and remove sources of library bottlenecking.

A key observation from this pilot study is that peptide barcodes when tagged to TPMT can alter protein abundance and likely its protein functions. Since functional inference relies heavily on the assumption that the peptide barcodes are inert, this work illustrates the importance of generating an optimal set of inert peptide barcodes. We employed the FACS stability assay with a library of peptide barcodes tagged to wildtype TPMT, which coupled with deep sequencing could prioritize peptide barcodes that likely will not impact protein stability and function. An alternative and complementary strategy is to perform the functional selections with the peptide barcode library tagged to wildtype protein and to the missense protein library. Peptide barcodes that alter the wildtype phenotype can be removed from the analysis of the mutational library. Lastly, having many peptide barcodes per protein variant could identify peptide barcodes that are not inert apparent by outlier function readouts.

The first implementation of peptide barcodes was implemented for binding assays of nanobody protein libraries. With our observations that peptide barcodes can alter protein stability

and abundance, one could postulate that some of their peptide barcodes could be affecting binding kinetics and thus altering their functional readout. In the context of nanobodies, which are generally thermally stable, it is unlikely peptide barcodes would alter stability. Having control experiments and optimizing peptide barcodes for "inertness" could improve experimental design for generalized implementation and decrease the probability of inaccurate functional readouts. Further experiments will need to be explored to assess whether a peptide barcode "inertness" translates across proteins and biochemical selections.

Much work needs to be done to generate an optimal peptide barcode library that ensures reliable functional annotation of missense variant libraries. However, this pilot study highlighted some key considerations for generating peptide barcodes that will be essential for final implementation. The key finding herein is that peptide barcodes can alter protein functions of the proteins they are representing. Thus, proper controls and development of assays to address peptide barcode "inertness" will be essential in defining an optimal peptide barcode library. Considering the development of an optimal peptide barcode library, we foresee exciting potential for a peptide barcode approach in annotating function of protein variants in a pooled format. A peptide barcode approach could accelerate the protein variant interpretation and provide evidence to resolve VUS, complementing deep mutational scanning approaches.

Contributions: Experiments and data analysis were performed by Ian Smith. Technical guidance was given by Lea Starita, Vijay Ramani, Beth Martin, and Ricard Rodriguez. Jay Shendure, Vijay Ramani, Judit Villén, and Ian Smith designed the experiments.

Chapter 6. COISOLATION PEPTIDE PAIRS FOR PEPTIDE IDENTIFICATION AND MS/MS-BASED QUANTIFICATION

6.1 ABSTRACT

SILAC-based metabolic labeling is a widely adopted proteomics approach that enables quantitative comparisons among a variety of experimental conditions. Despite its quantitative capacity, SILAC experiments analyzed with data dependent acquisition (DDA) do not fully leverage peptide pair information for identification and suffer from undersampling compared to label-free proteomic experiments. Herein, we developed a data dependent acquisition strategy that coisolates and fragments SILAC peptide pairs and uses y-ions for their relative quantification. To facilitate the analysis of this type of data, we adapted the Comet sequence database search engine to make use of SILAC peptide paired fragments and developed a tool to annotate and quantify MS/MS spectra of coisolated SILAC pairs. In an initial feasibility experiment, this peptide pair coisolation approach generally improved expectation scores compared to the traditional DDA approach. Fragment ion quantification performed similarly well to MS1-based quantification, and achieved more quantifications within less than 2-fold of the expected ratio. Lastly, our method enables reliable MS/MS quantification of SILAC proteome mixtures with overlapping isotopic distributions, which are difficult to deconvolute in MS1-based quantification. This study demonstrates the initial feasibility of the coisolation approach. Coupling this approach with intelligent acquisition strategies has the potential to improve SILAC peptide sampling and quantification.

6.2 INTRODUCTION

Stable-isotope labeling with amino acids in cell culture (SILAC)[131,132] is a powerful tool in quantitative proteomics for comparing biological samples. While chemical labeling strategies like tandem mass tags[133–135] have increased multiplexing capabilities, SILAC is still widely used for its quantitative accuracy[136] and its ability to metabolically-encode temporal information[137]. For instance, dynamic SILAC[26,27] and pulsed-SILAC[138,139] approaches have become the field standard methods for measuring protein turnover *in vivo* at a proteome-wide scale.

In a SILAC experiment, proteome mixtures with isotopically-labeled amino acids are analyzed by LC-MS/MS using DDA and the relative quantification of peptide pairs is obtained from the peptide precursor signals in MS1 scans[131]. Comparatively to label-free DDA samples, SILAC samples have better quantitative precision due to decreased technical variation, however SILAC suffers from redundant sampling of peptides and as a result fewer peptides are quantified.

As an alternative to traditional DDA, data independent acquisition (DIA) strategies[64,140,141] with quantification on the MS/MS have also been applied to SILAC samples, improving sampling reproducibility and quantification accuracy. Alternative MS acquisition methods such as BoxCarmax[142], which leverages a combination of BoxCar[143], multiplexed MS/MS DIA[144], and gas phase fractionation[145], further improve sampling and quantification. However, the wide-window isolation used in most DIA experiments can compromise SILAC quantification, given that asymmetric isolations of the peptide pair can lead to distorted SILAC ratios[146].

Data-dependent acquisitions that are informed by SILAC peptide pairs have also been developed. For instance, the Mann group demonstrated enhanced quantification of SILAC pairs with poorly defined ratios in the survey MS1 scan by triggering selected ion monitoring scans in

real-time[147]. Additionally, the Coon group demonstrated reliable quantification by targeted coisolation and fragmentation of the SILAC peptide pair in direct infusion MS experiments[148].

Experiments using other stable isotope labeling strategies, such as trypsin-catalyzed ^{16}O-to-^{18}O-exchange and d0/d3 methyl esterification, have leveraged peptide pairs for identification by comparing heavy and light peptide's spectra to assist in *de novo* peptide sequencing[149–151] and automated peptide search validation[152]. For quantification, Heller et al.[153] showcased the coisolation of heavy and light ^{16}O-^{18}O peptide pairs for MS/MS and utilized y-ion fragment pairs for relative quantification. Other methods using chemical[154] or metabolic[155] isotopologue labels have demonstrated the feasibility of using isotopically distinct fragment ions for relative MS/MS quantification with accuracy and precision. Analogously, coisolation of SILAC peptide pairs would result in a boost of b-ion fragment MS/MS signal and paired y-ion SILAC MS/MS signal that can be leveraged for database search identification and MS/MS-based quantification.

Here, we implement a MS acquisition to coisolate SILAC peptide pairs for MS/MS. To analyze coisolated MS/MS, we adapt Comet to perform peptide-spectral matching (PSM) using theoretical spectra of SILAC peptide pairs and we develop a tool to quantify SILAC y-ion pairs from MS/MS spectra. We demonstrate that our method can successfully identify SILAC peptide pairs, while enabling both MS1 and MS/MS-based quantification. We further expand the capabilities of our method to accurately quantify SILAC peptide pairs with overlapping isotopic distributions. Collectively, this work expands our proteomic toolkit for quantitative analysis of SILAC samples.

6.3 METHODS

Yeast growth and harvest

Two *Saccharomyces cerevisiae* yeast strains were used to generate SILAC proteome mixtures: DBY10144 and BY4742. *Saccharomyces cerevisiae* DBY10144 diploid strain (MATa/α) is a lysine prototroph from the FY (S288C) background (parental strains FY3G and FY4H). *Saccharomyces cerevisiae* BY4742 haploid strain (MATα) is a lysine auxotroph from the FY (S288C) background (parental strain FY2). Two starter cultures (synthetic complete media: 6.7 g/L yeast nitrogen base, 2% glucose, and 2 g/L of drop-out mix with all amino acids except lysine) were grown overnight at 30 ºC, one spiked with heavy $^{13}C_6$,$^{15}N_2$-lysine and other with light lysine $^{12}C_6$,$^{14}N_2$-lysine (both final concentration 0.872 mM for DBY10144 and 0.436 mM for BY4742). Both cultures were diluted using the same media composition (heavy- and light-lysine media respectively) to OD_{600}=0.1 (DBY10144) and OD_{600}=0.05 (BY4742). For DBY10144 pellets, yeast cell growth was stalled at OD_{600}=~0.75 (~8 doublings with overnight cultures) with 100% trichloroacetic acid (final concentration 10%) and cultures were harvested by centrifugation at 7,000 x *g* for 10 minutes at 4 ºC. Supernatants were decanted and cell pellets were washed with ~10 mL of chilled 100% acetone. Acetone-washed cell pellets were centrifuged at 7,000 x *g* for 10 minutes at 4 ºC, decanted, and cell pellets were snap frozen with liquid nitrogen and stored at -80 ºC. For BY4742 pellets, yeast were cultured overnight and harvested at OD_{600}=~0.85 (~8 doublings with overnight cultures) by centrifugation at 7,000 x *g* for 10 minutes at 4 ºC. Supernatants were decanted and cell pellets were washed with 2 mL chilled deionized water. Cell pellets were centrifuged at 7,000 x *g* for 10 minutes at 4 ºC, decanted, and cell pellets were snap frozen with liquid nitrogen and stored at -80 ºC.

Cell lysis, protein reduction and alkylation, and protein digestion

Cell pellets were resuspended on ice with 600 µL denaturation buffer composed of 8 M urea, 50 mM Tris pH 8.2, and 75 mM NaCl. Phosphatase inhibitors (10 mM sodium pyrophosphate, 50 mM of sodium fluoride, and 50 mM β-glycerophosphate) were added to the DBY10144 denaturation buffer. Cells were lysed by mechanical agitation using 0.5 mm zirconia/silica beads (60 second bead beating then 90 second rest on ice, repeated four times). Lysates were crudely clarified by centrifugation at 1,200 x *g* for 1 minute to remove beads followed by cell debris removal via centrifugation at 7,000 x *g* for 10 minutes at 4 ºC. Protein concentration for heavy- and light-lysine lysates was determined by BCA assay (Pierce). Clarified lysates were reduced at 5 mM dithiothreitol (DTT) for 30 minutes at 55 ºC, alkylated with 15 mM iodoacetamide in the dark with agitation for 30 minutes at room temperature, and quenched with an additional 5 mM DTT at room temperature for 30 minutes with agitation. Reduced and alkylated samples were diluted 1:1 (v:v) with 50 mM Tris pH 8.9 and 75 mM NaCl to a final pH ~8.5 (DBY10144 samples Tris pH 8.9 buffer contained phosphatase inhibitors: 10 mM sodium pyrophosphate, 50 mM of sodium fluoride, and 50 mM β-glycerophosphate). Lysates were digested with Lysyl endopeptidase (LysC; Wako Chemicals in HPLC grade water) 1:100 enzyme to protein substrate ratio overnight at room temperature. LysC digestion was quenched with trifluoroacetic acid (final concentration 1%), and digested peptides were placed at -80 ºC.

Peptide desalting

50 mg Sep-Pak tC18 polymer columns were used to clean up 1.5-1.7 mg of digested yeast lysate. 1 mL methanol was used to activate the column, then the following were used to equilibrate the column: 3x 1 mL 100% acetonitrile, 1 mL 70 % acetonitrile with 0.25% acetic

acid, 1 mL 40% acetonitrile with 0.5% acetic acid, and 3x 1 mL 0.1% trifluoroacetic acid (TFA). Peptides were loaded twice by gravity. 3x 1 mL 0.1% TFA and 1 mL 0.5% acetic acid were used to wash the column. Then, 600 µL 40% acetonitrile with 0.5% acetic acid and 400 µL 70% acetonitrile with 0.25% acetic acid were used to elute peptides. IMAC enrichments were performed on an aliquot of ~400 µg peptides (2 enrichments at 200 µg each). Multiple aliquots of 50 µg or 100 µg peptides were used to generate SILAC mixtures of lysine light to heavy ratios (10:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:10) for mass spectrometry analysis. All sample aliquots were dried by vacuum centrifugation and stored at -80 ºC.

Liquid chromatography mass spectrometry data acquisition strategies

SILAC peptide mixtures (500 ng) were subjected to nLC-MS/MS on a EASY-nLC 1200 (Thermo Fisher) coupled to a Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher). Peptides were loaded on a 100 µm x 3 cm trap column packed with 3 µm C18 beads (Dr. Maisch) and separated using a 90 minute reversed-phase gradient of 80% acetonitrile, 0.1% formic acid on a 100 µm x 30 cm analytical column packed with 1.9 µm C18 beads (Dr. Maisch). MS acquisitions were acquired using data-dependent acquisition with a cycle time of 3 seconds starting with a full MS1 scan on the Orbitrap over 300-1,500 m/z at 120,000 resolution, normalized automatic gain control set to Standard (100% - 4e5), and injection time set to Auto (max 50 ms). For traditional data-dependent acquisitions, MS/MS was acquired for the most intense m/z precursors (z=2-5) over the 3 second cycle time using the following parameters: dynamic exclusion of 30 seconds, 1.6 m/z isolation window of precursors, HCD fragmentation at 30 normalized collision energy (NCE), 30,000 resolution on the Orbitrap, normalized automatic gain control set to Standard (100% - 5e4), and injection time set to Auto (max 54 ms). For offset

left and right wide window scans (Coiso scans) applied to K0/K8 SILAC mixtures, the same acquisition parameters were used as the DDA MS/MS scan however each triggered precursor was isolated with a 6.5 m/z isolation window, offset -4 Da and 4 Da (left and right respectively). For comparing Coiso scans and DDA scans, the same precursor was subjected to left and right wide window scans and the DDA scan with the same respective parameters as above. For K6/K8 SILAC mixtures, the isolation offset was set to -1 Da and 1 Da (left and right) with a 5.0 m/z isolation window.

SILAC peptide pair database search approach and new parameters

The Comet[59,156] search engine was extended to perform peptide spectral matching on SILAC peptide coisolation theoretical fragments, controlled by the search parameter "silac_pair_fragments". In addition to specifying whether a standard database search or a coisolation search is performed, this parameter also controls whether to apply the coisolation fragment peaks on both the b- and y-ion series (silac_pair_fragments=1) or only on the y-ion series (silac_pair_fragments=2). Of note, we demonstrated that only paired y-ions should be considered (excluding possible paired b-ions) in peptide-spectral matching (Appendix E Supplementary Discussion) for it standardizes the possible theoretical fragments based on peptide length across all possible candidates, removes a bias towards missed cleaved decoys (Appendix E Supplementary Figure 1a), generates more PSMs (Appendix E Supplementary Figure 1b), and improves E-values (Appendix E Supplementary Figure 1c).

To perform a SILAC coisolation search, a static modification is applied to set the mass of lysine to the light SILAC mass and the mass difference between the light and heavy SILAC reagents is set as a variable modification on lysine residues (e.g. 8.014199 for K0/K8 or 1.99407

for K6/K8). All matching candidate peptides are scored by the fast cross-correlation (Xcorr) algorithm[157].

Standard E-value and Coisolation E-value for SILAC peptide pair prioritization

Comet calculates an expectation value (E-value) for each reported PSM. For a given PSM, the E-value is defined as the expected number of random peptides that score as well or higher than the PSM's cross-correlation score (Xcorr). The E-value is calculated based on modeling the incorrect score distribution which, in the case for Comet, is the histogram of random Xcorr scores for each spectrum query searched against candidate peptides from the sequence database. To calculate an E-value, the Xcorr histogram is converted to a cumulative distribution function (CDF) and a linear regression is fit to the right tail of the log10 transform of the CDF. Xcorr scores are extrapolated from the linear regression model to determine each PSM's E-value[59].

For a coisolation search, the reported E-value is based on the coisolation Xcorr. Additionally, Comet reports a second E-value (Coiso E-value or e.value_paired) based on Xcorr scores calculated on just the non-triggered fragment ion peaks e.g. only the light fragment ions if the MS/MS spectrum was triggered on a heavy precursor or only the heavy fragment ions if the MS/MS spectrum was triggered on a light precursor. This strategy is similar to calculating the delta Xcorr between the coisolation database search Xcorr (Coiso) and the traditional DDA database search Xcorr (DDA). For each peptide spectral match, all candidate peptides (target or decoy within 20ppm of targeted m/z) will have the delta Xcorr calculated building a delta Xcorr distribution. Similar to calculating a Comet E-value, we can calculate a Coiso E-value based on the expectation value at which the top PSM candidate's delta Xcorr intersects with the linear

regression of the right tail of the -$\log_{10}$(ΔXcorr CDF). This score accurately prioritized the correctly coisolated SILAC scan with a drastically lower E-value than the same precursor incorrectly coisolated. The Coiso E-value score is used to determine when a spectrum has correctly coisolated both light and heavy precursors or incorrectly coisolated only one of the two precursors.

Database searching S. cerevisiae data with Comet

Prior to database searching, MS raw files were converted to .mzML format using msconvert[158]. The .mzML files from Lys0/Lys8 proteome mixtures were searched with Comet, which can be located at https://sourceforge.net/p/comet-ms/code/1622/tree/branches/release_2019015_silacpair/. The following database search parameters were used: a SGD *S.cerevisiae* protein sequence database (July 2014), searching for b-ions, y-ions, and H_2O/NH_3 neutral loss fragments, LysC endopeptidase specificity (C-terminal to lysine; max 2 missed cleavages), fixed modification of cysteine carbamidomethyl and [+6.020129] on lysines only when $^{13}C_6$-lysine was the light SILAC label, variable modifications of oxidation on methionines, acetylation on protein N-terminus, and heavy lysine delta mass (based on mass difference between light to heavy lysine labels), mass tolerance of 20 ppm for precursor m/z, and mass tolerance of 0.02 Da for fragment ions.The parameter isotope_error was set to 3 for all SILAC ratios used except K6/K8 SILAC mixtures where isotope_error was set to 1. Coiso SILAC searches were designated silac_fragment_pairs = 1 or 2 (including and excluding paired b-ion fragments respectively, while traditional DDA searches had silac_fragment_pairs=0). Comet generated a pep.xml, pin, and tab delimited text output files for downstream analysis.

Peptide spectral match (PSM) FDR filtering and MS/MS spectral quantification

We developed a computational Python suite that integrates a variety of publicly available MS software with custom code to generate MS1 and MS/MS quantifications of PSMs from Comet database search results. Using .mzML spectral files, Dinosaur[159] was used to identify and quantify MS1 spectral features. The Python script takes .mzML files, Comet generated .pin files, and Dinosaur generated .feature.tsv files as inputs. Additionally, Comet generated .pin files could be filtered for PSM results pertaining to correctly coisolated wide window scans, based on Coiso E-value score prioritization, and its corresponding narrow window scan from the same precursor (resultant .pin taken as input). The following steps are conducted using our coisolation Python (v.3.8.1) script: (1) Using Pyteomics[160,161], relevant spectral header information and MS/MS spectra m/z and intensities are extracted from .mzML files. (2) PSM results from Comet database search engine are FDR filtered at 1% PSM FDR using mokapot[162], a Python algorithm similar to Percolator[60]. (3) After merging PSM FDR-filtered results with spectral data, custom code calculates theoretical masses of b-ion and paired y-ion fragments (z=1&2) for the monoisotopic and isotope errors 1&2. Then, MS/MS peaks were matched to these theoretical masses (maximum intensity observed within 50 ppm tolerance). (4) MS/MS spectral noise was determined using the median of all spectral peaks in the MS/MS scan from the .mzML file[163], and signal-to-noise ratios for heavy and light fragments were calculated for the average peak intensity of topN, top3, and top6 fragments with quantifiable heavy-light fragment pairs divide by noise. (5) Only annotated fragment isotopes (monoisotopic and/or isotope error peaks with intensity greater than MS/MS spectral noise intensity) that were observed for both heavy and light peptides were considered and, and isotope intensities observed in both heavy and light forms were summed to represent the fragment's heavy and light intensities. (6) MS/MS

quantifications were generated using the topN and top2-top6 paired heavy-light lysine paired fragments for the weighted average or median heavy/light ratios, excluding y^{1+} fragments. Top3-top6 paired fragment quantification heavy and light intensities were fit to a linear regression to extract SILAC ratio based on the slope with accompanying linear fit coefficient of determination (R^2). (7) Dinosaur features were mapped to mokapot PSM-filtered results using the following criteria: PSM with retention times within the bounds of the peptide MS1 feature and MS1 feature's m/z within 50 ppm of the theoretical PSM m/z. If multiple features map to a single PSM, the feature with the max intensity to the PSM m/z was chosen. (8) All steps were compiled into three output .csv files; one with PSM level summary MS1 and MS/MS quantifications, the second with all annotated heavy and light MS/MS fragments, the third with annotated pair y-ion fragments only used for quantification. The source code and coiso_silac Python package can be accessed on GitLab at https://gitlab.com/public_villenlab/coiso_silac.

Lysine6/Lysine8 MS/MS spectral deconvolution and quantification

For $^{12}C_6$-lysine (Lys6) and $^{13}C_6$,$^{15}N_2$-lysine (Lys8) SILAC proteome mixtures, the isotopic distributions of coisolated precursor MS1 and MS/MS peptide paired y-ion fragments overlap and require deconvolution for quantification. MS/MS spectral processing, FDR filtering, and Dinosaur MS1 feature mapping were applied as described above. To determine heavy and light fragment intensity contributions, all theoretical paired y-ion fragment intensities were extracted from the MS/MS scan for an extended isotopic distribution profile (monoisotopic peak and isotope errors 1-4) of the light peptide fragment based on the calculated theoretical m/z values. Missing isotope peaks were assigned 0 MS/MS intensity. Light fragment intensities contribute up to all 5 theoretical peaks (monoisotopic and isotopic errors 1-4), while the heavy

fragment intensities only contribute to isotopic error peaks 2-4. To demarcate the relative contribution of the light and heavy fragments to the isotope error intensities 2-4, the chemical composition of each y-ion fragment for light peptide sequence was determined, and using the Yergey et al. calculation[164], the theoretical isotopic distribution across the isotopic profile (normalized to the monoisotopic contribution) was calculated.

Using the mathematical approach developed by Chavez et al.[165], each y-ion fragment's observed isotopic distribution (containing heavy and light intensity contributions for the light monoisotopic peak and isotope error peaks 1-4) and the fragment's theoretical isotopic distribution serve as inputs to compute the fragment's optimal SILAC ratio (the heavy/light ratio that minimizes the ratio error in their model). Our implementation deviates from Chavez et al.'s approach in that we calculate each y-ion fragment's theoretical isotope distribution based on each fragment's molecular composition (compared to using the precursor theoretical distribution for all peptide's fragments) and we choose a different set of filters for accepting a fragment's ratio. We applied a filter that the quantification for each fragment can only be calculated if at least two isotopic peaks were observed across the isotopic profile and the heavy and light contributions to the optimal SILAC ratio calculation must be positive values. The PSM's optimal SILAC ratio and SILAC ratio error were designated by the median of the topN and top3 fragments' optimal ratio and ratio error for each PSM.

MS1-based quantification was determined via the deconvolution of MS1 precursor signals similarly to MS/MS peptide fragments, using the theoretical distribution of the precursor peptide and observed isotopic distribution for the nearest MS1 or apex MS1 scan for the closest 16 MS1 scans from the triggered MS/MS scan number. Top3 and Top5 MS1 observed isotopic distributions were calculated by summing the isotopic contributions for the 3 or 5 most abundant

isotopic distribution signals across subsequent scans. FDR-filtered PSMs with MS1 and MS/MS deconvoluted quantifications were returned as a summary .csv file.

Data analysis

Spectra for MS1 and MS/MS scans pertaining to figures were directly extracted from .mzML using custom code or Pyteomics[160,161] and MS/MS spectra were annotated with our custom Coiso annotation code or spectrum_utils[166]. Resultant spectra were plotted in R (version 3.6.1) and RStudio (version 1.4.1103), and all data figures were generated in Adobe Illustrator CS5 (version 15.0.0) and R. All code and data analysis can be accessed via GitLab at https://gitlab.com/public_villenlab/coiso_silac_analysis. All mass spectrometry data and analysis files generated for this manuscript will be deposited to the ProteomeXchange Consortium by the PRIDE partner upon submission.

6.4 Results

6.4.1 Rationale for coisolation SILAC acquisition and database searching

To isolate SILAC pairs, we use DDA to detect the precursor m/z and offset a 6.5 m/z isolation window to center the light and heavy pair for MS/MS. For Lys0:Lys8 SILAC mixtures, isolation centers are 4 Da from the precursor m/z based on the precursor charge state. Light and heavy precursors are thus coisolated when isolations are offset to the right and left, respectively. In initial MS acquisitions, we ensure pair coisolation by performing both offset isolations for MS/MS, and compare to DDA MS/MS for the same precursor (Figure 6.1a). Peptide-spectral

matching of the pair's MS/MS spectra (Figure 6.1b) can be performed using Comet for peptide assignment.

We expect the coisolation and analysis of peptide pairs will yield three improvements. First, fragmentation spectra of coisolated SILAC peptide pairs feature merged b-ions and split y-ions (Figure 6.1b top panel). In a Comet search, the increased ion representation should increase Xcorr values (Figure 6.1b). Second, to overcome the increased complexity in a wide isolation window, we apply a narrow mass tolerance to increase the specificity of the precursor. We expect these two features will prioritize the true pair over other candidates and paired decoys, thus improving Comet E-values and overall identifications. Third, quantification can be performed in the MS/MS using the y-ions, which are expected to have less interference than the precursor signals.

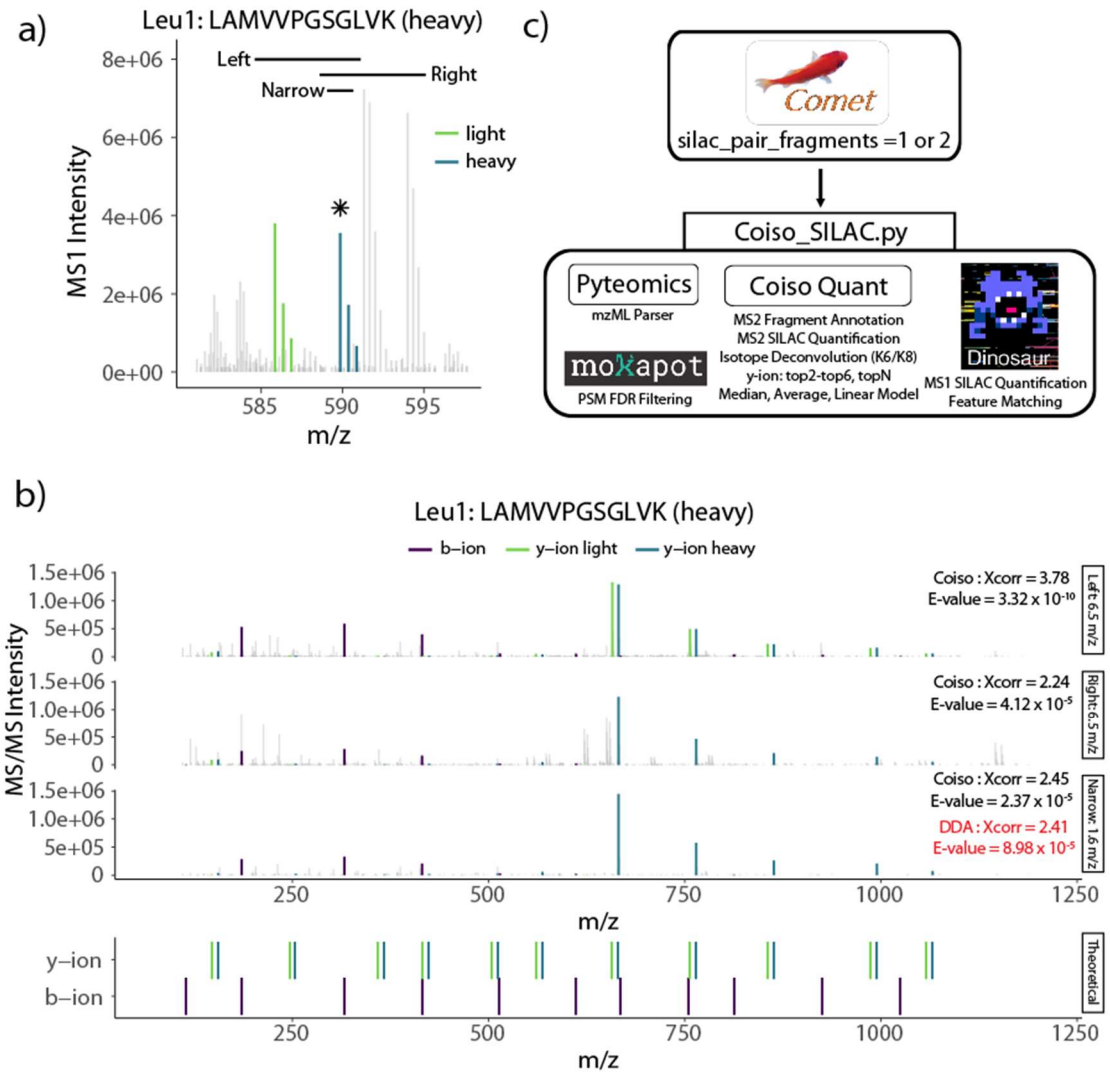


Figure 6.1: Coiso SILAC computation workflow and MS acquisitions. a) From the MS1 scan, Coiso 6.5-m/z wide isolation window (offset -4 Da left and +4 Da right) and DDA narrow window 1.6 m/z isolation window for a triggered heavy SILAC peptide from Leu1 (asterisk). Coiso MS acquisition aims to capture SILAC peptide pairs (light (green) and heavy(blue)) for fragmentation and MS/MS. **b)** MS/MS of left and right Coiso scans and narrow window DDA scan for the triggered Leu1 peptide in (**a**). Comet's E-values and Xcorr for the Coiso search are in black and standard DDA search are in red. Theoretical fragments for Coiso search are in the bottom panel at theoretical m/z values. Peptide spectral matched b-ions (purple), light y-ions (green), and heavy y-ions (blue) are highlighted and other MS/MS peaks are grey-scaled. **c)** Comet database searching for Coiso SILAC MS acquisitions performs peptide spectral matching to SILAC peptide paired fragments designated by the parameter "silac_pair_fragments". PSMs are matched to Pyteomics-parsed MS/MS spectral data, FDR-filtered with mokapot, mapped to Dinosaur MS1 features, and PSM y-ion paired fragments are annotated and quantified.

6.4.2 Adapting Comet database search engine for SILAC peptide pairs

To analyze SILAC pair data, a two step analysis pipeline was generated. First, Comet was adapted to perform peptide spectral matching (PSM) for SILAC pairs using *in silico* heavy

and light fragments. This feature is enabled by the "silac_pair_fragments" parameter. For all candidate targets and decoys within the MS1 ppm tolerance of the precursor m/z, theoretical paired spectra are generated and matched on-the-fly to calculate Xcorr scores (cross correlation score)[157]. With calculated Xcorr values, PSMs can be assigned an E-value, or an expectation score, as a metric to standardize the confidence of the reported PSMs and calibrate Xcorr values across scans. E-values serve as an effective single metric to differentiate target and decoy PSMs[59], thus enabling FDR-based PSM filtering to establish a high confidence set of identifications for downstream analysis.

The second component of the analysis pipeline is an open source Python package that performs "all-in-one" SILAC quantification with coisolation data. In the package, we integrate the publicly available software mokapot[162] to FDR filter Comet PSMs and Pyteomics[160,161] to extract MS spectral data. With custom code, we annotate peptide fragments to MS/MS spectra, calculate MS/MS-based SILAC ratios, and map Dinosaur[159] MS1 features to PSMs (Figure 6.1c). MS/MS SILAC ratios are generated for a variety of different summation strategies and variable number of most abundant topN paired y-ions. This quantification pipeline generates a result file containing 1% FDR filtered PSMs with scoring metrics and MS1 and MS/MS quantifications for downstream analysis.

We assess method feasibility by comparing peptide-spectral matching metrics and identifications between SILAC pair MS/MS and DDA MS/MS for the same precursors. Also, we evaluate quantification precision and accuracy for MS1 precursor and MS/MS y-ion pair features.

6.4.3 COISOLATING SILAC PAIRS IMPROVES IDENTIFICATION METRICS COMPARED TO DDA

To make SILAC pair identifications, we must couple the coisolation MS/MS acquisitions with the adapted Comet database search (Wide Coiso). We evaluate the performance of the coisolation method by comparing our peptide-spectral matching scores and identifications to single precursor isolations with a standard Comet search (Narrow DDA). We leveraged the MS isolation schema from Figure 6.1a to analyze seven *S. cerevisiae* proteome mixtures across a large dynamic range of SILAC ratios (10:1, 4:1, 2:1, 1:1, 1:2, 1:4, and 1:10; lysine+0: $^{13}C_6$,$^{15}N_2$-lysine+8). Initial coisolation MS acquisitions were designed based on SILAC peptides with a single lysine. Thus, Wide Coiso PSMs that did not have one lysine were removed, along with the precursor's corresponding Narrow MS/MS scan. Since both Wide offset scans are analyzed for each precursor, only the Wide offset scan containing the SILAC pair was compared to the matching Narrow MS/MS scan (See Methods).

In the 1:1 SILAC labeled sample, we observed larger Xcorr scores (96%) for Wide Coiso compared to Narrow DDA (Figure 6.2a). This is as expected because the additional paired y-ions and the boost in b-ion signal should result in a more representative signal for Wide Coiso. Importantly, the standardized E-value metric showed a 42% improvement for Wide Coiso compared to Narrow DDA for matching PSMs at 1% FDR (Figure 6.2b). Generally, the E-value distributions across all considered PSMs between Wide Coiso and Narrow DDA reflected similar distributions (Appendix E Supplementary Figure 6.2a). For Coiso Wide, we observed increases in the distribution of $-\log_{10}$(E-values) for decoys (Appendix E Supplementary Figure 6.2b) and targets (Appendix E Supplementary Figure 6.2c). Coiso Wide targets particularly demonstrated increases at the right tail of the $-\log_{10}$(E-value) distribution. A receiver operating characteristic

(ROC) plot demonstrates that Wide Coiso E-value is a similarly predictive metric for distinguishing between targets and decoys PSMs as Narrow DDA E-values (Figure 6.2c). For total PSM identifications at 1% FDR, Wide Coiso captures ~92% for 1:10 and 10:1 SILAC ratios, ~94% for 1:4 and 4:1 SILAC ratios, ~97% for 1:2 and 2:1 SILAC ratios, and ~97% for 1:1 SILAC ratio compared to Narrow DDA (Figure 6.2d). Thus, we conclude that our coisolation pair method provides a similar capacity to identify SILAC peptide pairs proteome-wide as DDA.

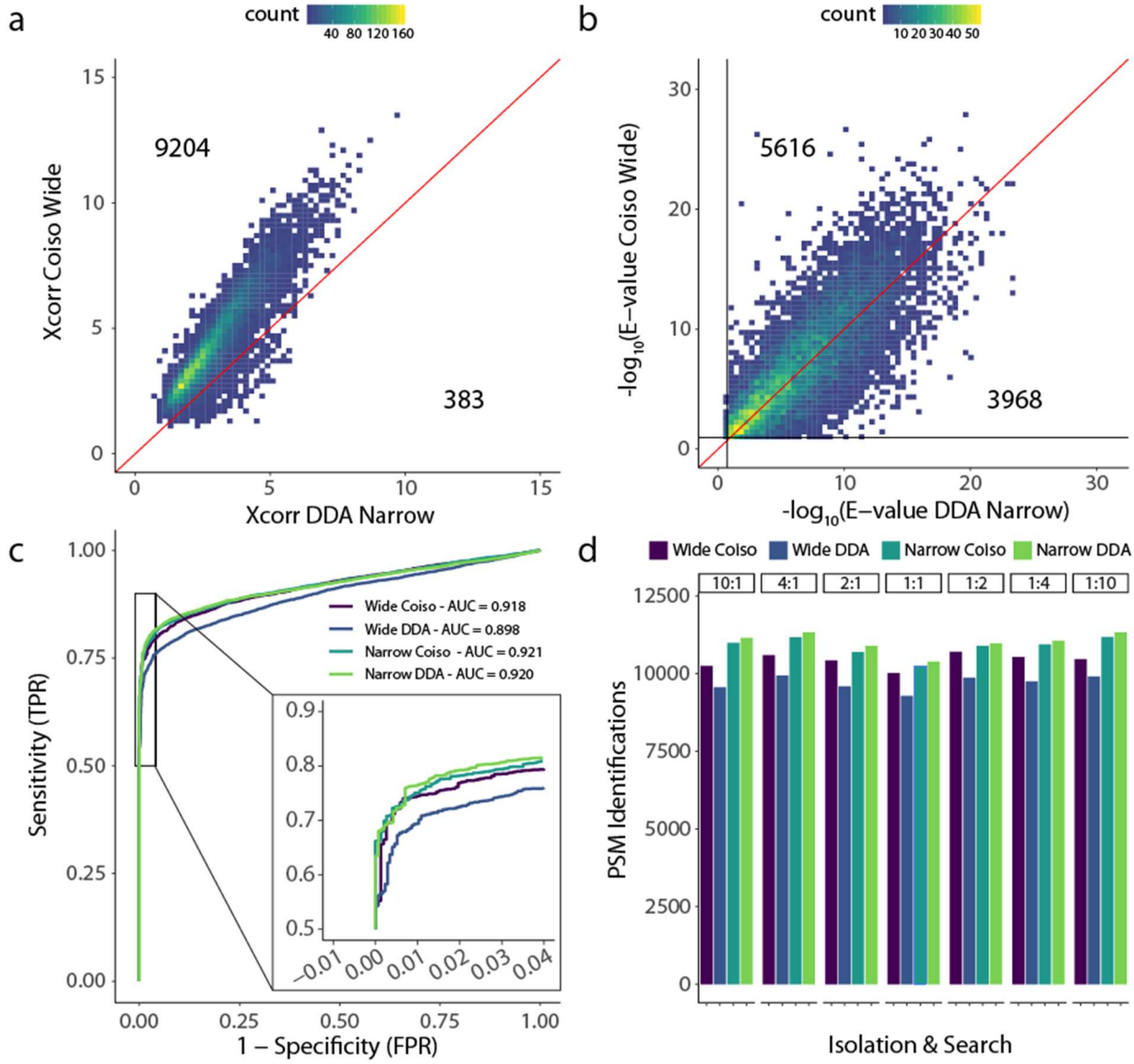


Figure 6.2: Coiso SILAC improves Comet identification metrics compared to traditional DDA. a) Density-binned scatter plot comparison of Xcorr from correctly coisolated wide window scan searched with Coiso algorithm and narrow window scan searched with DDA parameters for matching targeted precursor for 1:1 *S. cerevisiae* SILAC mixture. **b)** Same as (**a**) for Comet E-values. **c)** ROC plot for all combinations: correctly coisolated Coiso scans and narrow window scans searched with Coiso or traditional DDA Comet search parameters for 1:1 *S. cerevisiae* SILAC mixture. Inlet zooms over 0.01 cut-off. **d)** PSM identifications at 1% FDR based on Comet E-values for PSMs with one lysine across the scan acquisition:Comet search parameter combinations in **(c)** for seven *S. cerevisiae* SILAC ratio proteome mixtures.

6.4.4 COISOLATION SCORE FOR IDENTIFYING SCANS THAT COISOLATE SILAC PAIRS

While our coisolation approach generally improves PSM E-values, E-values alone are not strong predictors of whether the SILAC peptide pair is coisolated. For example, Wide window

isolations that exclude one of the SILAC features can generate a PSM E-value that passes the PSM-level 1% FDR filter (Figure 6.1c 2nd panel: right isolation). Therefore, we sought to generate a coisolation score that could prioritize Wide window MS/MS that coisolate the SILAC pair.

For SILAC pair isolation, y-ion spectral representation from the non-targeted SILAC precursor must be observed. Thus, to generate the coisolation score, we measure the added Xcorr contributions from the pair's y-ions of the non-targeted SILAC precursor. Then, we compare the PSM's non-targeted y-ion Xcorr for each scan against all candidates within the MS1 ppm tolerance to calculate an expectation score or Coiso E-value (Appendix Supplementary Figure 6.3). Our Coiso E-value should calibrate the non-targeted precursor's Xcorr contribution across scans and standardize our confidence in SILAC pair detection.

To test whether the Coiso E-value can correctly prioritize coisolated SILAC peptide pairs, we leveraged the MS acquisition in Figure 6.1a but removed the narrow window scan. Since both Wide offset MS/MS scans are analyzed for each precursor, Coiso E-values should prioritize coisolated Wide scans over the respective non-coisolated scan. For a 1:1 (Lysine+0:Lysine+8) SILAC *S. cerevisiae* proteome mixture, we should observe improved Coiso E-values for heavy precursors with left offset and light precursors with right offset Wide MS/MS scans. With the 6.5 m/z Wide windows, coisolation can occur for all charge states when one lysine is present and charges z=4-6 when two lysines are present. All other lysine-charge combinations will not coisolate for both offset MS/MS.

The standard Comet E-value can prioritize coisolated scans from non-coisolated scans based on higher $-\log_{10}$(E-values) (Appendix E Supplementary Figure 6.4). However, without both Wide offset scans for the same precursor, it is difficult to tell from Comet E-value alone

whether the SILAC pair is coisolated. Also, we observe high -$\log_{10}$(E-values) for many PSMs with two lysines and z=2-3, which indicates high confidence PSMs without SILAC pair coisolation.

Alternatively, when exploring the same data with Coiso E-value, we observed clear separation of coisolated and non-coisolated SILAC pair spectra. The data demonstrates the anticipated Coiso E-value relationship relative to precursor charge state and number of lysines (Figure 6.3a). The number of lysine-charge combinations that should not have coisolation demonstrate Coiso E-values similar to that of decoys (z=2-3 & 2 Lys; z=2-6 & 3 Lys). Combinations that should have only one offset produce coisolation clearly separate between non-coisolated and coisolated scans with the expected heavy or light peptide assignments (z=2-3 & 1 Lys; z=3-6 & 2 Lys). Lastly, combinations that should coisolate with both offset MS/MS demonstrate Coiso E-values along the diagonal (z=4-6 & 1 Lys) (Figure 6.3a). Thus, our Coiso E-value can serve as a useful predictor for correctly coisolated SILAC peptide pairs.

Since the non-targeted precursor's y-ion signal is essential to calculate SILAC ratios, we can extend the Coiso E-value to prioritize PSMs for quantification. In a 1:1 SILAC-labeled *S. cerevisiae* dataset, the target PSMs demonstrate a bimodal distribution with respect to -$\log_{10}$(Coiso E-value). Target PSMs either overlap with the decoy distribution or shift towards high values (Appendix E Supplementary Figure 6.5). High -$\log_{10}$(Coiso E-values) indicate robust pair y-ion signals, likely enabling SILAC ratio calculation. Similarly to using Comet E-values for PSM FDR-control, a FDR cut-off for Coiso E-values can be used to filter PSMs for high-quality quantifications. Thus, our Coiso E-value enables unprecedented FDR-control of SILAC quantifications, improving the reliability of quantification in SILAC experiments.

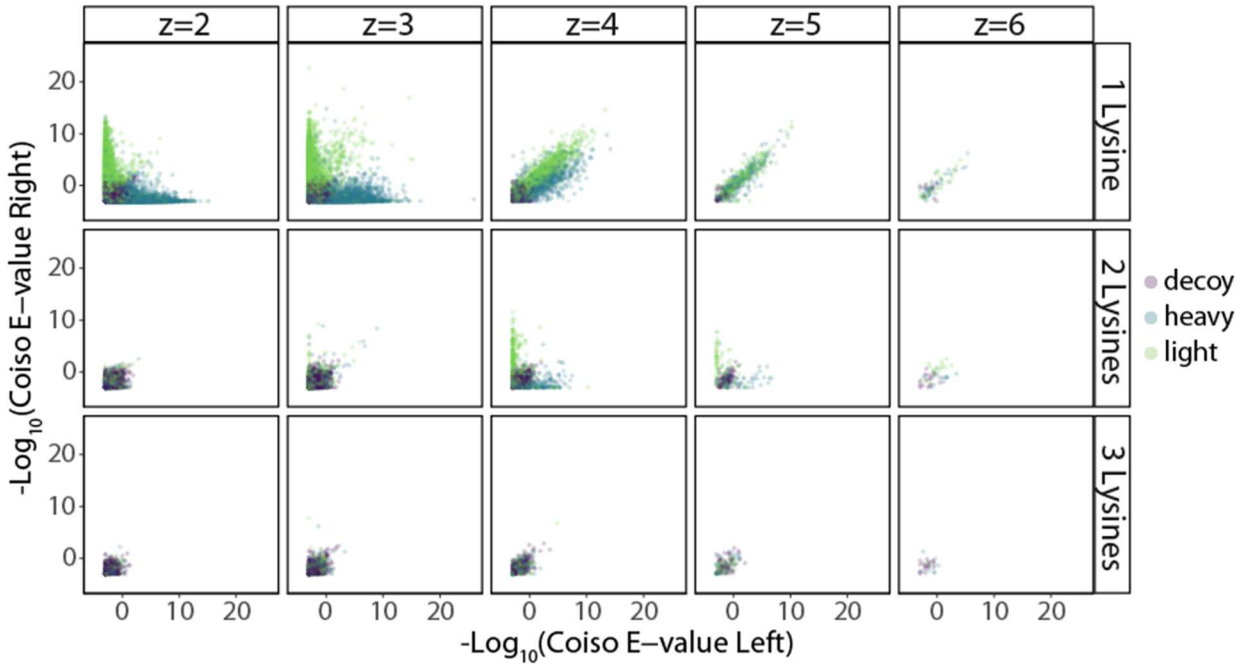


Figure 6.3: Coiso E-value for prioritizing SILAC peptide pair coisolation and predictive quantifiability. Scatterplot for Coiso E-value of the 1:1 *S. cerevisiae* SILAC proteome mixture faceted by PSM charge state and number of lysines for matching left and right offset Coiso scans. PSM assignment either heavy (blue), light (green), or decoy (purple) based on correctly Coiso scans PSM sequence.

6.4.5 MS/MS QUANTIFICATION OF SILAC PEPTIDE PAIRS

To assess quantification performance, we analyzed seven SILAC *S. cerevisiae* proteome mixtures (Lys0:Lys8) using Wide MS/MS with both offsets. PSMs were filtered for the offset with the higher -log$_{10}$(Coiso E-value) and peptides with one lysine (See Methods). For each PSM, we calculate SILAC ratios from precursor MS1 and paired y-ion MS/MS signals, which are compared for precision and accuracy. MS/MS-based quantification is calculated from the most abundant y-ion pair signals. The number of quantified PSMs by MS/MS and its overlap with quantified MS1 vary based on the top y-ion pairs used for quantification (Appendix E Supplementary Figure 6.6-6.7). Here, we highlight the Top3 and Top4 most abundant y-ions for MS/MS quantification for comparison with MS1 quantification. We concluded filters for Top3

and Top4 y-ion pairs exemplified the best balance between number of PSMs quantified and quantification accuracy and precision.

Since our coisolation method enriches SILAC pair signals, we expect more PSMs with calculated SILAC ratios from observed y-ion pairs compared to precursor pairs. In a 1:1 Lys0:Lys8 *S. cerevisiae* proteome mixture, we observe a 14% increase in quantified PSMs by Top3 y-ion pair MS/MS compared to MS1 (Figure 6.4a**)**. The number of quantified PSMs by MS/MS improves the greater the SILAC ratio deviates from 1:1. We observe a 15% improvement with 1:2 and 2:1 SILAC ratios, 23% improvement with 1:4 and 4:1 SILAC ratios, and 40-51% improvement at 1:10 and 10:1 SILAC ratios for top3 fragment ions (Top4: 1:1 at 11%, 1:2&2:1 at 12%, 1:4&4:1 at 15%, and 1:10&10:1 at 18-30% improvements). For the Top3 and Top4 y-ion pairs, MS/MS quantified PSMs largely overlap with MS1 counterparts (Figure 6.4b-c). Also, often both MS1 and MS/MS quantifications are available for a given PSM, which could be leveraged together for SILAC ratio calculations.

For each PSM, MS/MS-based SILAC ratios were calculated by either sum, median, or linear regression across a number of TopX y-ion pairs (Appendix E Supplementary Figure 6.8). For Top3 and Top4 y-ion pairs, median-based MS/MS quantifications demonstrated narrower distributions than MS1 quantifications (Figure 6.4d), suggesting improved precision. For 1:1, 1:2, and 2:1 Lys0:Lys8 SILAC mixtures, MS/MS quantification distributions accurately center the expected SILAC ratios. However, SILAC ratios further from 1:1 demonstrated mild (1:4 and 4:1) to moderate (1:10 and 10:1) ratio compression for MS/MS quantifications. Ratio compression did not correlate with MS/MS fragment ion signal, and the relationship between SILAC ratios and MS intensities were similar between MS1 and MS/MS quantification (Appendix E Supplementary Figure 6.9).

Due to MS1 complexity, precursor signals are susceptible to poor signal-to-noise resulting in outlier quantifications. We anticipate that our coisolation method's gas phase enrichment should improve signal-to-noise resulting in fewer outlier quantifications. To compare, we define an outlier by the percentage of quantified PSMs outside a two-fold difference from the expected SILAC ratio. The percentage of outliers varies by MS/MS quantification approach and top y-ion pair filters (Appendix E Supplementary Figure 6.10). For SILAC ratios 4:1 to 1:4 (Lys0:Lys8), Top3 and Top4 median-based MS/MS quantifications had substantially less outliers compared to MS1 quantifications (Figure 6.4e). However, the percentage of two-fold outliers for 10:1 and 1:10 SILAC ratios was greater for MS/MS qualifications. This is likely driven by ratio compression shifting quantifications from the expected SILAC ratio. At these extreme ratios, both MS1 and MS/MS quantifications demonstrated ~9% two-fold outliers from distribution medians, suggesting similar precision.

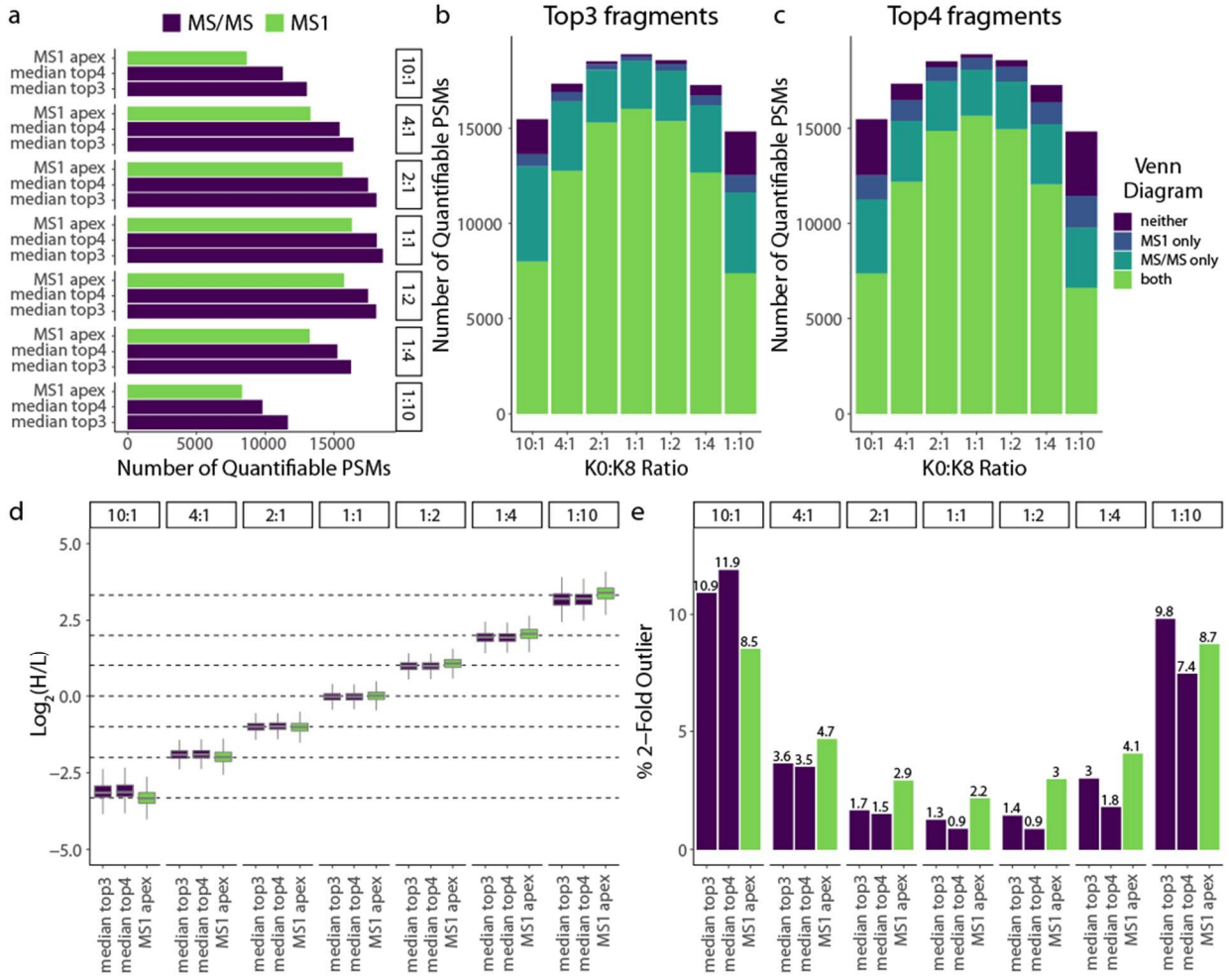


Figure 6.4: Coiso SILAC enables more quantifications with improved precision and less outliers. a) Barplot of the number of quantifiable PSMs for MS1 (green) and the two best Coiso MS/MS quantification methods (purple). Plots consider the same set of PSMs for each of the SILAC *S.cerevisiae* proteome mixtures (K0:K8 ratio respectively). **b-c)** Venn diagram for the same samples as in (**a**) depicted as a stacked barplot outlining the overlap between MS1 quantifications and MS/MS quantifications considering top3 (**b**) or top4 (**c**) quantifiable y-ion fragment MS/MS-based quantification filter. Venn diagram designations are the following: both : quantifiable MS1 and MS/MS (green); neither: not quantifiable in MS1 and MS/MS (purple), MS1 only (dark blue), MS/MS only (light blue). **d)** Box plots for peptide-spectral matches for the same samples as in (**a**). MS/MS-based quantification methods (purple) filtered for top3 or top4 quantifiable paired y-ion fragments. MS1-matched SILAC features from Dinosaur (green) using either apex or sum-based quantification. Box plots represent the distribution from all quantifiable PSMs from (**a**). In the box plot, the horizontal line represents the median, box designates the IQR, and the whiskers indicate 1.5 x IQR from the box ends. **e)** Bar plots of percentage peptide-spectral matches that are two-fold outliers from the expected value across SILAC *S.cerevisiae* proteome mixtures (K0:K8 respectively) as in (**a**). MS/MS quantification methods (purple) filtered by top3 or top4 quantifiable paired y-ion fragments and Dinosaur MS1-derived quantifications (green) are represented here.

6.4.6 QUANTIFYING SILAC PEPTIDE PAIRS WITH OVERLAPPING ISOTOPIC DISTRIBUTIONS

A strategy to improve the SILAC coisolation method is to reduce the MS/MS spectral complexity by decreasing the size of the MS isolation window. This can be achieved by using

SILAC labels with a smaller delta mass. Particularly, SILAC labels separated by 2Da can improve DDA sampling because the SILAC precursor pairs will have overlapping isotopic distributions. In this scenario, the SILAC pairs are detected as a single feature and are subjected to dynamic exclusion from a single MS/MS event. However, MS1-based quantification for SILAC mixtures with overlapping isotopic distributions can be challenging to deconvolute. The coisolation method could enable improved SILAC quantification due to reduced isotopic overlap of peptide fragment ions in the MS/MS. Thus, we set out to explore the feasibility of the coisolation method for proteomes with SILAC labels separated by a 2Da mass difference.

We generated $^{13}C_6$-lysine (Lys6) and $^{13}C_6$,$^{15}N_2$-lysine (Lys8) SILAC-labeled *S. cerevisiae* proteome mixtures across a dynamic range of SILAC ratios (10:1, 4:1, 2:1, 1:1, 1:2, 1:4, and 1:10 Lys6:Lys8 ratios). We piloted our SILAC coisolation method using 1Da left and right offsets with 5 m/z Wide MS/MS acquisitions. PSMs were assigned using the Comet adapted for SILAC pairs. As above, PSMs were filtered for the offset MS/MS with a higher -$\log_{10}$(Coiso E-value) and one lysine (See Methods).

In Lys6:Lys8 SILAC mixtures, the second light ^{13}C isotope shares the same mass as the heavy monoisotopic peak, aggregating their signals. This overlap continues for the heavier adjacent ^{13}C isotopes which all must be deconvoluted. For precursors and fragments, we deconvolute the light contribution to the heavy intensities using the mathematical approach in Chavez et al.[165]. This approach fits the theoretical isotopic distribution[164] to the observed isotopic spectra to calculate the optimal SILAC (R_{opt}) ratio (Figure 6.5a).

For MS/MS-based quantification, we applied this deconvolution approach to calculate R_{opt} ratios for each paired y-ion fragment. We surmised a SILAC ratio for each PSM as the median R_{opt} among the top3 and topN most abundant y-ion pairs. For MS1-based quantification,

we applied the deconvolution method for overlapping precursor signals, generating a R_{opt} for MS1 apex and a median R_{opt} for the top3 MS1 scans (See Methods).

In 4:1 to 1:4 Lys6:Lys8 and Lys0:Lys8 SILAC mixtures, we observed a similar number of quantified PSMs based on y-ion pairs (Figure 6.5b, Appendix E Supplementary Figure 6.11a). At the most extreme SILAC ratios, more PSMs were quantified by MS/MS in Lys6:Lys8 samples to Lys0:Lys8 samples (Appendix E Supplementary Figure 6.11a). Also, the deconvolution approach resulted in similar numbers of quantified MS1 and MS/MS features in Lys6:Lys8 samples (Figure 6.5b). Surprisingly, the Lys6:Lys8 mixture with the most quantified PSMs was the 1:2 Lys6:Lys8 ratio. The SILAC 1:2 ratio could reflect the greatest "pair" signal for MS/MS triggering, due to the asymmetrical light isotopic contribution boosting the heavy precursor signal.

We observed Lys6:Lys8 MS/MS-based quantifications accurately mapped to the expected SILAC ratios (Figure 6.5c) and demonstrated similar quantification precision to Lys0:Lys8 mixtures (Appendix E Supplementary Figure 6.11b). In 1:4 and 1:10 Lys6:Lys8 SILAC mixtures, we observe mild SILAC ratio compression. Alternatively, the 4:1 and 10:1 SILAC ratio distributions were shifted towards more negative, extreme ratios. This is due to the asymmetrical nature of the Lys6:Lys8 deconvolution. The model likely is calculating extreme SILAC ratios when little to no heavy signal is observed in the isotopic profile. This notion is supported in the 10:1 Lys6:Lys8 sample where there is a clear association between smaller peptide length and more negative, extreme $\log_2$(heavy/light) ratios (Appendix E Supplementary Figure 6.12). This is because shorter peptides on average would have shorter quantified y-ion fragments, which naturally have less light ^{13}C contribution to the heavy isotopes.

We demonstrate the initial feasibility of the coisolation method for MS/MS-based quantification of SILAC labels with overlapping isotopic distributions. Collectively, we demonstrate that our coisolation method offers the community an exciting new alternative to comprehensively identify and quantify SILAC proteome mixtures.

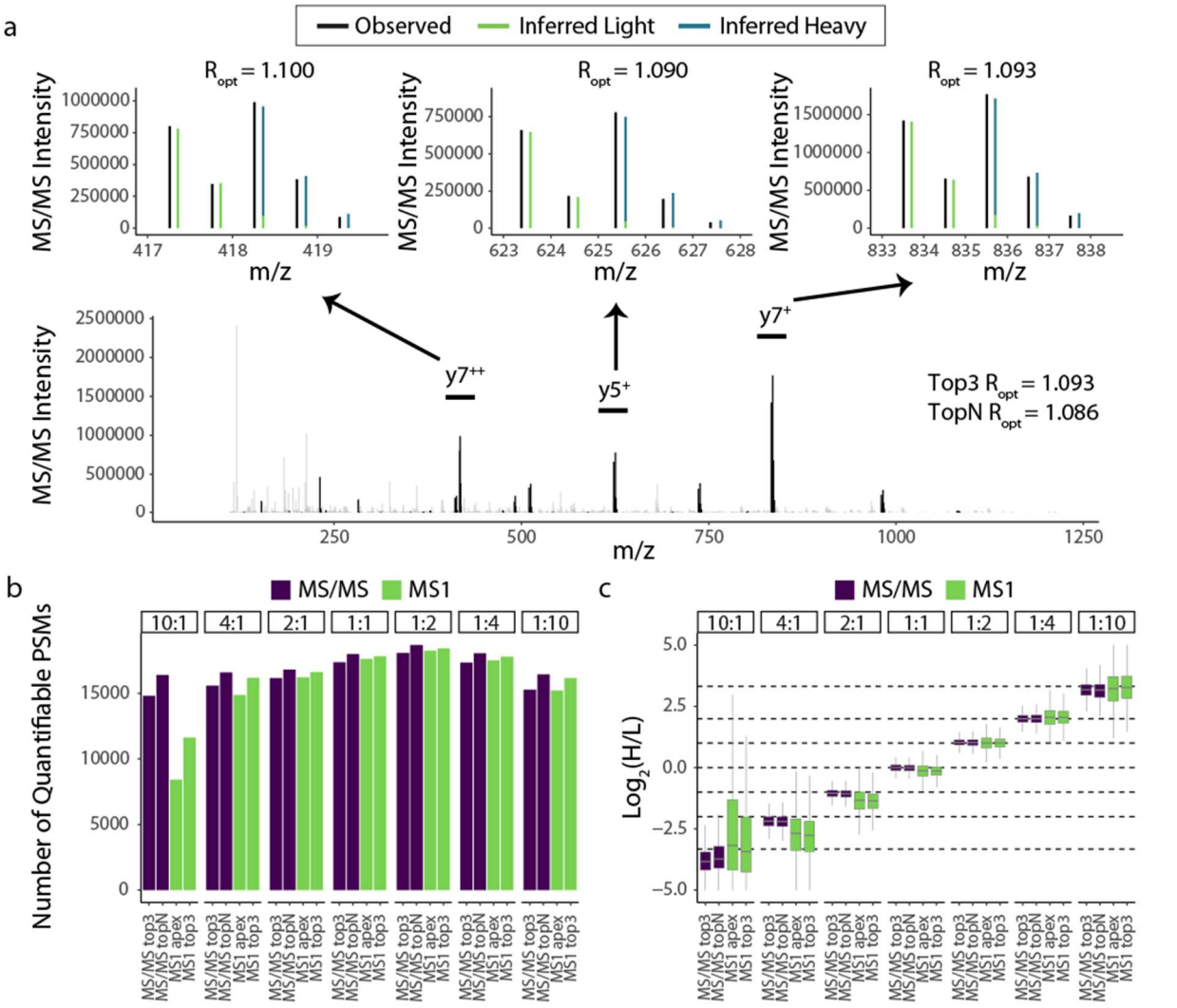


Figure 6.5: Coiso SILAC of K6:K8 mixtures via isotope deconvolution for robust MS/MS quantifications. a) Coiso SILAC MS/MS spectra (bottom panel) of PSA1 ETFPILVEEK light peptide with peptide fragments (black) and unannotated spectra (grey). Top3 peptide paired fragments ($y7^{++}$: left, $y5^{+}$: middle, $y7^{+}$:right) zoomed in over the fragment's observed isotope distribution (black) and the inferred light (green) and inferred heavy (blue) fragment intensity contributions based on the model-defined optimal SILAC ratio ($R_{opt} = Log_2$(heavy/light)). **b)** Bar plot of number of quantifiable PSMs for topN and top3 fragment ion filter requirement (faceted by proteome mixtures defined by SILAC K6:K8 ratios). **c)** Box plots for the distribution of all quantifiable PSMs from (**b**). MS/MS-based quantification methods (purple) filtered for top3 or topN quantifiable paired y-ion fragments. In the box plot, the horizontal line represents the median, box designates the IQR, and the whiskers indicate 1.5 x IQR from the box ends.

6.5 Discussion

Herein, we developed a SILAC coisolation analysis platform equipped with novel MS/MS acquisition schema, database searching strategy, and quantification pipeline. The offset, wide window MS isolations enable gas phase enrichment of the SILAC peptide pair for MS/MS analysis. The adapted Comet search for SILAC pairs leverages the precursor's light and heavy fragments for peptide-spectral matching. The coisolation method improves Comet identification scores and achieves a similar number of PSMs compared to traditional DDA. We generated a Coiso E-value score that prioritizes MS/MS with successfully coisolated peptide pairs and enables FDR-control for SILAC quantification. Also for quantification, our coisolation method offers interference-free MS/MS-based quantification of y-ion pairs. Coisolation quantification outperformed MS1 quantification in the number of quantified PSMs, quantification precision, and percentage of outlier quantifications, however the method demonstrated ratio compression and less accuracy at extreme SILAC ratios. Of note, the coisolation method retains MS1 quantifications with the added benefit of MS/MS quantification.

The two main limitations of this study are related to the current coisolation MS/MS acquisition. First, we perform wide window MS/MS for both offsets to ensure coisolation in at least one of the offsets. Second, wide window coisolations result in high complexity MS/MS for Lys0:Lys8 SILAC mixtures. These limitations likely result in more proteome undersampling and likely fewer PSMs than a traditional DDA, respectively.

To address the complexity, we demonstrate that the coisolarion method can be extended to proteome mixtures using SILAC labels separated by 2Da. In Lys6:Lys8 labeled proteome mixtures, MS/MS quantification greatly outperformed MS1 quantification. However, further optimization of the isotope deconvolution method will be necessary for extreme light peptide-

dominant Lys6:Lys8 mixtures. We anticipate these improvements are feasible by filtering out y-ion R_{opt} calculations with high ratio error.

The intent of the study was to demonstrate the feasibility of the coisolation method when the SILAC peptide pair is successfully coisolated. Now, we can look to explore alternative strategies to decrease spectral complexity and improve proteome sampling. For example, offline HPLC methods can be employed to fractionate SILAC labeled proteomes. Fractionation will decrease the complexity of the wide window scans by decreasing the complexity of the proteome in each MS sample. Also, online ion mobility separations (IMS) could separate SILAC peptide pairs from other pairs by their collisional-cross section (CCS). DIA-SIFT[167], which couples SILAC-DIA with IMS, demonstrated robust precision and accuracy for SILAC MS/MS quantification, suggesting merit if applied to the coisolation method.

With API access on the modern-day mass spectrometers, intelligent MS acquisition strategies enable users to interface with MS acquisition on-the-fly[168]. We could leverage the API to on-the-fly detect MS1 SILAC features and perform a wide MS/MS scan to ensure SILAC pair coisolation. This adjustment could remove the need to acquire both offset MS/MS scans for each precursor. To improve proteome sampling, new dynamic exclusion parameters could ensure MS/MS of a SILAC precursor pair only once. With a “SILAC pair" dynamic exclusion, we could theoretically improve SILAC proteome sampling to the same depth as label-free. Also, sequential isolations of light and heavy SILAC precursors with a MSX approach[144] could minimize the MS/MS spectral complexity and further enrich SILAC pair signals. Deconvolution of the y-ion pair signal from coisolation MSX MS/MS can be deduced from fill times following peptide spectral-matching. Collectively, we anticipate the largest improvements to the

coisolation method can be achieved by intelligent MS acquisitions to increase proteome sampling and reduce MS/MS complexity.

An alternative approach to improve peptide-spectral matching would be to increase the number of SILAC labels within the wide window MS/MS. A 3-plex or 4-plex SILAC labeling schema would increase the peptide's spectral representation, likely improving Xcorrs to a greater extent than other candidate peptides. Intelligent acquisition strategies for detecting 3-plex or 4-plex MS1 features could enable comprehensive sampling and "SILAC plex" dynamic exclusion. With the success of Lys6:Lys8 quantifications, we envision a 4-plex with Lys0:Lys2:Lys6:Lys8 would be feasible. All 3-plex combinations of Lys0:Lys2:Lys4:Lys6:Lys8 except labels with three successive 2Da spacing (e.g. Lys0:Lys2:Lys4, Lys2:Lys4:Lys6, and Lys4:Lys6:Lys8) would be also possible. A more sophisticated model has been developed for labels with 3 or more successive 2Da spacings, suggesting potential feasibility[169]. Use of isotopologues could improve multiplexing capabilities for quantification but likely would not be leveraged in the coisolation Comet searches due to MS/MS mass tolerances.

The coisolation SILAC method will offer unique advantages to MS analysis of SILAC phosphoproteomes. Phosphopeptides with multiple phospho-acceptor amino acids can be isobaric and have different localizations. During peptide-spectral matching, diagnostic fragment ions distinguishing between multiple phosphosite localizations must be observed to localize a phosphosite. Since coisolation MS/MS results in paired y-ions, a phosphosite's diagnostic y-ions will be doubled which could provide confidence in or assign a specific phosphosite. Also, traditional MS1 quantification of SILAC phosphopeptides assumes a single phosphosite localization can be associated with a single MS1 feature. However, coeluting phosphopeptides with different phosphosite localizations can have interfering MS1 signals resulting in inaccurate

SILAC ratios. The coisolation approach avoids this assumption by enabling quantification of multiple coeluting isobaric phosphopeptides in a single MS/MS scan. Each phosphosite's diagnostic paired y-ions can be used to assign accurate SILAC ratios specific to each phosphosite localization.

This proof-of-principle study demonstrates a coisolation MS method that can feasibly identify and quantify SILAC labeled proteomes as SILAC peptide pairs. This approach expands our proteomics toolkit for analyzing SILAC samples and offers exciting new opportunities for future development.

Contributions: Experiments and data analysis were performed by Ian Smith. Jimmy Eng developed the Comet database search engine for searching SILAC peptide pairs with assistance from Ian Smith. Judit Villén, Jimmy Eng, Ricard Rodriguez, and Ian Smith designed the experiments.

Chapter 7. CONCLUSION

7.1 IMPACT OF PRESENTED WORK

7.1.1 FUNCTIONAL INSIGHT INTO PHOSPHORYLATED PROTEOFORMS

In this dissertation, I developed two novel proteomic methods to functionally prioritize protein phosphorylation events that rely on comparing protein thermal stability (Chapter 2) and protein turnover (Chapter 3) to that of the unmodified protein. This work addresses a bottleneck in the field pertaining to the ~97% of the 100,000's of phosphorylation events that have unknown function[34]. Both methods can assay thousands of phosphorylation events for these protein properties in a single experiment, accelerating the functional prioritization of phosphosites for deeper characterization.

Of note, the protein thermal stability and protein turnover assays can be applied to different environmental or genetic perturbations to elucidate functional phosphorylation events whose altered protein thermal stability and protein turnover are condition specific. Also, both assays can be easily extended to other post-translational modifications (e.g., acetylation, ubiquitination) and model systems (human, mouse cell lines).

For phosphorylation's role on protein thermal stability, we found a number of phosphorylation sites that likely alter protein-protein interactions or protein conformational stability. To date, many studies have examined the structural context of phosphorylation events and uncovered many regulatory phosphosites that alter protein conformation and promote crosstalk with other post-translational modifications to induce functional changes[35,170,171]. However, most of these studies associate the location of the phosphorylation event on the protein structure to infer conformational changes, but we have limited experimental data to prioritize

which phosphorylation events likely alter protein structure and how they do so. When mapping destabilizing phosphosites to known structures, we observed that many of these phosphosites occur at protein interfaces, likely disrupting protein-protein interactions. Thus, this work can complement *in silico* structural modeling to inform how single amino acid changes can alter protein structure.

For phosphorylation's role on protein turnover, we explored the flux of pulsed lysine and found many phosphoproteins with differences in protein turnover compared to its unmodified protein. In Chapter 3, we proposed that differences in our experimentally-derived protein turnover metric could not confidently be ascribed to differences in protein degradation or synthesis. However, the kinetic differences of amino acid incorporation between a phosphorylated isoform and its protein could implicate biased phosphorylation towards proteins of a certain protein age, which warrants further exploration. This protein age-biased phosphorylation model was presented to explain phosphorylated proteoforms with faster turnover readouts, whose presence alone contradicts a simplified steady state model that includes protein phosphorylation. In the protein age-biased phosphorylation model, phosphorylated proteoforms with faster turnover could result from an asymmetrical phosphorylation of protein molecules biased toward newly synthesized proteins. Follow-up validation of individual faster turnover phosphosites will be necessary to test our novel hypothesis.

As others have proposed[64,71], phosphorylated proteoforms with different R_{TO} than the unmodified protein can be functionally prioritized. However, our method cannot assign a phosphosite to a specific function. I think the exciting implications of this work are in the new questions raised: Is protein age an important determinant of phosphorylation and how might

protein age influence protein function? Also, is there a relationship between a protein age and a protein's functional state?

7.1.2 NOVEL METHODS TO EXPLORE CONSEQUENCES OF PROTEIN CLEAVAGES

From work in Chapter 2[33] and from Zecha et al.[31], peptide-level readouts for protein thermal stability and protein turnover, respectively, recapitulated natural protein cleavage events and demonstrated functional differences between its resultant cleaved proteoform products. This prompted the work presented in Chapter 4 where we applied protein turnover and protein thermal stability assays to detect host protein cleavage events driven by the protease activity of NSP5 of the SARS-CoV-2 virus.

Some proteomic methods to identify protein cleavages rely on detecting neo-N-terminal peptides[98–100]. However, these methods are limited to neo-N-terminal peptides that are easily detected by MS and thus result in many protein cleavages not identified. Also, the detection of the neo-N-termini peptides does not provide any functional information. Our peptide thermal stability and protein turnover assays applied to detecting protein cleavage events can overcome these limitations. These assays can define a narrow region of the protein cleavage event without requiring the detection of a precise cleaved peptide. Also, differing protein turnover and thermal stability of the cleaved proteoforms could indicate functional changes to a protein's degradation and stability, sometimes due to the action of other proteins (e.g., when a protein loses its interaction with cleaved host NSP5 protein substrate).

While the protein turnover and thermal stability assays can detect protein cleavage events, the protein cleavage field could progress by using the approaches in Chapter 4 in tandem with N-terminomics approaches. They are complementary methods that can bridge the gap between large-scale protein cleavage detection, where N-terminomics excels, and large-scale

functional cleavage assessment, where our methods excel. Together, these methods can begin to identify and prioritize host NSP5 substrates with likely loss-of-function or altered function. These NSP5 substrates could be valuable candidates for follow-up functional validation to determine whether the cleavage event might mediate or modulate SARS-CoV-2 viral infection and propagation.

7.1.3 PROTEOMIC TECHNOLOGIES TO ADDRESS FUNCTIONAL IMPACT OF MUTATIONS

In Chapter 5, we developed a peptide barcoding approach to facilitate the detection and functional characterization of thousands of missense protein variants in a single experiment. Limitations of bottom-up proteomics makes variant detection across a large library of missense protein variants inefficient. For instance, in a missense library, variant proteoforms are low abundance and dominated by wildtype proteoform peptides. Also, variant-specific peptides have different ionization efficiencies making abundance comparisons between variants difficult. To address this limitation, we can tag each protein variant with a unique molecular identifier, or peptide barcode, and use the MS detection of the peptide barcode to represent its tagged variant. Molecular phenotyping assays, such as protein thermal stability (Chapter 2), could be applied to barcoded protein variant libraries and use peptide barcode MS measurements to compare protein variants.

In Chapter 5, we presented initial feasibility experiments of the peptide barcoding approach, which demonstrated that: (1) peptide barcodes can accurately reflect biological properties of their tagged proteins, (2) peptide barcodes can be enriched to simplify sample complexity and enhance the detectability of peptide barcodes, and (3) a majority of peptide barcodes do not alter the stability of the tagged protein. This work provides a foundation for

further development and implementation of the peptide barcoding approach to characterize the functional impact of mutations at scale. However, substantial work is still required to bring it to fruition.

There are far-reaching implications to the successful implementation of the peptide barcoding approach. First, peptide barcoding will be one of the first works of its kind to identify and functionally assay protein variants directly by mass spectrometry. Second, a single optimized peptide barcode library and single MS method can be applied to characterize any protein variant library. Third, a single peptide barcoded protein variant library can be easily extended to 10's of biochemical or biological assays for function, enabling extensive characterization at single amino acid resolution. While a majority of the functional selections are generalizable, other molecular selections can be easily adapted to a protein of interest (e.g. affinity pulldown of a protein's interacting partner). Fourth, by applying the functional selection and performing the measurement at the protein-level, we have no constraints for assaying protein libraries. To the contrary, the field standard deep mutational scanning approach[124] requires labor intensive tuning of a cellular phenotype to conduct these functional assays. Last, our MS assays using peptide barcodes could provide experimental functional evidence to resolve variants of unknown significance (VUS) in the clinic.

7.1.4 A NOVEL STRATEGY FOR PEPTIDE-SPECTRAL MATCHING

For the protein thermal stability experiment in Chapter 2 and protein turnover assays I developed in Chapters 3 and 4, we rely on the Stable Isotope Labeling with Amino acids in Cell culture (SILAC)[172] for quantification. Generally, duplex SILAC labeling of the proteome increases the complexity of the peptide sample by 2-fold, as a set of redundant peptide identities with different labeling schemes. As a consequence, MS analysis of SILAC samples identifies

fewer unique peptides and covers fewer proteins compared to samples with no labeling. Unfortunately, no increases to MS acquisition speed or improvements to analysis have adequately compensated for the somehow redundant complexity.

To improve the performance of our SILAC-based assays, we developed a novel MS acquisition method and data analysis strategy/pipeline to identify and quantify SILAC labeled peptides. Our approach, presented in Chapter 6, consists of a data-dependent wide window isolation to selectively capture SILAC peptide pairs for joined fragmentation and MS/MS analysis. We modified the Comet[59,156] protein sequence database search algorithm to match paired y-ions from MS/MS for peptide-spectral matching. Our coisolation peptide pair method improved PSM scoring metrics and improved precision for MS/MS quantification.

This coisolation SILAC peptide pair method provides a unique perspective to the traditional MS acquisition and database searching approach. Traditional protein sequence database searching strategies generate a single peptide assignment per MS/MS scan. Alternatively, our approach performs PSM identification for SILAC peptide pairs, whose paired y-ions and shared b-ion spectral features generally improve our confidence in PSM assignments. Also, compared to the traditional analysis of SILAC proteomes involving quantification on the MS1 signals, our method generates quantitative information at both the MS1 and MS/MS levels, boosting our confidence for quantification. Lastly, using our novel Coiso E-value metric, we perform unprecedented FDR-control for SILAC quantification, ensuring meaningful quantifications for downstream analysis.

The method presented in Chapter 6 also provides a foundation for further expansions. When applied to SILAC labeled phosphopeptides, SILAC paired y-ions increase the number of diagnostic fragment ions for phosphosite localization compared to single precursor MS/MS.

Also, coisolation MS/MS analysis of co-eluting, isobaric SILAC phosphopeptides could enable SILAC ratio calculations for multiple phosphorylation localizations in a single MS/MS scan. When coupling our approach to intelligent MS acquisition approaches[168], we could identify and dynamically exclude peptide pairs after a single MS/MS, thus improving proteome sampling to the same depth as label-free. Additionally, our method is amenable to SILAC multiplexes greater than two and amenable to quantification of SILAC labels separated by as little as 2 Da. Lastly, we could develop novel chemical labeling approaches that keep pairs isobaric in MS1 and paired in MS/MS. This would enable similar complexity to label-free approaches while maximizing the identification and quantification benefits of our coisolation method.

We anticipate the coisolation peptide pair method will have general applicability to all SILAC labeled proteomic experiments. When applying SILAC labeling to biological samples, we expect our method (with some optimization in MS acquisition) to improve the accuracy and precision of quantifications, which can translate to reliable biological findings in downstream analysis.

7.2 LOOKING FORWARD

7.2.1 SHIFTING FROM PROTEIN-CENTRIC TO PROTEOFORM-CENTRIC

Generally in proteomics, we analyze data from a "protein-centric approach". The origin of this perspective likely derives from the central dogma as the final product of the genetic code. Thus, when analyzing mass spectrometry data, we often consolidate all proteoforms to a single protein readout. Aggregating the abundance of all proteoforms to one protein can be detrimental for identifying important protein signatures in our experiments. For instance, protein abundance may remain constant across two conditions, however differences in proteoform abundance or

stoichiometry among proteoforms could indicate a functional change. Additionally, if we observe a protein abundance change, one cannot interpret whether a modified proteoform is the source of that change.

In this thesis, we have improved upon traditional MS protein abundance readouts, migrating from a protein-centric perspective towards a proteoform-centric perspective. With a proteoform-centric perspective, we remove protein-level consolidation of peptides which relies on an oversimplified approach to protein inference. Thus, we retain proteoform information by performing our proteome analysis at the peptide-level. Peptide-level analysis allows for the comparison of different proteoforms using peptides that are proteoform specific. Importantly, we extend beyond the molecular phenotype of abundance because differences in peptide ionization efficiency make direct abundance comparisons between proteoforms difficult. Therefore, we leveraged the molecular phenotypes of protein thermal stability and turnover to compare proteoforms. Using these approaches, we perform a functional selection at the protein-level and maintain the proteoform specific thermal stability and turnover at the peptide-level. The readouts are unaffected by peptide ionization efficiencies and are directly comparable between different proteoforms. We can use differences in thermal stability and turnover between proteoforms as indicators for functional differences between them. Thus, from a proteoform-centric perspective, we can begin to expose the underlying functional complexity of proteoforms that is largely ignored from a protein-level perspective.

In the near future, I foresee the next logical step is to associate MS assay prioritized proteoforms to specific annotated functional roles. To do so, we must implement many broadly applicable functional MS experiments to assay a range of functional proxies at scale. First steps, including this work, focusing on proteoform functional differences during steady state

conditions. However, proteins and proteoforms are dynamic functional molecules that can change structure, interactions, stoichiometries, and subcellular localizations in different cellular conditions. Thus, our next steps will require extending our functional selections to different cellular contexts to refine our proteoform atlas with annotated functions. Lastly, the functional comparisons thus far have been done in isolation between proteoforms of a single protein-coding gene. The long-term objective will be to explore the interplay between proteoforms across protein-coding genes and how the proteoform interactions underlie cellular phenotypes. In the next section, I will outline exciting MS methods and future directions that could expand the repertoire of functional assays to apply to proteoforms.

I want to note that I have spoken about proteoforms in this context as synonymous with the term "peptidoform". True proteoforms require detection of the full length proteins by MS which is difficult at scale to date. However, the curation of the Blood Proteoform Atlas[173] highlights that advances in top-down approaches could address this bottleneck in the future. The goals and objectives outlined here to encode functional context into our MS assays and migrating away from the "protein-centric approach" will only benefit the community when our technology enables us to reliably measure full length proteoforms at scale.

7.2.2 SCALING MOLECULAR PHENOTYPING ASSAYS FOR PROTEOFORM FUNCTION FURTHER AND ENABLE PREDICTION

Looking forward, I would like to present a number of functional proxies that can be explored for proteoforms that are amenable to analysis by MS. The essential facet for a useful MS approach to assay proteoforms is that the functional selection must be performed on the protein-level *en masse*, and upon proteome digestion the readouts can be preserved at the peptide-level. Selection encoded peptide-level readouts, unique to specific peptidoforms, enable

functional comparisons between modified proteoforms. As mentioned prior, these biochemical readouts are directly comparable between proteoforms, which ensure all observed peptides have similar assay accuracy independent of potential differences in digestion efficiency and MS ionization.

In this thesis, I have applied functional assays of protein turnover and thermal stability, also successfully implemented by others[31,33,38,64,71,72]. However, many other functional protein properties are available for proteoform comparisons by MS. For instance, phosphorylation and alternative spliced proteoforms have been assayed for their role in complex formation using size exclusion chromatography coupled with MS[71,174]. Also, many proteoforms can specifically participate in the RNA interactome, and mass spectrometry assays such as OOPS[175] and quantitative RNA interactome capture (iRIQ)[176] could be leveraged to identify proteoforms with altered RNA interactome participation, with the later study already demonstrating success for phosphorylated proteoforms. One could identify modified proteoforms that mediate specific protein-protein interactions by coupling mass spectrometry with affinity purification (AP-MS)[177], polysome proteome profiling[45], or proximity labeling techniques[178]. Additionally, a proteoform's subcellular localization can define its function, with many protein modifications inducing subcellular translocation. We can implement several subcellular localization separation[179,180] or labeling strategies[178] coupled with MS to identify proteoforms that have altered subcellular residences. Additionally, many of these methods can be applied in combination with environmental, drug, and small molecule perturbations to expand the cellular states and reach to proteoforms that have condition-specific functions.

Hopefully, these functional assays can facilitate the curation of a functional proteoform atlas. This atlas would include detailed annotation at single amino acid resolution provided by a

wide diversity of functional assays and will span missense mutation, PTM, and protein cleavage proteoforms. While this is a lofty goal, a functional proteoform atlas will be a valuable resource for the prediction of structure and function relationships. For instance, deep mutational scanning[124] approaches have greatly benefited from building generalized models for functional variant prediction[181,182]. Our MS functional data could complement DMS data and help improve the power of our current models for variant interpretation. Also, the wealth of MS experimental data has already been able to predict a functional score of phosphorylation sites in humans[68], and the addition of our functional MS data could only improve the accuracy of the functional score. Together, functional MS assays and the development of functional proteoform atlas could provide the community with the important tools and resources necessary to uncover functional complexity of proteins and the modified proteome.

7.2.3 CLOSING REMARKS

In summary, I have presented a collection of high-throughput MS methods to assay the functions of phosphorylated, cleaved, and missense variant proteoforms. I foresee the proteomics community has benefited and will continue to improve by exploring informative molecular phenotypes at the peptide-level to further understand protein and proteoform functions in the cell. I hope that my thesis work is part of the beginning of harnessing these methods' exciting potential for exploring the next dimension of proteomics: Functional Proteoformics.

BIBLIOGRAPHY

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).
3. Bludau, I. & Aebersold, R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat. Rev. Mol. Cell Biol.* **21**, 327–340 (2020).
4. Smith, L. M. & Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).
5. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
6. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
7. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).
8. Chait, B. T. Mass Spectrometry: Bottom-Up or Top-Down? *Science* (2006) doi:10.1126/science.1133987.
9. Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
10. Liu, Y. A peptidoform based proteomic strategy for studying functions of post-translational modifications. *PROTEOMICS* 2100316 (2021) doi:10.1002/pmic.202100316.
11. Alberts, B. *et al.* Protein Function. *Mol. Biol. Cell 4th Ed.* (2002).
12. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).

13. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Lo Muzio, L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *Int. J. Mol. Med.* **40**, 271–280 (2017).
14. Humphrey, S. J., James, D. E. & Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab. TEM* **26**, 676–687 (2015).
15. Willis, A., Jung, E. J., Wakefield, T. & Chen, X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene* **23**, 2330–2338 (2004).
16. Von Heijne, G. Patterns of Amino Acids near Signal-Sequence Cleavage Sites. *Eur. J. Biochem.* **133**, 17–21 (1983).
17. Martoglio, B. & Dobberstein, B. Signal sequences: more than just greasy peptides. *Trends Cell Biol.* **8**, 410–415 (1998).
18. Hegde, R. S. & Bernstein, H. D. The surprising complexity of signal sequences. *Trends Biochem. Sci.* **31**, 563–571 (2006).
19. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346**, 1255784 (2014).
20. Martinez Molina, D. *et al.* Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science* **341**, 84–87 (2013).
21. Tan, C. S. H. *et al.* Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells. *Science* **359**, 1170–1177 (2018).
22. Mateus, A. *et al.* Thermal proteome profiling in bacteria: probing protein state in vivo. *Mol. Syst. Biol.* **14**, e8242 (2018).
23. Muslin, A. J., Tanner, J. W., Allen, P. M. & Shaw, A. S. Interaction of 14-3-3 with signaling

proteins is mediated by the recognition of phosphoserine. *Cell* **84**, 889–897 (1996).

24. Lin, K., Hwang, P. & Fletterick, R. Distinct phosphorylation signals converge at the catalytic center in glycogen phosphorylases. *Structure* (1997) doi:10.1016/S0969-2126(97)00300-6.
25. Collins, M. O., Yu, L., Campuzano, I., Grant, S. G. N. & Choudhary, J. S. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Mol. Cell. Proteomics MCP* **7**, 1331–1348 (2008).
26. Pratt, J. M. *et al.* Dynamics of Protein Turnover, a Missing Dimension in Proteomics. *Mol. Cell. Proteomics* **1**, 579–591 (2002).
27. Doherty, M. K., Hammond, D. E., Clague, M. J., Gaskell, S. J. & Beynon, R. J. Turnover of the Human Proteome: Determination of Protein Intracellular Stability by Dynamic SILAC. *J. Proteome Res.* **8**, 104–112 (2009).
28. Swaney, D. L. *et al.* Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. *Nat. Methods* **10**, 676–682 (2013).
29. Martin-Perez, M. & Villén, J. Determinants and Regulation of Protein Turnover in Yeast. *Cell Syst.* **5**, 283–294 (2017).
30. Gawron, D., Ndah, E., Gevaert, K. & Van Damme, P. Positional proteomics reveals differences in N-terminal proteoform stability. *Mol. Syst. Biol.* **12**, 858 (2016).
31. Zecha, J. *et al.* Peptide Level Turnover Measurements Enable the Study of Proteoform Dynamics. *Mol. Cell. Proteomics* **17**, 974–992 (2018).
32. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
33. Smith, I. R. *et al.* Identification of phosphosites that alter protein thermal stability. *Nat. Methods* **18**, 760–762 (2021).

34. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).

35. Beltrao, P. *et al.* Systematic Functional Prioritization of Protein Posttranslational Modifications. *Cell* **150**, 413–425 (2012).

36. Studer, R. A. *et al.* Evolution of protein phosphorylation across 18 fungal species. *Science* **354**, 229–232 (2016).

37. Dissmeyer, N. & Schnittger, A. The age of protein kinases. *Methods Mol. Biol. Clifton NJ* **779**, 7–52 (2011).

38. Huang, J. X. *et al.* High throughput discovery of functional protein modifications by Hotspot Thermal Profiling. *Nat. Methods* **16**, 894–901 (2019).

39. Potel, C., Kurzawa, N., Beecher, I., Mateus, A. & Savitski, M. M. Impact of phosphorylation on thermal stability of proteins. *bioRxiv* (2020) doi:10.1101/2020.01.14.903849.

40. Gaetani, M. *et al.* Proteome Integral Solubility Alteration: A High-Throughput Proteomics Assay for Target Deconvolution. *J. Proteome Res.* **18**, 4027–4037 (2019).

41. Finley, D., Bartel, B. & Varshavsky, A. The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis. *Nature* **338**, 394–401 (1989).

42. Kane, P. M. *et al.* Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* **250**, 651–657 (1990).

43. Holt, L. J. *et al.* Global Analysis of Cdk1 Substrate Phosphorylation Sites Provides Insights into Evolution. *Science* **325**, 1682–1686 (2009).

44. Dephoure, N. *et al.* A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci.* **105**, 10762–10767 (2008).

45. Imami, K. *et al.* Phosphorylation of the Ribosomal Protein RPL12/uL11 Affects Translation during Mitosis. *Mol. Cell* **72**, 84-98.e9 (2018).

46. Viéitez, C. *et al.* Towards a systematic map of the functional role of protein phosphorylation. *bioRxiv* (2019) doi:10.1101/872770.

47. Leutert, M., Rodriguez-Mias, R. A., Fukuda, N. K. & Villén, J. R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. *Mol. Syst. Biol.* **15**, e9021 (2019).

48. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

49. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

50. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37**, 825–831 (2009).

51. The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.

52. Groll, M. *et al.* Structure of 20S proteasome from yeast at 2.4Å resolution. *Nature* **386**, 463–471 (1997).

53. Bulfer, S. L., Brunzelle, J. S. & Trievel, R. C. Crystal structure of Saccharomyces cerevisiae Aro8, a putative α-aminoadipate aminotransferase. *Protein Sci.* **22**, 1417–1424 (2013).

54. Jogl, G., Rozovsky, S., McDermott, A. E. & Tong, L. Optimal alignment for enzymatic proton transfer: Structure of the Michaelis complex of triosephosphate isomerase at 1.2-Å resolution. *Proc. Natl. Acad. Sci.* **100**, 50–55 (2003).

55. Garcia-Saez, I., Kozielski, F., Job, D. & Boscheron, C. Structure of GAPDH 3 from S. cerevisiae at 2.0 Å resolution. (2010) doi:10.2210/pdb3PYM/pdb.

56. Didierjean, C. *et al.* Crystal Structure of Two Ternary Complexes of Phosphorylating Glyceraldehyde-3-phosphate Dehydrogenase from Bacillus stearothermophilus with NAD and d-Glyceraldehyde 3-Phosphate. *J. Biol. Chem.* **278**, 12968–12976 (2003).

57. Kim, H., Feil, I. K., Verlinde, C. L. M. J., Petra, P. H. & Hol, W. G. J. Crystal Structure of Glycosomal Glyceraldehyde-3-phosphate Dehydrogenase from Leishmania mexicana: Implications for Structure-Based Drug Design and a New Position for the Inorganic Phosphate Binding Site. *Biochemistry* **34**, 14975–14986 (1995).

58. Wagih, O. *et al.* A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol. Syst. Biol.* **14**, e8430 (2018).

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. 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).

62. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

63. Humphrey, S. J., James, D. E. & Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **26**, 676–687 (2015).

64. Wu, C. *et al.* Global and Site-Specific Effect of Phosphorylation on Protein Turnover. *Dev. Cell* **56**, 111-124.e6 (2021).

65. Martin-Perez, M. & Villén, J. Feasibility of Protein Turnover Studies in Prototroph Saccharomyces cerevisiae Strains. *Anal. Chem.* **87**, 4008–4014 (2015).

66. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787 (2009).

67. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. *R Package Version 2560* R package version 2.56.0 (2020).

68. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).

69. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).

70. Christiano, R. *et al.* A Systematic Protein Turnover Map for Decoding Protein Degradation. *Cell Rep.* **33**, 108378 (2020).

71. Zecha, J. *et al.* Linking post-translational modifications and protein turnover by site-resolved protein turnover profiling. *Nat. Commun.* **13**, 165 (2022).

72. Potel, C. M. *et al.* Impact of phosphorylation on thermal stability of proteins. *Nat. Methods* **18**, 757–759 (2021).

73. 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).

74. Wu, R. *et al.* A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat. Methods* **8**, 677–683 (2011).

75. Christiano, R., Nagaraj, N., Fröhlich, F. & Walther, T. C. Global Proteome Turnover

Analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Rep.* **9**, 1959–1965 (2014).

76. Amanchy, R. *et al.* A curated compendium of phosphorylation motifs. *Nat. Biotechnol.* **25**, 285–286 (2007).
77. Leuenberger, P. *et al.* Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* **355**, (2017).
78. Smith, I. R. *et al.* Identification of phosphosites that alter protein thermal stability. *bioRxiv* 2020.01.14.904300 (2020) doi:10.1101/2020.01.14.904300.
79. Tsang, C. K., Liu, Y., Thomas, J., Zhang, Y. & Zheng, X. F. S. Superoxide dismutase 1 acts as a nuclear transcription factor to regulate oxidative stress resistance. *Nat. Commun.* **5**, 3446 (2014).
80. Parisi, E., Yahya, G., Flores, A. & Aldea, M. Cdc48/p97 segregase is modulated by cyclin-dependent kinase to determine cyclin fate during G1 progression. *EMBO J.* **37**, (2018).
81. Albertyn, J., Hohmann, S., Thevelein, J. M. & Prior, B. A. GPD1, which encodes glycerol-3-phosphate dehydrogenase, is essential for growth under osmotic stress in Saccharomyces cerevisiae, and its expression is regulated by the high-osmolarity glycerol response pathway. *Mol. Cell. Biol.* **14**, 4135–4144 (1994).
82. Jung, S., Marelli, M., Rachubinski, R. A., Goodlett, D. R. & Aitchison, J. D. Dynamic Changes in the Subcellular Distribution of Gpd1p in Response to Cell Stress. *J. Biol. Chem.* **285**, 6739–6749 (2010).
83. Lee, Y. J., Jeschke, G. R., Roelants, F. M., Thorner, J. & Turk, B. E. Reciprocal Phosphorylation of Yeast Glycerol-3-Phosphate Dehydrogenases in Adaptation to Distinct Types of Stress. *Mol. Cell. Biol.* **32**, 4705–4717 (2012).
84. Effelsberg, D., Cruz-Zaragoza, L. D., Tonillo, J., Schliebs, W. & Erdmann, R. Role of

Pex21p for Piggyback Import of Gpd1p and Pnc1p into Peroxisomes of Saccharomyces cerevisiae. *J. Biol. Chem.* **290**, 25333–25342 (2015).

85. McShane, E. *et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167**, 803–815 (2016).
86. Salovska, B. *et al.* Isoform-resolved correlation analysis between mRNA abundance regulation and protein level degradation. *Mol. Syst. Biol.* **16**, e9170 (2020).
87. Stukalov, A. *et al.* Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. 2020.06.17.156455 (2021) doi:10.1101/2020.06.17.156455.
88. Bojkova, D. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* **583**, 469–472 (2020).
89. Selkrig, J. *et al.* SARS-CoV-2 infection remodels the host protein thermal stability landscape. *Mol. Syst. Biol.* **17**, e10188 (2021).
90. Bouhaddou, M. *et al.* The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* **182**, 685-712.e19 (2020).
91. Klann, K. *et al.* Growth Factor Receptor Signaling Inhibition Prevents SARS-CoV-2 Replication. *Mol. Cell* **80**, 164-174.e4 (2020).
92. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
93. Laurent, E. M. N. *et al.* Global BioID-based SARS-CoV-2 proteins proximal interactome unveils novel ties between viral polypeptides and host factors involved in multiple COVID19-associated mechanisms. 2020.08.28.272955 (2020) doi:10.1101/2020.08.28.272955.
94. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*

582, 289–293 (2020).

95. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412 (2020).
96. Roe, M. K., Junod, N. A., Young, A. R., Beachboard, D. C. & Stobart, C. C. Targeting novel structural and functional features of coronavirus protease nsp5 (3CLpro, Mpro) in the age of COVID-19. *J. Gen. Virol.* **102**, (2021).
97. Flynn, J. M. *et al.* Comprehensive fitness landscape of SARS-CoV-2 Mpro reveals insights into viral resistance mechanisms. 2022.01.26.477860 (2022) doi:10.1101/2022.01.26.477860.
98. Pablos, I. *et al.* Mechanistic insights into COVID-19 by global analysis of the SARS-CoV-2 3CLpro substrate degradome. *Cell Rep.* **37**, 109892 (2021).
99. Meyer, B. *et al.* Characterising proteolysis during SARS-CoV-2 infection identifies viral cleavage sites and cellular targets with therapeutic potential. *Nat. Commun.* **12**, 5553 (2021).
100. Kleifeld, O. *et al.* Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat. Biotechnol.* **28**, 281–288 (2010).
101. Kleifeld, O. *et al.* Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat. Protoc.* **6**, 1578–1611 (2011).
102. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
103. 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).

104. McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150–7158 (2014).

105. Eng, J. K. *et al.* A deeper look into Comet – implementation and features. *J. Am. Soc. Mass Spectrom.* **26**, 1865–1874 (2015).

106. 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).

107. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).

108. Miettinen, T. P. *et al.* Thermal proteome profiling of breast cancer cells reveals proteasomal activation by CDK4/6 inhibitor palbociclib. *EMBO J.* **37**, e98359 (2018).

109. Childs, D. *et al.* Nonparametric Analysis of Thermal Proteome Profiles Reveals Novel Drug-binding Proteins. *Mol. Cell. Proteomics MCP* **18**, 2506–2515 (2019).

110. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

111. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).

112. Feller, G. Protein stability and enzyme activity at extreme biological temperatures. *J. Phys. Condens. Matter Inst. Phys. J.* **22**, 323101 (2010).

113. Ward, W. W., Prentice, H. J., Roth, A. F., Cody, C. W. & Reeves, S. C. Spectral Perturbations of the Aequorea Green-Fluorescent Protein. *Photochem. Photobiol.* **35**, 803–

808 (1982).

114. Moustaqil, M. *et al.* SARS-CoV-2 proteases PLpro and 3CLpro cleave IRF3 and critical modulators of inflammatory pathways (NLRP12 and TAB1): implications for disease presentation across species. *Emerg. Microbes Infect.* **10**, 178–195 (2021).

115. Schmidt, N. *et al.* The SARS-CoV-2 RNA–protein interactome in infected human cells. *Nat. Microbiol.* **6**, 339–353 (2021).

116. Kamel, W. *et al.* Global analysis of protein-RNA interactions in SARS-CoV-2-infected cells reveals key regulators of infection. *Mol. Cell* **81**, 2851-2867.e7 (2021).

117. Labeau, A. *et al.* Characterization and functional interrogation of SARS-CoV-2 RNA interactome. *bioRxiv* 2021.03.23.436611 (2021) doi:10.1101/2021.03.23.436611.

118. Green, L., Houck-Loomis, B., Yueh, A. & Goff, S. P. Large ribosomal protein 4 increases efficiency of viral recoding sequences. *J. Virol.* **86**, 8949–8958 (2012).

119. Banerjee, A. K. *et al.* SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. *Cell* **183**, 1325-1339.e21 (2020).

120. Schaffert, N., Hossbach, M., Heintzmann, R., Achsel, T. & Lührmann, R. RNAi knockdown of hPrp31 leads to an accumulation of U4/U6 di-snRNPs in Cajal bodies. *EMBO J.* **23**, 3000–3009 (2004).

121. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

122. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

123. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-985 (2014).

124. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

125. Fayer, S. *et al.* Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. *Am. J. Hum. Genet.* **108**, 2248–2258 (2021).

126. Egloff, P. *et al.* Engineered peptide barcodes for in-depth analyses of binding protein libraries. *Nat. Methods* **16**, 421–428 (2019).

127. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

128. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45**, e102 (2017).

129. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinforma. Oxf. Engl.* **26**, 966–968 (2010).

130. Yen, H.-C. S., Xu, Q., Chou, D. M., Zhao, Z. & Elledge, S. J. Global protein stability profiling in mammalian cells. *Science* **322**, 918–923 (2008).

131. Ong, S.-E. *et al.* Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics*. *Mol. Cell. Proteomics* **1**, 376–386 (2002).

132. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21**, 315–318 (2003).

133. Thompson, A. *et al.* Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* **75**, 4942–4942 (2003).

134. McAlister, G. C. *et al.* Increasing the Multiplexing Capacity of TMTs Using Reporter Ion Isotopologues with Isobaric Masses. *Anal. Chem.* **84**, 7469–7478 (2012).

135. Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399–404 (2020).

136. Hanke, S., Besir, H., Oesterhelt, D. & Mann, M. Absolute SILAC for Accurate Quantitation of Proteins in Complex Mixtures Down to the Attomole Level. *J. Proteome Res.* **7**, 1118–1130 (2008).

137. Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **7**, 952–958 (2006).

138. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).

139. Schwanhäusser, B., Gossen, M., Dittmar, G. & Selbach, M. Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* **9**, 205–209 (2009).

140. Salovska, B. *et al.* Isoform-resolved correlation analysis between mRNA abundance regulation and protein level degradation. *Mol. Syst. Biol.* **16**, e9170 (2020).

141. Pino, L. K., Baeza, J., Lauman, R., Schilling, B. & Garcia, B. A. Improved SILAC Quantification with Data-Independent Acquisition to Investigate Bortezomib-Induced Protein Degradation. *J. Proteome Res.* **20**, 1918–1927 (2021).

142. Salovska, B., Li, W., Di, Y. & Liu, Y. BoxCarmax: A High-Selectivity Data-Independent Acquisition Mass Spectrometry Method for the Analysis of Protein Turnover and Complex Samples. *Anal. Chem.* **93**, 3103–3111 (2021).

143. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat.*

Methods **15**, 440–448 (2018).

144. Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods* **10**, 744–746 (2013).

145. Vincent, C. E. *et al.* Segmentation of precursor mass range using 'tiling' approach increases peptide identifications for MS1-based label-free quantification. *Anal. Chem.* **85**, 2825–2832 (2013).

146. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, e8126 (2018).

147. Graumann, J., Scheltema, R. A., Zhang, Y., Cox, J. & Mann, M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol. Cell. Proteomics MCP* **11**, M111.013185 (2012).

148. Meyer, J. G., Niemi, N. M., Pagliarini, D. J. & Coon, J. J. Quantitative shotgun proteome analysis by direct infusion. *Nat. Methods* **17**, 1222–1228 (2020).

149. Shevchenko, A. *et al.* Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **11**, 1015–1024 (1997).

150. Qin, J., Herring, C. J. & Zhang, X. De novo peptide sequencing in an ion trap mass spectrometer with 18O labeling. *Rapid Commun. Mass Spectrom.* **12**, 209–216 (1998).

151. Goodlett, D. R. *et al.* Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun. Mass Spectrom.* **15**, 1214–1221 (2001).

152. Volchenboum, S. L., Kristjansdottir, K., Wolfgeher, D. & Kron, S. J. Rapid Validation of Mascot Search Results via Stable Isotope Labeling, Pair Picking, and Deconvolution of Fragmentation Patterns*. *Mol. Cell. Proteomics* **8**, 2011–2022 (2009).

153. Heller, M., Mattou, H., Menzel, C. & Yao, X. Trypsin catalyzed 16O-to-18O exchange for comparative proteomics: tandem mass spectrometry comparison using MALDI-TOF, ESI-QTOF, and ESI-ion trap mass spectrometers. *J. Am. Soc. Mass Spectrom.* **14**, 704–718 (2003).

154. Bamberger, C., Pankow, S., Park, S. K. R. & Yates, J. R. Interference-Free Proteome Quantification with MS/MS-based Isobaric Isotopologue Detection. *J. Proteome Res.* **13**, 1494–1501 (2014).

155. Merrill, A. E. *et al.* NeuCode Labels for Relative Protein Quantification*. *Mol. Cell. Proteomics* **13**, 2503–2512 (2014).

156. Eng, J. K. *et al.* A Deeper Look into Comet—Implementation and Features. *J. Am. Soc. Mass Spectrom.* **26**, 1865–1874 (2015).

157. Eng, J. K., Fischer, B., Grossmann, J. & MacCoss, M. J. A Fast SEQUEST Cross Correlation Algorithm. *J. Proteome Res.* **7**, 4598–4602 (2008).

158. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

159. 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).

160. 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).

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

162. Fondrie, W. E. & Noble, W. S. mokapot: Fast and Flexible Semisupervised Learning for

Peptide Detection. *J. Proteome Res.* **20**, 1966–1971 (2021).

163. Bakalarski, C. E. *et al.* The Impact of Peptide Abundance and Dynamic Range on Stable-Isotope-Based Quantitative Proteomic Analyses. *J. Proteome Res.* **7**, 4756–4765 (2008).

164. Yergey, James., Heller, David., Hansen, Gordon., Cotter, R. J. & Fenselau, Catherine. Isotopic distributions in mass spectra of large molecules. *Anal. Chem.* **55**, 353–356 (1983).

165. Chavez, J. D., Keller, A., Mohr, J. P. & Bruce, J. E. Isobaric Quantitative Protein Interaction Reporter Technology for Comparative Interactome Studies. *Anal. Chem.* **92**, 14094–14102 (2020).

166. Bittremieux, W. spectrum_utils: A Python Package for Mass Spectrometry Data Processing and Visualization. *Anal. Chem.* **92**, 659–661 (2020).

167. Haynes, S. E., Majmudar, J. D. & Martin, B. R. DIA-SIFT: A Precursor and Product Ion Filter for Accurate Stable Isotope Data-Independent Acquisition Proteomics. *Anal. Chem.* **90**, 8722–8726 (2018).

168. 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).

169. Chavez, J. D., Keller, A., Wippel, H. H., Mohr, J. P. & Bruce, J. E. Multiplexed Cross-Linking with Isobaric Quantitative Protein Interaction Reporter Technology. *Anal. Chem.* **93**, 16759–16768 (2021).

170. Strumillo, M. J. *et al.* Conserved phosphorylation hotspots in eukaryotic protein domain families. *Nat. Commun.* **10**, 1977 (2019).

171. Bludau, I. *et al.* The structural context of PTMs at a proteome wide scale. *bioRxiv* 2022.02.23.481596 (2022) doi:https://doi.org/10.1101/2022.02.23.481596.

172. Ong, S.-E. *et al.* Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).

173. Melani, R. D. *et al.* The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* **375**, 411–418 (2022).

174. Bludau, I. *et al.* Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat. Commun.* **12**, 3810 (2021).

175. Queiroz, R. M. L. *et al.* Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* **37**, 169–178 (2019).

176. Vieira-Vieira, C. H., Dauksaite, V., Gotthardt, M. & Selbach, M. Proteome-wide quantitative RNA interactome capture (qRIC) identifies phosphorylation sites with regulatory potential in RBM20. 2021.07.12.452044 (2021) doi:10.1101/2021.07.12.452044.

177. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).

178. Rhee, H.-W. *et al.* Proteomic Mapping of Mitochondria in Living Cells via Spatially-Restricted Enzymatic Tagging. *Science* **339**, 1328–1331 (2013).

179. Mulvey, C. M. *et al.* Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat. Protoc.* **12**, 1110–1135 (2017).

180. Martinez-Val, A. *et al.* Spatial-proteomics reveals phospho-signaling dynamics at subcellular resolution. *Nat. Commun.* **12**, 7113 (2021).

181. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* **6**, 116-

124.e3 (2018).

182. Dunham, A. & Beltrao, P. Exploring amino acid functions in a deep mutational landscape. *Mol. Syst. Biol.* **17**, e10305 (2021).

183. Post, H. *et al.* Robust, Sensitive, and Automated Phosphopeptide Enrichment Optimized for Low Sample Amounts Applied to Primary Hippocampal Neurons. *J. Proteome Res.* **16**, 728–737 (2017).

184. Savitski, M. M. *et al.* Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. *J. Proteome Res.* **12**, 3586–3598 (2013).

185. Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: 'the good, the bad and the ugly'. *J. Proteome Res.* **8**, 5347–5355 (2009).

186. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).

187. Hogrebe, A. *et al.* Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.* **9**, 1–13 (2018).

188. Franken, H. *et al.* Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.* **10**, 1567–1593 (2015).

189. Kasari, V. *et al.* A role for the Saccharomyces cerevisiae ABCF protein New1 in translation termination/recycling. *Nucleic Acids Res.* **47**, 8807–8820 (2019).

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

APPENDIX A

Appendix A: Supplementary information for Chapter 2: Identification of phosphosites that alter protein thermal stability

SUPPLEMENTARY DISCUSSION

Experimental workflow for phosphopeptide enrichment and TMT labeling

HTP directly compares melting temperatures between phosphopeptide isoforms and their corresponding proteins. However, critical steps in the experimental workflow were conducted separately for phosphopeptides and proteins and also for the different temperature channels to trace melting curves, possibly introducing technical error. We indeed observe in the Huang et al. dataset low correlation between replicates (mean $R^2 = 0.43$ for proteins, mean $R^2 = 0.22$ for phosphopeptide isoforms), between phosphopeptide isoforms and proteins ($R^2 = 0.18$ using the supplementary data provided; $R^2 = 0.20$ with our reanalysis), and between proteins and unmodified peptides identified in the phospho-enriched samples ($R^2 = 0.18$ with our reanalysis).

One potential source of error could be in the phosphopeptide enrichment, where previous work[183] has shown that the ratio of peptide to TiO_2 or IMAC stationary phase is an important parameter to optimize for successful enrichment, and deviations from the optimal ratio may result in variable peptide binding and/or recovery. In a thermal proteome profiling experiment, the soluble protein fraction recovered from the temperature treatment is very different across temperatures. The HTP method enriches the derived peptides from each temperature separately

but using the same amount of TiO_2 material. Our expectation with this setup would be highly variable phosphopeptide enrichment across the different temperatures. While we could not assess this directly, we observe variable phosphopeptide enrichment efficiencies across replicates ranging from 52 to 78%, with a substantial number of unmodified peptides in the phosphopeptide-enriched samples that appear to have significantly different T_m compared to their reference protein.

Peptides for protein analysis and phosphopeptides were TMT-labeled and desalted separately. While we would expect uniform and high yield labeling reactions across all samples, the workflow does not control for any potential differences. To minimize all these sources of error, we suggest conducting TMT labeling of peptides first and apply phosphopeptide enrichment after the samples from the different temperature channels have been combined. This scheme would minimize the distortion of the melting curves for phosphopeptides[39].

Our workflow introduces an isotopically-labeled 30^{o}C sample to control for any technical differences occurring after temperature treatment, including phosphopeptide enrichment and sample cleanup. As a result, we observe high correlation between replicates, between proteins and unmodified peptides identified in the phospho-enriched sample (R^2=0.92), and between phosphorylation isoforms and proteins (R^2= 0.81).

TMT ratio and T_m compression

HTP applies MS2-based TMT quantification of unmodified proteome and phosphoproteome samples in single shot injections. Many studies[184,185] have reported TMT ratio compression, and attributed this to interfering TMT signals derived from peptides that have been co-isolated for MS/MS fragmentation with the precursor of interest. We expect interference and ratio compression would increase along with sample complexity, and would be high for the experimental settings used in the Huang et al. study, consisting of whole human proteome measurements over 2h (replicates 1 and 2) or 4h (replicates 3-6) LC-MS/MS run times. Indeed, we observe the technical replicates of the longer runs to be better correlated (R^2=0.7) than those for the shorter runs (R^2=0.5). We expect TMT ratio compression will translate into a compression of protein melting curves and melting temperatures trending towards the sample T_m average. The low reproducibility between LC-MS/MS repeat injections (average proteome R^2=0.62, average phosphoproteome R^2=0.36, Figure 2.1b) supports that TMT ratio compression could be a source of error.

Several approaches have been described to mitigate TMT ratio compression including the use of an orthogonal fractionation of the proteome to reduce complexity of each MS injection, filtering the data to only include TMT quantifications derived from isolating a single predominant precursor signal[184], decreasing the isolation window, and/or applying a sequential isolation step in the MS to quantify TMT reporters in an MS3 scan[186].

In our method, we fractionated the proteome into 5 fractions to reduce complexity and obtain deeper coverage. We also chose to use a SILAC-based quantification approach, which has proven to provide more accurate phosphopeptide quantifications than TMT[187] and does not suffer

from ratio compression. We think quantification accuracy becomes more important with peptide or phosphopeptide isoform analysis than with protein analysis, given that the final quantitative measurement aggregates only one to a few measurements.

TPP melting curve fitting

We found that in order to recapitulate the Huang et al. result of the phosphopeptide isoform T_m and protein T_m having the same median T_m all data must be fit together, so sample-to-sample normalization in the TPP package[188] could adjust all phosphorylation and unmodified protein sample consensus curves towards a single representative curve. We believe that this is not the correct way to apply the TPP package because it may obviate global differences in protein and phosphoprotein melting. Separate analyses of proteins and phosphopeptide isoforms for the Huang et al. data indeed revealed existing differences in the average melting curve. Rather, unmodified peptides that are common between enriched and unenriched samples could be used for data normalization[39].

Statistical analysis

The study by Huang et al. encompassed five biological replicates for phosphorylation analysis and six biological replicates for unmodified proteome analysis, and each was analyzed twice on the mass spectrometer. When conducting the t-test to compare the T_m of phosphopeptide isoforms and proteins, the authors treated mass spectrometer repeat injections as independent biological replicates, thus artificially increasing the power of their statistical tests. Additionally, the authors led with the assumption that the T_m variances of the unmodified protein and the phosphoprotein were equal, while we calculated these as 2.4 and 9.5 respectively. Lastly, the authors did not adjust p-values for multiple testing.

With these considerations, we reimplemented the t-test on the Huang et al. dataset by applying median consolidation of reanalyzed samples, assuming unequal variances, and adjusting p-values for multiple testing using the Benjamini-Hochberg method. This resulted in a dramatic decrease in the number of phosphopeptide isoforms that significantly alter protein T_m, from 719 to 20 (<1% of the reported 2,883 high-quality measurements) (Extended Data Figure 2.2).

Using unmodified peptides in phosphopeptide-enriched samples to assess technical variation

We conducted some analysis that uses unmodified peptides that were detected in phosphopeptide-enriched samples to assess technical variation and interrogate if peptide and phosphopeptide samples from the HTP workflow can be compared. We would expect the unmodified peptides to have the same T_m as the bulk unmodified protein and we tested this assumption with two metrics: T_m correlations and MS intensity distributions.

To test if these unmodified peptides are good representations of the protein T_m, we classified the peptides measured in the HTP protein samples in two groups, according to their detection (yes/no) in the phosphopeptide samples. We calculated the correlations between measured T_m's for peptides and the T_m's for the corresponding proteins for the two peptide groups. We found that the two peptide groups had similar correlation to the protein T_m and the correlation was slightly higher for the group of unmodified peptides that were also detected in the phosphopeptide samples ($R^2 = 0.59$ *vs.* $R^2 = 0.42$). This indicates that these unmodified peptides were as good or better reporters of the protein T_m as any other unmodified peptide from the same protein.

Additionally, we compared these unmodified peptides to phosphopeptides with regards to their MS1 intensity and TMT reporter ion intensity in the HTP phosphopeptide samples. We observed that the two groups of peptides had similar intensity distributions for both MS1 precursors and TMT reporter fragment ions, indicating that similar measurement error would be expected for both groups.

These results indicate that the unmodified peptides in the HTP phosphopeptide-enriched samples provide a representative measurement of the protein T_m. This result is further supported by the Dali data where the R_s correlation between the unmodified peptides in the phosphopeptide samples and the protein R_s is $R^2 = 0.92$ (Extended Data Figure 2.1b).

HTP workflow comparison

In response to our commentary, Huang et al. implemented three workflows (EL-HTP, LE-HTP, and LFE-HTP) and provided data to address whether phosphorylation enrichment and TMT labeling order impacted T_m readouts and if TMT ratio compression was distorting T_m readouts. From our global analysis of their supplementary data, we came to four conclusions that validate our concerns.

First, the LFE-HTP (label-fractionate-enrich) protocol is considerably more reproducible than EL-HTP (enrich-label as the original HTP) and LE-HTP (label-enrich) (Appendix A Supplementary Figure 1a). This indicates that TMT labeling prior to phosphopeptide enrichment, fractionation of TMT-labeled peptides and phosphopeptides, and SPS-MS3 acquisition methods (as we suggest above) collectively help improve the reproducibility of the method.

Second, the order of TMT-labeling and phosphopeptide enrichment matters. EL-HTP (as the original HTP) shifts measured T_m's to lower values (less stable) relative to LE-HTP (-1.5^{o}C

averaged shift) and LFE-HTP (-1.8ºC averaged shift) (Appendix A Supplementary Figure 1b). (Note that protein T_m samples and results for LE and EL are the same). As a consequence, phosphosites will be erroneously seen as more destabilizing. We think this could be due to variable phosphopeptide enrichment efficiencies across EL-HTP temperature channels.

Third, the differences in T_m between phosphorylation isoform and protein for the LE-HTP and LFE-HTP protocols do not correlate (Appendix A Supplementary Figure 1c). This suggests that there could be substantial interference in the TMT quantification in the LE-HTP method, which uses single-shot injections, impacting the accuracy of the T_m readouts. Those interference issues were alleviated by peptide pre-fractionation and MS analysis using the SPS-MS3 approach (as we suggest above).

The LFE-HTP dataset demonstrates that phosphorylation alters protein thermal stability much less than published in the HTP study. The correlation between phosphorylated isoform and protein T_m's is R^2=0.58 for LFE-HTP (Appendix A Supplementary Figure 1d) and is comparable to the result from Savitski's lab and to our result in yeast extracts. This correlation is much higher than found for other implementations: original HTP (R^2=0.18), EL-HTP (R^2=0.20), and LE-HTP (R^2=0.23). Thus, considering the HTP improvements, all studies agree and suggest that phosphorylation impacts protein thermal stability to a lesser extent than previously reported.

Considering these four conclusions, the LFE-HTP approach, an experimental design that we suggest, mitigates many of our reliability and reproducibility concerns. For future studies interrogating the effects of phosphorylation on protein stability, we strongly recommend using LFE-HTP, Phospho-TPP, or Dali over the EL-HTP or LE-HTP approaches.

Examples of phosphosites altering thermal stability identified with Dali

We identified stabilizing phosphosites that may play a role in protein translation (Ser38 on RPL12/uL11 and Thr1191 on NEW1) and phosphosites that may modulate the kinetics of key glycolytic enzymes (Thr331 on PGK1 destabilizing, Ser149 on GAPDH stabilizing) (Figure 2.2, Extended Data Figure 2.4). NEW1 phosphorylation at Thr1191 stabilized the protein with a delta $R_s = 1.23$ (Figure 2.2c). NEW1 is a translation factor that binds to the ribosome at a position analogous to eEF3 and fine-tunes the efficiency of translation termination[189]. The identified NEW1 phosphosite fits a CK2 consensus motif, is located within its acidic C-terminal sequence, and is highly conserved. A T1191A mutant has growth defects[46] suggesting that phosphorylation is important for NEW1 function.

For the glycolytic enzymes, we measured the stability for six phosphosites on PGK1, of which only Thr331 showed significantly decreased stability (Extended Data Figure 2.5a). This observation agrees with the predicted stability effects of phosphomimetic substitutions on PGK1[58] (Extended Data Figure 2.5a). We identified a stabilizing phosphosite at Ser149 in the GAPDH isozymes TDH1, TDH2 and TDH3 (Extended Data Figure 2.5b). Ser149 is adjacent to catalytic Cys150 and to the binding sites of glyceraldehyde-3-phosphate (G3P) and inorganic phosphate. Interestingly, Ser149 phosphorylation could occupy the inorganic phosphate binding site (Extended Data Figure 2.5b). Additionally, it has been recently reported that a TDH3 S149A mutant exhibits a growth defect with doxorubicin compared to wild-type and decreases TDH3 activity to a greater extent than a TDH3 knockout[68]. Our results raise the possibility that S149 phosphorylation may increase the stability of apo-GAPDH, the GAPDH-G3P reaction intermediate and aid phosphate transfer by enhancing product release.

Dali results in yeast vs HTP results in human

Our commentary has two separate sections. In the first section, we reanalyze published HTP data collected in a human cell line to identify technical issues. In the second section, we present an alternative method, Dali, which we applied to identify phosphosites altering protein thermal stability in the *S. cerevisiae* proteome. We do not intend to conduct a 1-to-1 comparison of phosphosites and their effects between HTP conducted in human cells and Dali conducted in yeast lysates. These effects may be different due to different biology of the two organisms, to different cell states, or to biochemical differences in conducting the protein melting.

However, it is valid to compare two methods developed for the same application based on the figures of merit of each method (e.g. reproducibility and reliability). We show that Dali has good reproducibility and reliably identifies phosphopeptides significantly altering protein thermal stability, thus validating our method. We wanted to show these metrics in our method because these were our main concerns with the HTP approach. We avoided a comparison between method sensitivities, which could be influenced by the different size and complexity of the two proteomes.

Limitations of Dali

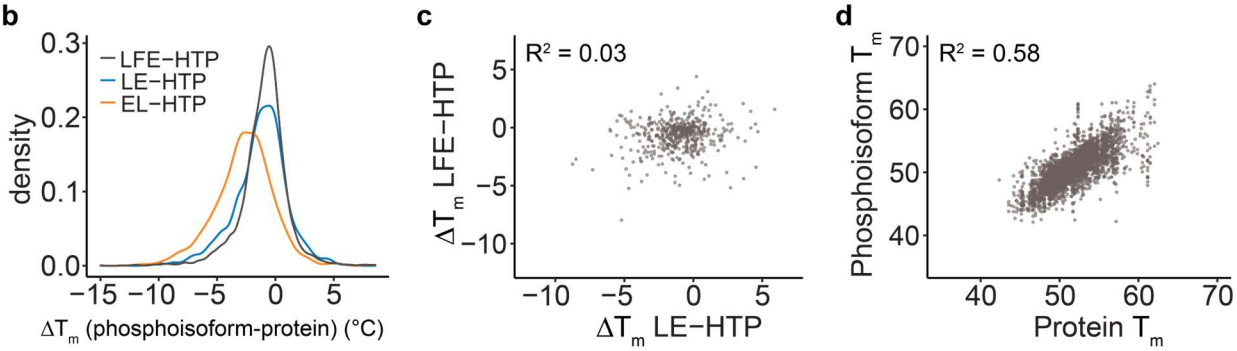
We acknowledge some limitations inherent to the Dali approach we developed. First, each sample combines two independent cell cultures grown in light and heavy SILAC media. It is possible that there may be some differences in protein abundance that may introduce variability across replicate measurements of protein and phosphopeptide isoform R_s values. We attempt to control for this by swapping the isotopic labeling scheme for half of the replicates and partially mitigate the issue by removing the 5% most variable data, which alongside filters out poor-quality quantifications.

A second limitation is on the magnitude of the R_s values, which provide a measure of the relative stability of the protein (or phosphoprotein) to unfold and aggregate under a temperature gradient centered around 50ºC *vs.* 30ºC. Therefore, R_s values are relative to the T_m of each protein, and our ability to detect changes to R_s values depends on the temperature gradient chosen, relative to the protein T_m. In our study, we selected a T_m gradient centered around the median T_m for the *S. cerevisiae* proteome. Thus, we expect to have missed in our study functionally relevant phosphosites in proteins with extreme melting temperatures. In order to capture these, additional experiments would be required where the temperature gradient is globally shifted towards lower or higher temperatures.

SUPPLEMENTARY FIGURES

a

Average pairwise replicate Pearson's correlation	Original HTP	EL-HTP	LE-HTP	LFE-HTP	Dali
phosphopeptide isoform	R^2 = 0.21	R^2 = 0.23	R^2 = 0.22	R^2 = 0.68	R^2 = 0.65
proteins	R^2 = 0.42	R^2 = 0.36		R^2 = 0.73	R^2 = 0.76

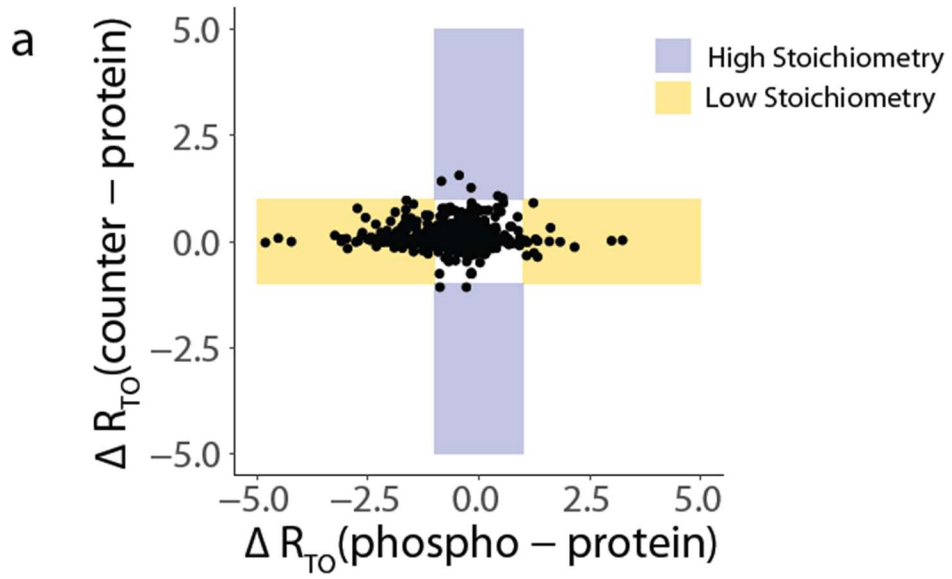


Supplementary Figure 1. Comparison of HTP workflows. a, Mean pairwise Pearson correlations of T_m (HTP) and R_s (Dali) values for phosphoisoforms and proteins between replicates (original HTP n=6 for protein and n=5 for phosphoisoforms; EL-HTP and LE-HTP n=4; LFE-HTP n=3; Dali n=6 biological replicates). **b,** Distribution of T_m differences between phosphoisoforms and unmodified proteins (EL-HTP n=711; LE-HTP n=1,110; LFE-HTP n=4,024 data points). **c,** Scatter plot and Pearson correlation comparing T_m differences measured using the LFE-HTP and LE-HTP workflows (n=478 data points). **d,** Scatter plot and Pearson correlation between T_m of phosphopeptide isoforms and T_m of the corresponding protein for the LFE-HTP workflow (n=4,024 data points).

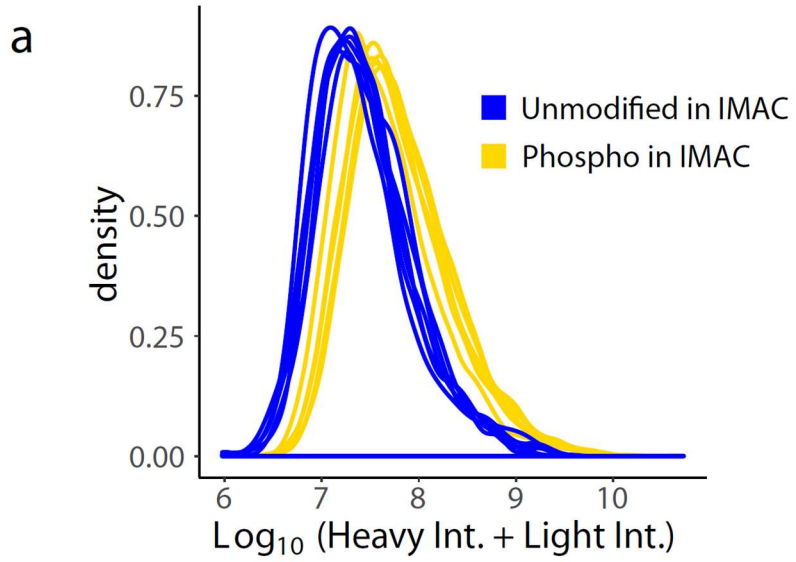
APPENDIX B

Appendix B: Supplementary information for Chapter 3: Kinetic analysis of phosphorylation impact on experimentally-derived protein turnover and protein age-biased phosphorylation

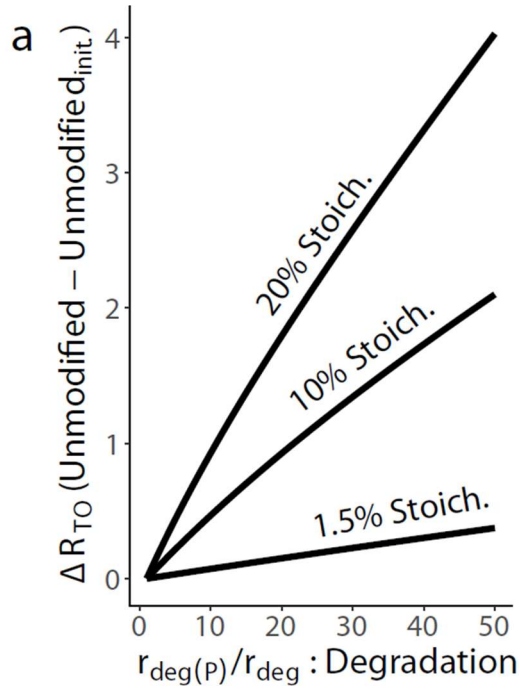
SUPPLEMENTARY FIGURES



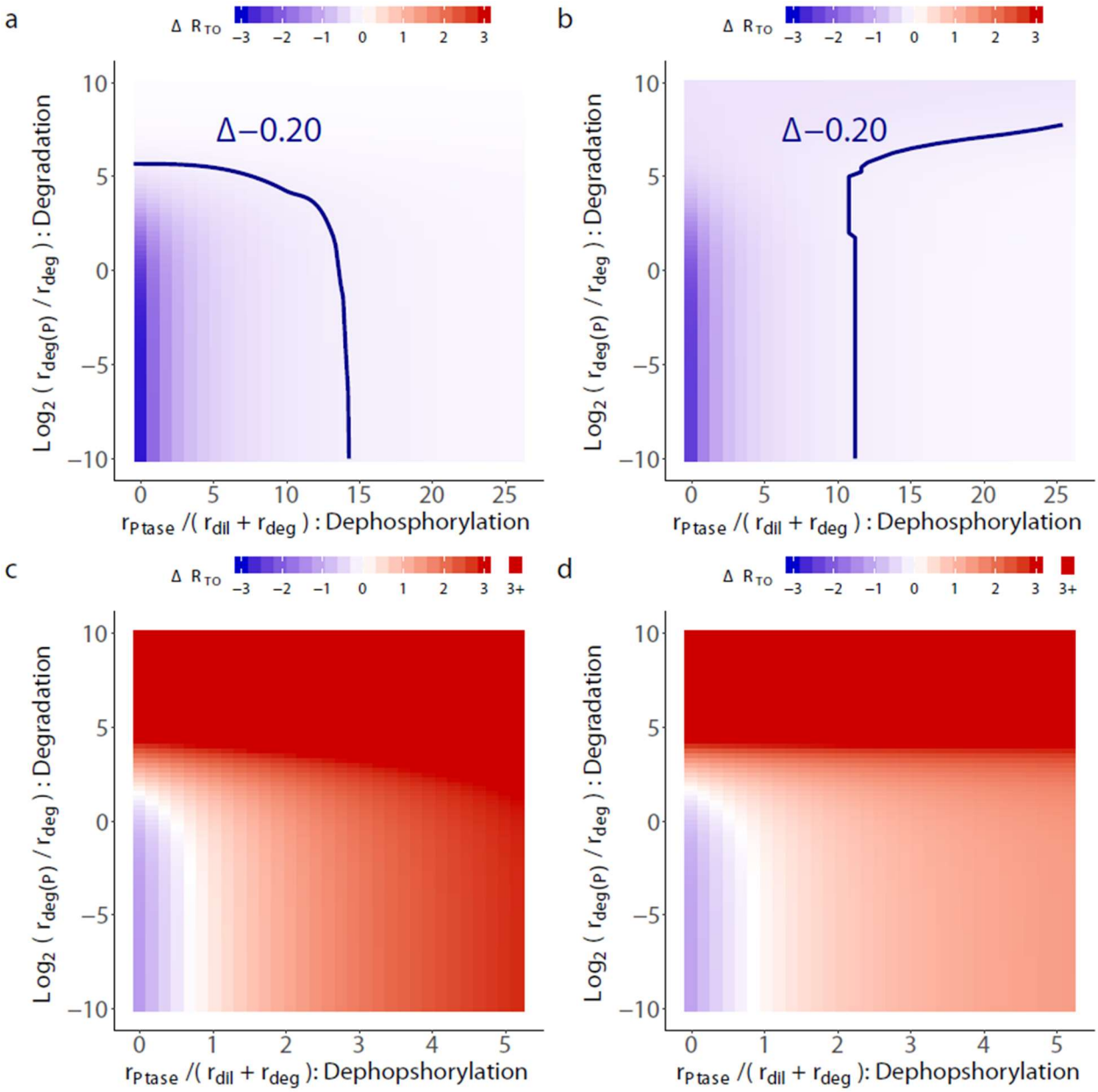
Supplementary Figure 3.1: Most large ΔR_{TO} phosphorylation events are low stoichiometry. a) Scatterplot of ΔR_{TO} (counterpart proteoform - its protein) to ΔR_{TO} (phosphorylation proteoform - its protein). Highlighted pale blue suggests high stoichiometry phosphorylation, where phosphorylation proteoform R_{TO} is more similar to its protein R_{TO} than counterpart proteoform R_{TO}. Highlighted pale yellow would suggest low stoichiometry phosphorylation.



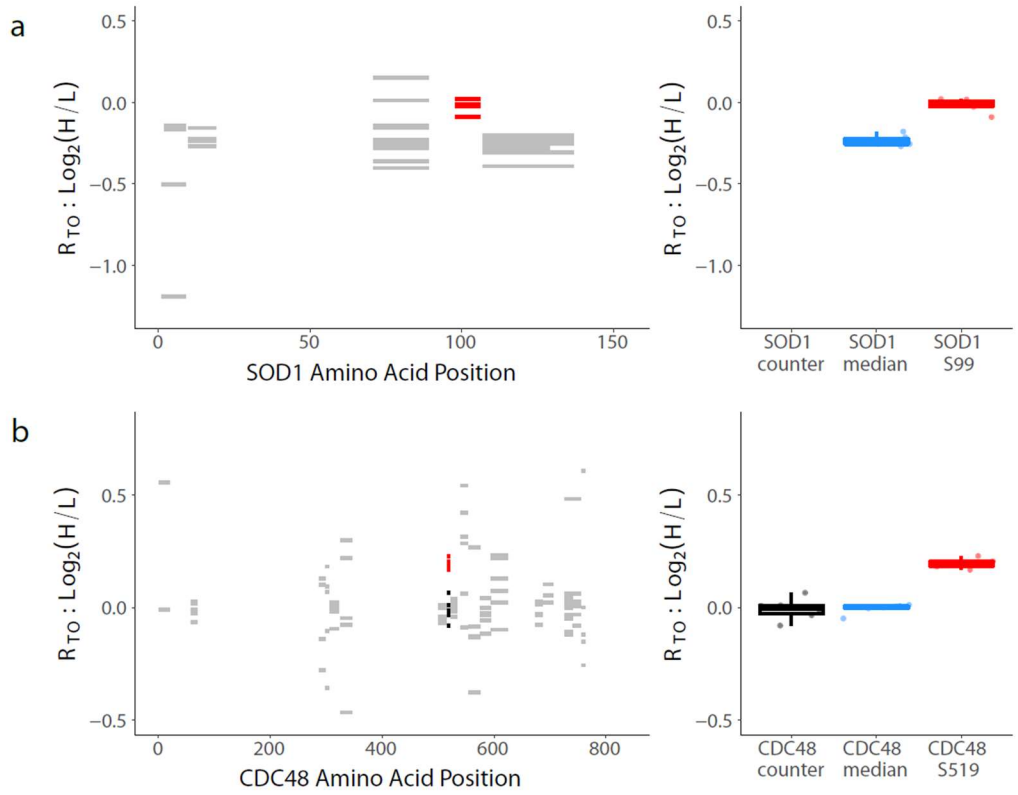
Supplementary Figure 3.2: Unmodified peptides are low abundance in phosphorylation-enriched samples. a) Log_{10} of summed MS1 intensity of heavy- and light-lysine containing phosphopeptides and unmodified peptides in phosphorylation-enriched IMAC samples. Each line represents the distributions of phosphopeptides (gold) and unmodified peptides (blue) for each replicate.



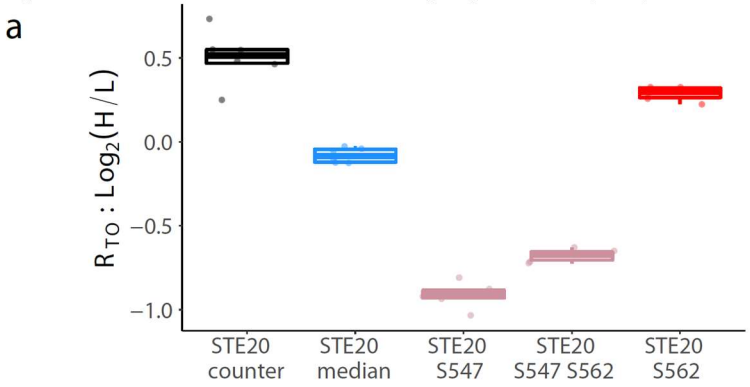
Supplementary Figure 3.3: Increasing phosphorylation isoform degradation rates increase unmodified protein R_{TO} proportional to phosphorylation stoichiometry. a) Simulation using the traditional model with the extension of phosphorylation (Figure 3.3a). Unmodified protein degradation rate (r_{deg}) was set to average degradation of proteome (based on Figure 3.3b) and r_{dil} was set to cellular doubling. Phosphorylation degradation rate ($r_{deg(P)}$) is increased relative to its unmodified protein degradation rate (r_{deg}). Simulation of ΔR_{TO} of the R_{TO} unmodified protein at set increased phosphorylation isoform degradation rate (Unmodified) - R_{TO} of the unmodified protein where $r_{deg} = r_{deg(P)}$ ($Unmodified_{init.}$). Black lines based on ΔR_{TO} for range of $r_{deg(P)}/r_{deg}$ given phosphorylation stoichiometries: 1.5%, 10%, and 20%.



Supplementary Figure 3.4: Kinetic analysis of traditional and age-dependent models for protein turnover with phosphorylation. a) Simulation of ΔR_{TO} of (phosphorylation isoform - its protein) (color: blue < 0, red > 0) for range of phosphatase rates and phosphorylation isoform degradation rates using the traditional model from Figure 3.3a. Phosphatase rate is relative to unmodified protein removal from the cell rates ($r_{dil} + r_{deg}$), and phosphorylation degradation rate ($r_{deg(P)}$) is depicted relative to unmodified protein counterpart degradation rate (r_{deg}). Phosphorylation stoichiometry was set to 1.5%, unmodified degradation rate (r_{deg}) was set to average degradation of proteome (solid line from Figure 3.3b), and r_{dil} was set to cellular doubling. Blue line represents rate combinations that result in -0.20 ΔR_{TO}. **b)** Simulation same as (**a**) but the phosphorylation stoichiometry was set to 20%. **c)** Simulation of ΔR_{TO} of (phosphorylation isoform - its protein) (color: blue < 0, red > 0) for range of phosphatase rates and phosphorylation isoform degradation rates for the age-biased phosphorylation model in Figure 3.3g. Phosphatase rate is relative to unmodified protein removal from the cell rates ($r_{dil} + r_{deg}$), and phosphorylation degradation rate ($r_{deg(P)}$) is depicted relative to unmodified protein degradation rate (r_{deg}). Phosphorylation isoform, unmodified protein pool A, and unmodified protein pool B stoichiometry was set to 1.5%, 20%, 78.5% (respectively), unmodified degradation rate (r_{deg}) was set to average degradation of proteome (solid line from Figure 3.3b), $r_{BtoA} = 0$, and r_{dil} was set to cellular doubling. **d)** Simulation same as (**c**) but the phosphorylation isoform, unmodified protein pool A, and unmodified protein pool B stoichiometry was set to 20%, 20%, 60% (respectively).



Supplementary Figure 3.5: Known functional phosphorylation isoforms have faster R_{TO}. a) (*left*) Replicate R_{TO} values for observed SOD1 unmodified peptides identified in the proteome sample (grey) and SOD1 phospho-serine phosphorylation proteoform (red) displayed across the length of SOD1. (*right*) Boxplot of replicate R_{TO} for unmodified protein (blue) and phosphorylation proteoform S38 (red). **b)** Same as (**a**) for CDC48 and its faster turnover phosphorylation proteoform S519. Replicate R_{TO} values for unmodified counterpart proteoform (black) were added to left and right plots.



Supplementary Figure 3.6: Ste20 unmodified counterpart proteoform to S562 phosphorylation proteoform is low stoichiometry. a) Boxplot of replicate R_{TO} for Ste20 unmodified counterpart proteoform (black), Ste20 unmodified protein (blue), phosphorylation proteoform S562 (red), phosphorylation proteoform S547 (pale red), and dual phosphorylation proteoform S547 S562 (pale red).

SUPPLEMENTARY EQUATIONS FOR SIMULATION

$$Prot_T = Prot_L^U + Prot_H^U + Prot_L^M + Prot_H^M \tag{6}$$

Dynamics

All cells are grown initially in light media so that no heavy protein is present, resulting in the following initial conditions,

$$t = 0, \quad Prot_L^U + Prot_L^M = Prot_T, \quad Prot_H^U + Prot_H^M = 0 \tag{7}$$

We assume that synthesis is governed by a given rate constant, r_{synth}, when a cell is in steady state. When heavy media is added, we further posit that protein synthesis entirely switches over to this source, providing only a single input to the system,

$$RNA \rightarrow Prot_H^U \tag{8}$$

Heavy and light pools of protein are able to reversibly convert between their modified and unmodified forms,

$$Prot_L^U \leftrightharpoons Prot_L^M, \quad Prot_H^U \leftrightharpoons Prot_H^M, \tag{9}$$

Given the post translational modifications studied here, the heavy and light unmodified forms are expected to behave similarly in their modification reactions. Thus, the rate costants which govern the phosphorylation and dephosphorylation rates of modifications, r_{kinase} and r_{Ptase}, can be applied similarly to both isotopic pools.

Protein from all pools can be lost to both degradation processes as well as dilution as cells divide. For the rate of dilution, we do not expect any difference between the modified and unmodified pools, and we expect the pool specific rate to be goverened by stoichiometry,

$$r_{dil}^U = (1-\theta) * r_{dil}, \quad r_{dil}^M = \theta * r_{dil} \tag{10}$$

For the degradation rate, however, we are interested in additional parameter, the change in degradation rate which is a consequence of modification. We model this as the multiplicative factor ρ,

$$r_{deg}^U = (1-\theta) * r_{deg}, \quad r_{deg}^M = \rho * \theta * r_{deg} \tag{11}$$

This gives ρ an interpretation as the relative increase or decrease in degradation rate over what we would expect from stoichiometry alone. The total rate of disapearence is then the sum of all of these,

$$r_{dis} = r_{dil}^U + r_{dil}^M + r_{deg}^U + r_{deg}^M \tag{12}$$

Substituting in eq. 10 and eq. 11 into eq. 12, rearranging terms, and applying the fact that the rate of protein removal (degradation + dilution) must match the rate of synthesis to maintain steady state protein levels, thus allowing us to derive the protein steady state equation (equivalent to eq. 5),

$$r_{synth} = (1-\theta)(r_{dil} + r_{deg}) + \theta(r_{dil} + \rho * r_{deg}) \tag{13}$$

We define the term on the right of eq. 13 to be the specific rate of disapearence of modified protein, $r_{dis}^M = \theta(r_{dil} + \rho * r_{deg})$, which along with the modification dephosphorylation rate defined above, affects the total level of modified protein in the cell. Thus, in order to satisfy our assumption of steady state stoichiometry defined in 1, we set

$$(1-\theta) * r_{mod} = \theta * r_{Ptase} + r_{dis}^M = \theta * r_{Ptase} + \theta(r_{dil} + \rho * r_{deg}) \tag{14}$$

Rate Equations

In order to build our models, we need to derive 4 rate equations, $\frac{dProt_H^U}{dt}$, $\frac{dProt_L^U}{dt}$, $\frac{dProt_H^M}{dt}$, $\frac{dProt_L^M}{dt}$. Each of these is derived similarly given the above rates, so we will explicitly derive only $\frac{dProt_H^U}{dt}$. This requires us to both define the rate of input, $\frac{dInput_H^U}{dt}$, as well as the rate of output, $\frac{dOutput_H^U}{dt}$. $Prot_H^U$ recieves input from $Prot_H^M$ when post translational modifications are removed, and as stated in eq. 8, is the only pool recieving new protein from sythesis. For the heavy unmodified pool derivation, all rates are similarly impacting both heavy and light pools, thus the rates only need to be corrected by the proportion (heavy/total). Same correction can be applied to light protein pools for respective light protein pools (unmodified and modified). For minimizing notation, summed heavy and light of a protein pool ($Prot_H^X + Prot_L^X$) is equivalent to $Prot_{H+L}^X$.

Thus, we can define the total rate of input into $Prot_H^U$ as,

$$\frac{dInput_H^U}{dt} = r_{synth} + \theta * r_{Ptase}\frac{Prot_H^M}{Prot_H^M + Prot_L^M} = r_{synth} + \theta * r_{Ptase}\frac{Prot_H^M}{Prot_{H+L}^M} \tag{15}$$

Similarly, we can derive the output of the system by considering the fact that rate of disapearence governing the unmodified fraction of protein, $(1-\theta)(r_{dil} + r_{deg})$, also needs to be modified by the proportion of the unmodified pool made up of heavy protein,

$$\frac{dOutput_H^U}{dt} = -(1-\theta)(r_{dil} + r_{deg})\frac{Prot_H^U}{Prot_{H+L}^U} - (1-\theta)r_{kinase}\frac{Prot_H^U}{Prot_{H+L}^U} \tag{16}$$

Finally, eq. 15 and eq. 16 can be added together to give,

$$\frac{dProt_H^U}{dt} = r_{synth} + \theta * r_{Ptase}\frac{Prot_H^M}{Prot_{H+L}^M} - (1-\theta)(r_{dil} + r_{deg})\frac{Prot_H^U}{Prot_{H+L}^U} - (1-\theta)r_{kinase}\frac{Prot_H^U}{Prot_{H+L}^U} \tag{17}$$

The rate equation is almost exactly the same for $Prot_L^U$, without the added rate of sythesis,

$$\frac{dProt_L^U}{dt} = \theta * r_{Ptase}\frac{Prot_L^M}{Prot_{H+L}^M} - (1-\theta)(r_{dil} + r_{deg})\frac{Prot_L^U}{Prot_{H+L}^U} - (1-\theta)r_{kinase}\frac{Prot_L^U}{Prot_{H+L}^U} \tag{18}$$

Switching r_{kinase} for r_{Ptase} and considering r_{dis}^M allows us to derive the rate equations for the modified pools,

$$\frac{dProt_H^M}{dt} = (1-\theta)r_{kinase}\frac{Prot_H^U}{Prot_{H+L}^U} - \theta * r_{Ptase}\frac{Prot_H^M}{Prot_{H+L}^M} - \theta(r_{dil} + \rho * r_{deg})\frac{Prot_H^M}{Prot_{H+L}^M} \tag{19}$$

$$\frac{dProt_L^M}{dt} = (1-\theta)r_{kinase}\frac{Prot_L^U}{Prot_{H+L}^U} - \theta * r_{Ptase}\frac{Prot_L^M}{Prot_{H+L}^M} - \theta(r_{dil} + \rho * r_{deg})\frac{Prot_L^M}{Prot_{H+L}^M} \tag{20}$$

In order to better understand the complex observed results from the dynamic SILAC experiment with phosphorylation, we generated a simulator using the above rate equations. The simulation iterated through small step-wise protein molecule additons on the scale of every 6 seconds to accurately estimate the abundance of all protein pools over the 90 minute experiment. The strategy of the simulation was to walk through the protein pools one at a time starting in order of synthesis. In this model, protein input rates for the unmodified protein were added to the pool defining a new proportion of light and heavy molecules for that pool. Output protein molecules were defined from new proportions and respective output protein pool rates. Same simulation process order: (add input molecules (input rates), new proportion heavy/light proportion calculation, and subtract output molecules (output rates and new proportion)) was applied next pool since synthesis, which in this case is the modified protein pool. Finally, heavy and light protein abundance ratios for unmodified and modified protein pools were calculated to generate a theoretically observed R_{TO}.

Age-biased phosphorylation turnover model

One hypothesis for the observation of faster turnover phosphorylation sites is that there is a phosphorylation bias towards "newly-synthesized" unmodified proteins. This alteration to the model cannot be implemented into the tradational model above. The simplest steady state model to enable this phenomena requires the generation of a "newer" (since synthesis) unmodified protein pool and an "older" unmodified protein pool, with only the "newer" unmodified pool being phosphorylated.

Thus, we consider the three protein pools, $Prot_T = Prot^{new} + Prot^{old} + Prot^{phos}$ and the following stoichiometry terms $\theta^{phos} = Prot^{phos}/Prot_T$, $\theta^{new} = Prot^{new}/Prot_T$, $\theta^{old} = Prot^{old}/Prot_T$ under the following constraints:

$$\frac{dProt_T}{dt} = 0, \quad \frac{d\theta^{phos}}{dt} = 0, \quad \frac{d\theta^{new}}{dt} = 0, \quad \frac{d\theta^{old}}{dt} = 0 \tag{21}$$

Given that the total protein concentration and the stoichiometries remain constant,

$$\frac{dProt^{phos}}{dt} = 0, \quad \frac{dProt^{new}}{dt} = 0, \quad \frac{dProt^{old}}{dt} = 0, \tag{22}$$

Given this system, in a dynamic SILAC experiment:

$$Prot_T = Prot_L^{new} + Prot_H^{new} + Prot_L^{old} + Prot_H^{old} + Prot_L^{phos} + Prot_H^{phos} \tag{23}$$

At the moment of the heavy lysine pulse,

$$t = 0, \quad Prot_T = Prot_L^{new} + Prot_L^{old} + Prot_L^{phos}, \quad Prot_H^{new} + Prot_H^{old} + Prot_H^{phos} = 0 \tag{24}$$

For this system only the heavy "newer" unmodified protein pool can be synthesized:

$$RNA \rightarrow Prot_H^{new} \tag{25}$$

The age-biased phosphorylation model has phosphorylated protein being generated from only the "newly synthesized" unmodified protein pool via phosphorylation (r_{kinase}), and the new unmodified protein regenerated via the dephosphorylation (r_{Ptase}) rate.

$$Prot_L^{new} \leftrightharpoons Prot_L^{phos}, \quad Prot_H^{new} \leftrightharpoons Prot_H^{phos}, \tag{26}$$

In the simplest form of the age dependent model, only the "newly synthesized" unmodified protein pool can generate the "older" unmodified protein pool. The generation from "new" to "old" is defined by r_{AtoB} and generation from "old" to "new" is defined by r_{BtoA}.

$$Prot_L^{new} \leftrightharpoons Prot_L^{old}, \quad Prot_H^{new} \leftrightharpoons Prot_H^{old}, \tag{27}$$

Additionally, in this model the "newly synthesized" unmodified protein pool can undergo three fates: generate the "older" unmodified protein pool, be removed (degraded or diluted) from the cell, or generate the phosphorylated protein pool. Phosphorylated protein and "older" unmodifed protein degrade and dilute from the system as observed in previous model.

Thus the steady state equations are the following:

For synthesis equals total disappearence from the system:

$$r_{synth} = \theta^{new} * (r_{deg} + r_{dil}) + \theta^{old} * (r_{deg} + r_{dil}) + \theta^{phos} * (\rho * r_{deg} + r_{dil}) \tag{28}$$

Steady state for phosphorylation event:

$$\theta^{new} * r_{kinase} = \theta^{phos} * r_{Ptase} + \theta^{phos} * (\rho * r_{deg} + r_{dil}) \tag{29}$$

Steady state for "new" to "old" unmodified protein and vice versa conversion:

$$\theta^{new} * r_{AtoB} = \theta^{old} * r_{BtoA} + \theta^{old} * (r_{deg} + r_{dil}) \tag{30}$$

Rate equations for all protein pools in Age-biased phosphorylation model:

$$\frac{dProt_H^{new}}{dt} = r_{synth} + \theta^{phos} * r_{Ptase} \frac{Prot_H^{phos}}{Prot_{H+L}^{phos}} + \theta^{old} * r_{BtoA} \frac{Prot_H^{old}}{Prot_{H+L}^{old}} - \theta^{new} * r_{kinase} \frac{Prot_H^{new}}{Prot_{H+L}^{new}} - \theta^{new} * r_{AtoB} \frac{Prot_H^{new}}{Prot_{H+L}^{new}} - \theta^{new} * (r_{deg} + r_{dil}) \frac{Prot_H^{new}}{Prot_{H+L}^{new}} \tag{31}$$

$$\frac{dProt_L^{new}}{dt} = \theta^{phos} * r_{Ptase} \frac{Prot_L^{phos}}{Prot_{H+L}^{phos}} + \theta^{old} * r_{BtoA} \frac{Prot_L^{old}}{Prot_{H+L}^{old}} - \theta^{new} * r_{kinase} \frac{Prot_L^{new}}{Prot_{H+L}^{new}} - \theta^{new} * r_{AtoB} \frac{Prot_L^{new}}{Prot_{H+L}^{new}} - \theta^{new} * (r_{deg} + r_{dil}) \frac{Prot_L^{new}}{Prot_{H+L}^{new}} \tag{32}$$

$$\frac{dProt_H^{phos}}{dt} = \theta^{new} * r_{kinase} \frac{Prot_H^{new}}{Prot_{H+L}^{new}} - \theta^{phos} * r_{Ptase} \frac{Prot_H^{phos}}{Prot_{H+L}^{phos}} - \theta_{phos}(r_{dil} + \rho * r_{deg}) \frac{Prot_H^{phos}}{Prot_{H+L}^{phos}} \tag{33}$$

$$\frac{dProt_L^{phos}}{dt} = \theta^{new} * r_{kinase}\frac{Prot_L^{new}}{Prot_{H+L}^{new}} - \theta^{phos} * r_{Ptase}\frac{Prot_L^{phos}}{Prot_{H+L}^{phos}} - \theta_{phos}(r_{dil} + \rho * r_{deg})\frac{Prot_L^{phos}}{Prot_{H+L}^{phos}} \tag{34}$$

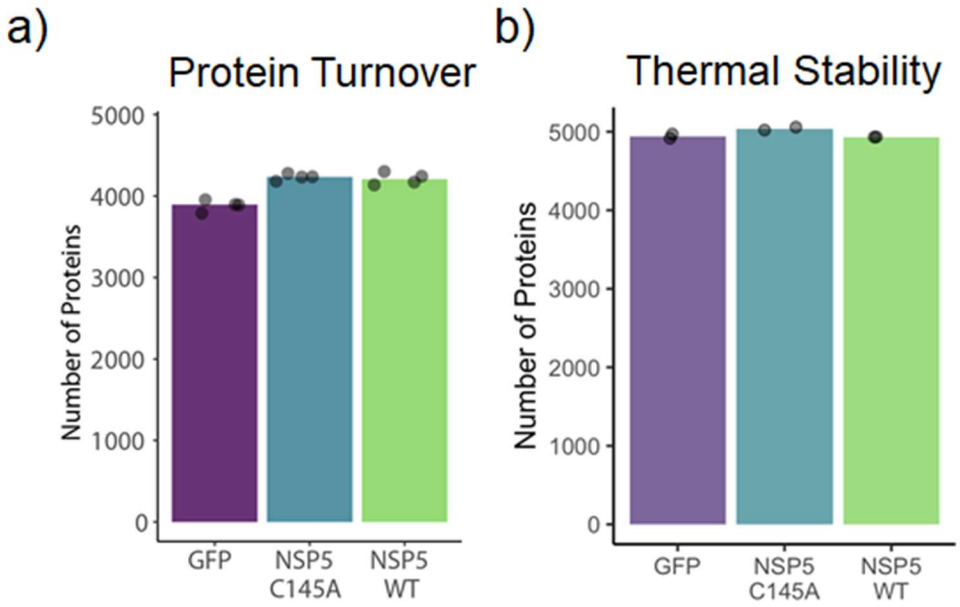
$$\frac{dProt_H^{old}}{dt} = \theta^{new} * r_{AtoB}\frac{Prot_H^{new}}{Prot_{H+L}^{new}} - \theta^{old} * r_{BtoA}\frac{Prot_H^{old}}{Prot_{H+L}^{old}} - \theta_{old}(r_{dil} + r_{deg})\frac{Prot_H^{old}}{Prot_{H+L}^{old}} \tag{35}$$

$$\frac{dProt_L^{old}}{dt} = \theta^{new} * r_{AtoB}\frac{Prot_L^{new}}{Prot_{H+L}^{new}} - \theta^{old} * r_{BtoA}\frac{Prot_L^{old}}{Prot_{H+L}^{old}} - \theta_{old}(r_{dil} + r_{deg})\frac{Prot_L^{old}}{Prot_{H+L}^{old}} \tag{36}$$

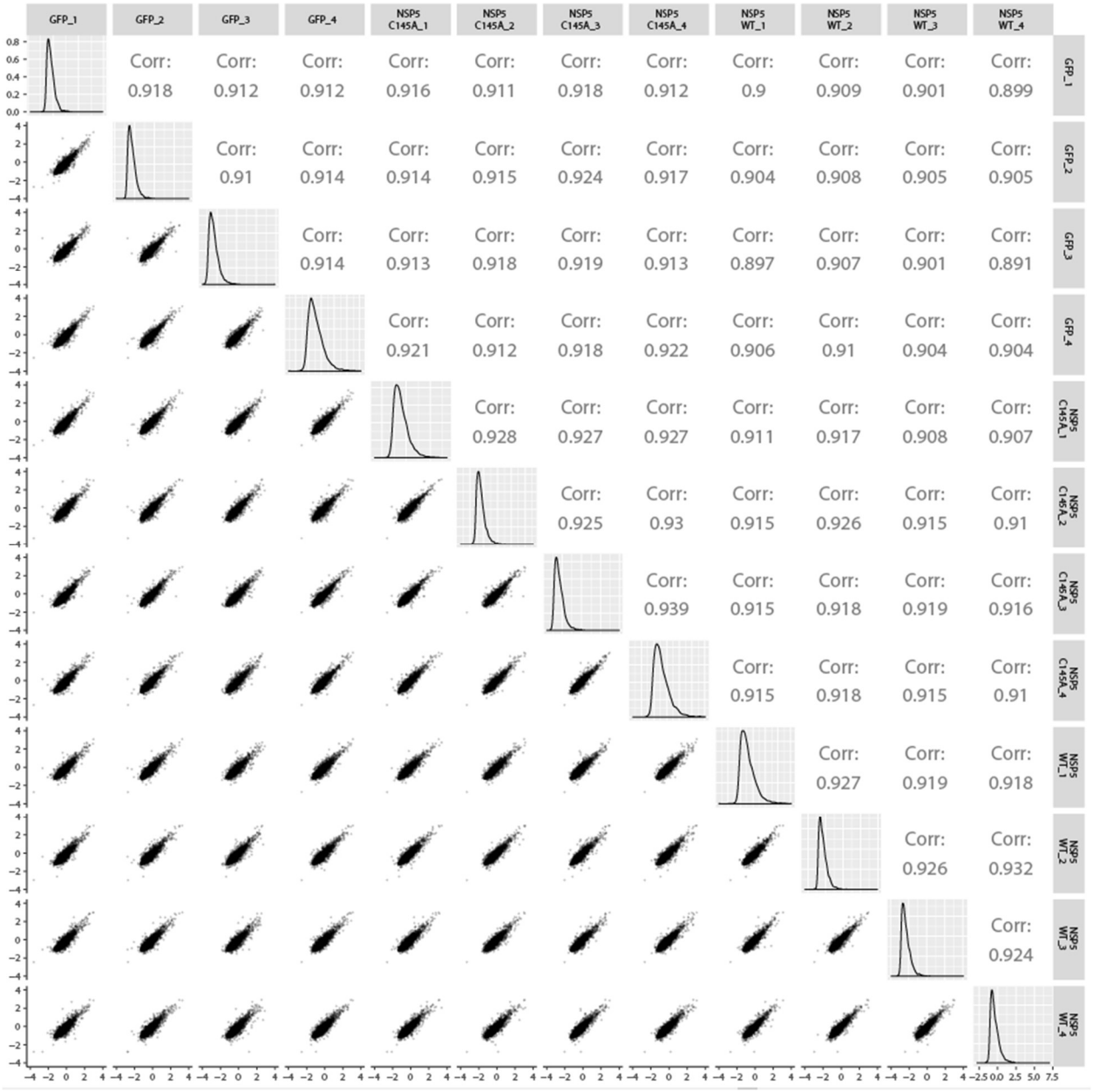
APPENDIX C

Appendix C: Supplementary information for Chapter 4: Identification of SARS-CoV-2 NSP5 host protease substrates by protein turnover or thermal stability

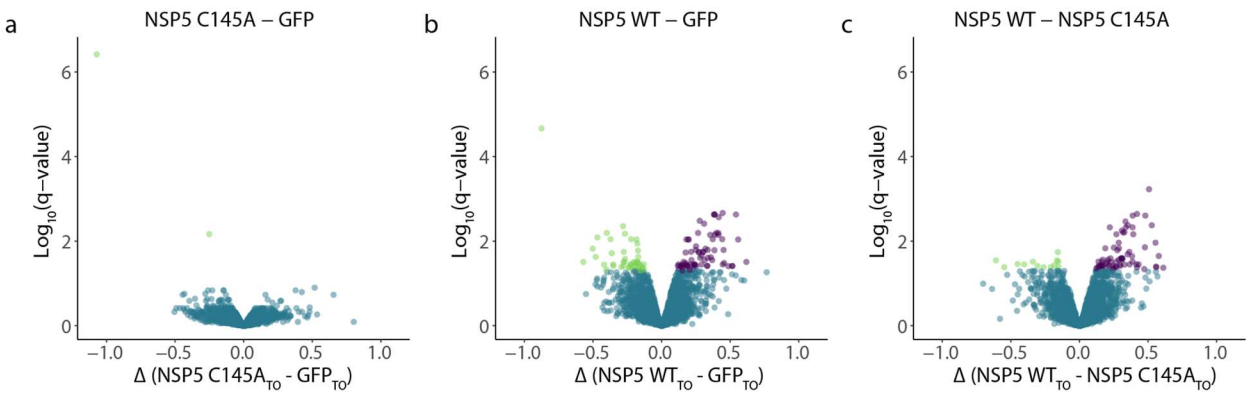
SUPPLEMENTARY FIGURES



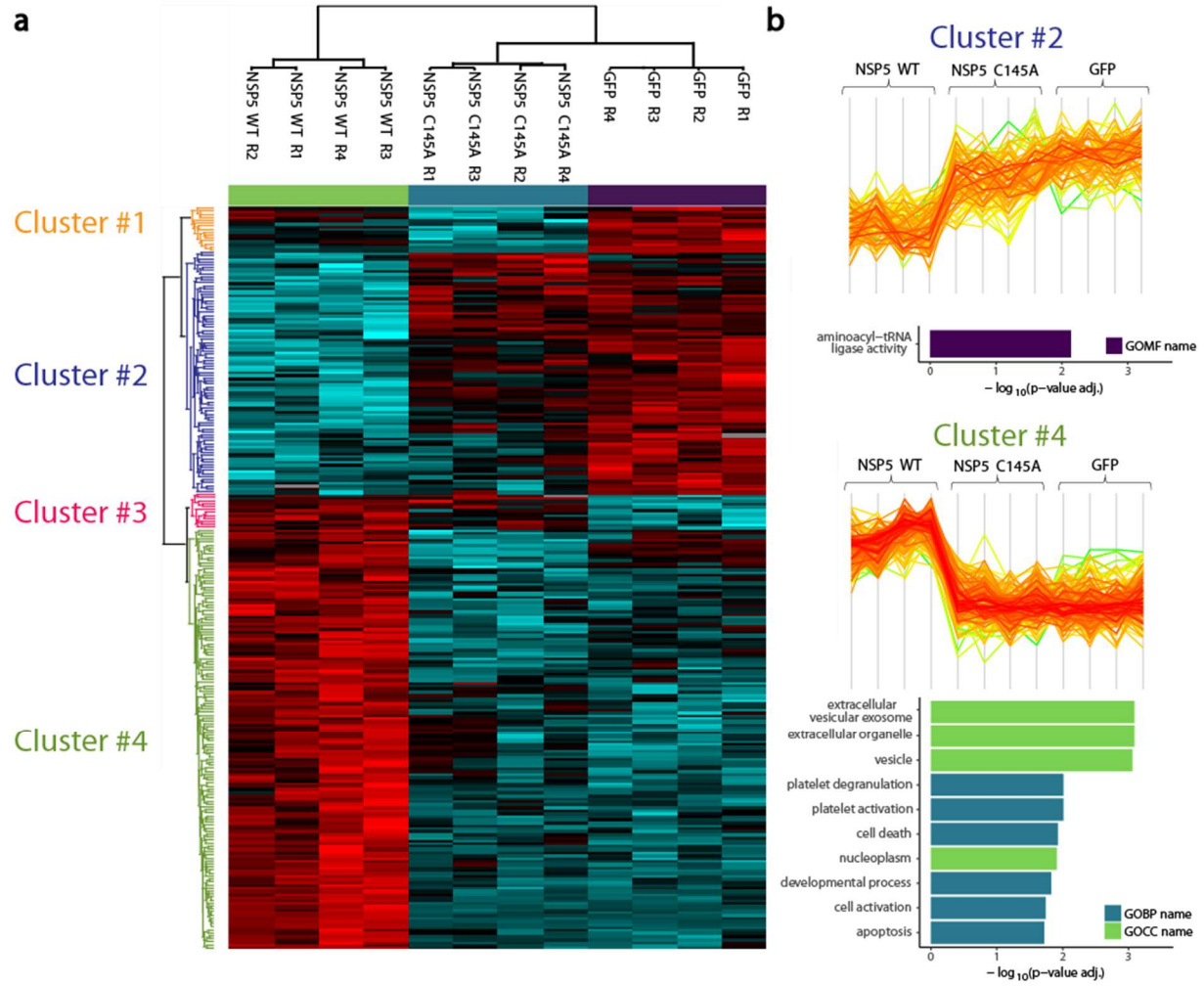
Supplementary Figure 4.1: Protein identifications for protein turnover and protein thermal stability assay. a) For protein turnover, bar plot of protein identifications of HEK293T proteomes overexpressing GFP (purple), NSP5 C145A (blue), and NSP5 wildtype (green) with points representing identifications for each replicate. **b)** Same as (**a**) for protein thermal stability assay (thermal proteome profiling: TPP).



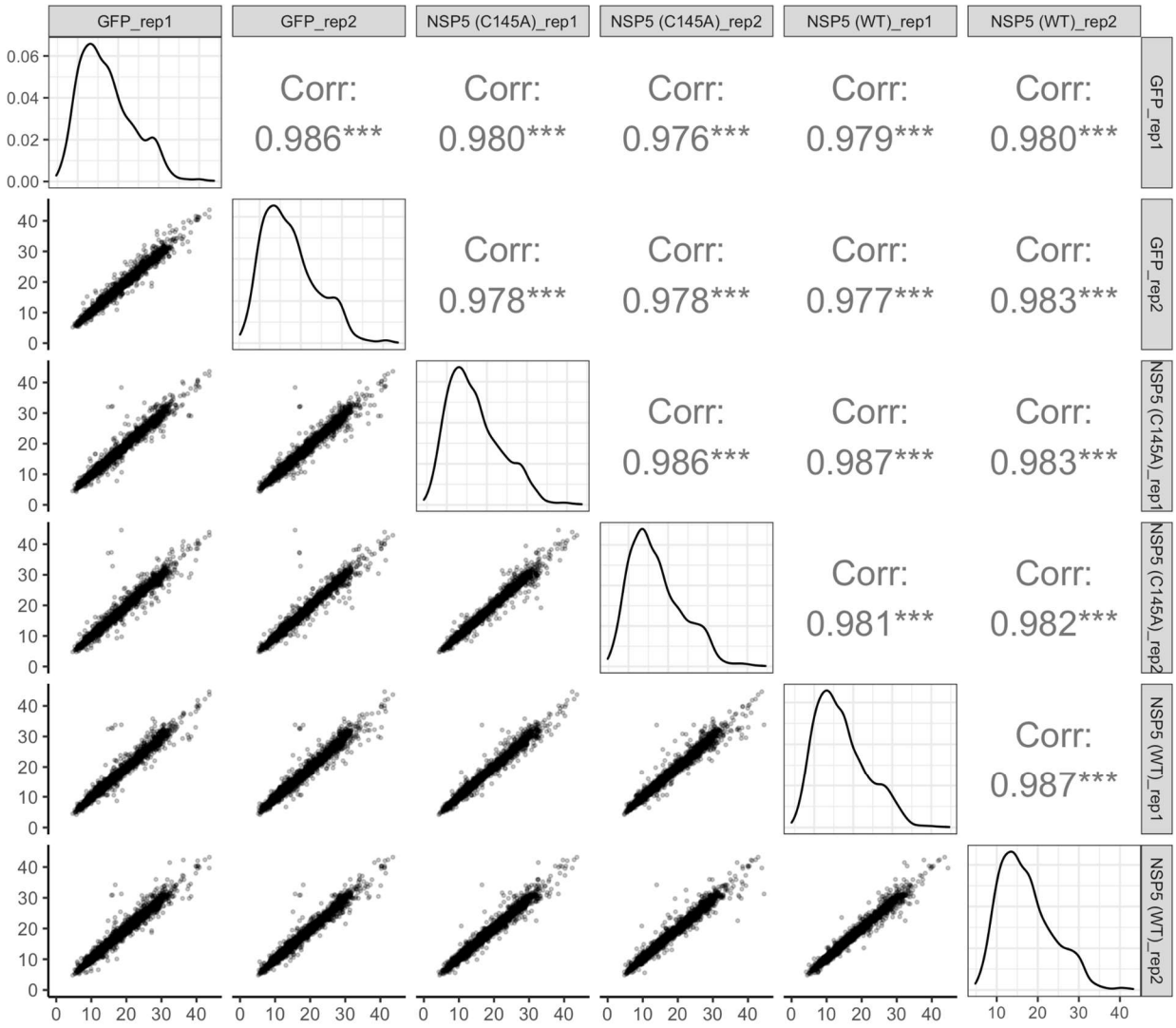
Supplementary Figure 4.2: Dynamic SILAC replicate reproducibility. a) The lower triangle contains scatter plots comparing R_{TO} for all pairwise replicates (within and across protein overexpression conditions). Each point represents a protein. The density plots along the diagonal are for each replicate's R_{TO} distribution across conditions. The upper triangle contains the Pearson Correlation R for all pairwise replicates across all the overexpression conditions.



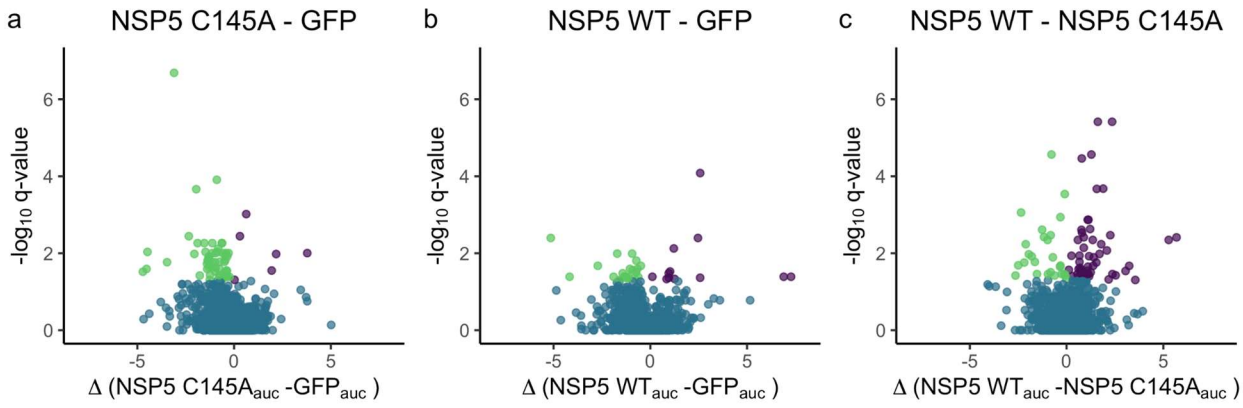
Supplementary Figure 4.3: Pairwise Limma analysis of overexpression conditions. a) Pairwise statistical comparison between NSP5 C145A and GFP with Limma. The x-axis designates the difference in median replicate R_{TO} between the conditions or ΔR_{TO}. The y-axis designates the negative log10 of the Benjamini-Hochberg adjusted Limma p-value (q-value). Significantly faster (purple) and slower (green) R_{TO} proteins are designated by a q-value < 0.05. Proteins with no significant difference in R_{TO} are designated in blue. **b)** Same as in (**a**) but for the pairwise R_{TO} comparison between NSP5 wildtype and GFP. **c)** Same as in (**a**) but for the pairwise R_{TO} comparison between NSP5 wildtype and NSP5 C145A.



Supplementary Figure 4.4: Hierarchical clustering of samples and ANOVA significant proteins for GO enrichment analysis. a) Dendrogram that uses hierarchical clustering to cluster sample replicates (columns) and ANOVA significant proteins (rows). Bar plot above the dendrogram is colored according to the overexpression condition (purple:GFP ; blue:NSP5 C145A ; green:NSP5 wildtype). ANOVA significant proteins were clustered into 4 groups with its tree and label colored accordingly. Protein values are Z-score scaled across all condition's replicates (row-wise) scaled from higher R_{TO} as red and lower R_{TO} as blue. **b)** Extracted protein R_{TO} profiles (Z-score scaled values as in (**a**)) are projected as line graphs for Cluster 2 (top) and Cluster 4 (bottom). Gene ontology enrichments for Clusters 2 and 4 (top and bottom respectively) are plotted as bar plots with its adjusted p-value plotted on the x-axis (all adjusted p-value < 0.05) and it's GO term on the y-axis. Bars are colored according to their broad GO term designation (GO Molecular Function:GOMF ; GO Biological Process:GOBP ; GO Cellular Compartment:GOCC).



Supplementary Figure 4.5: Crude Thermal Proteome Profiling replicate reproducibility. a) The lower triangle contains scatter plots comparing protein area under the melting curve for all pairwise replicates (within and across protein overexpression conditions). Each point represents a protein. The density plots along the diagonal are for each replicate's AUC distribution across conditions. The upper triangle contains Pearson Correlation R for all pairwise replicates across all the overexpression conditions.



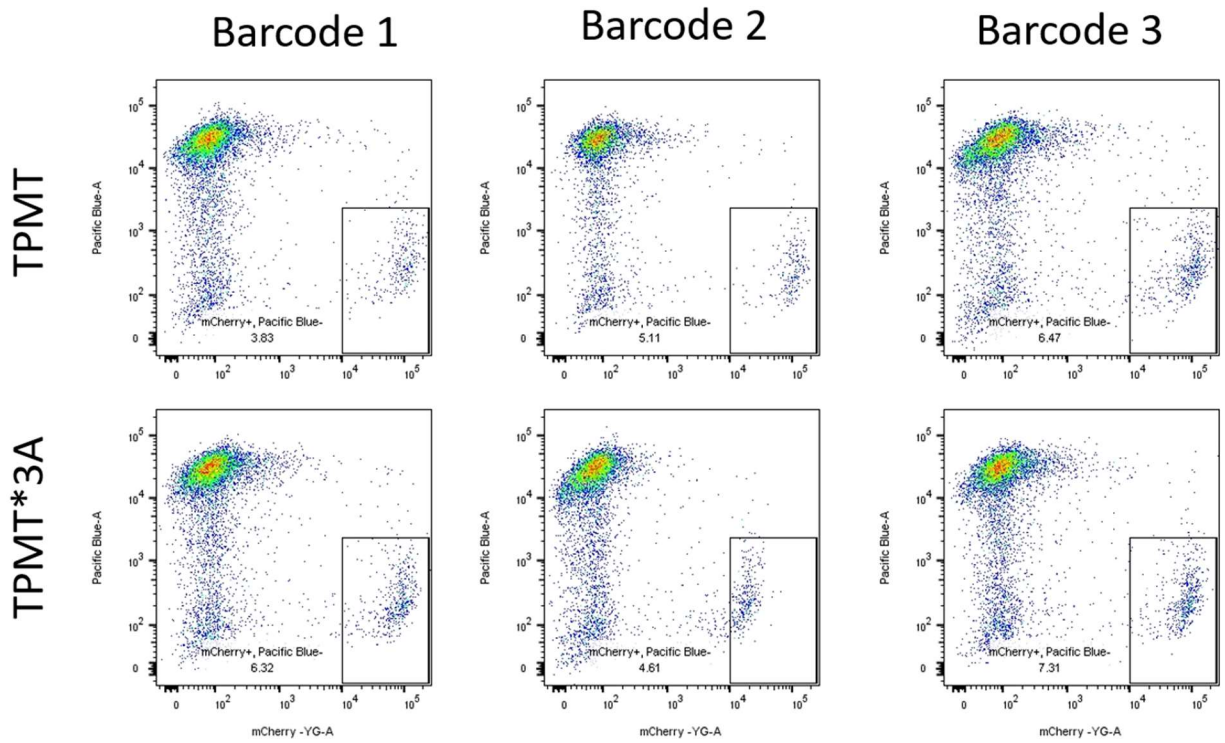
Supplementary Figure 4.6: Pairwise NPARC analysis of overexpression conditions. a) Pairwise statistical comparison between **(a)** NSP5 C145A and GFP; **(b)** NSP5 WT and GFP; **(c)** NSP WT and NSP5 C145A with non-parametric analysis of response curves (NPARC). The x-axis designates the difference in median replicate area under the melting curve (auc) between the conditions. The y-axis designates the negative log10 of the Benjamini-Hochberg adjusted NPARC p-value (q-value). Significantly stabilized (purple) and destabilized (green) proteins are designated by a q-value < 0.05. Proteins with no significant difference in R_{TO} are designated in blue.

APPENDIX D

Appendix D: Supplementary information for Chapter 5: Developing a peptide barcoding method to assess stability of thousands of TPMT protein variants

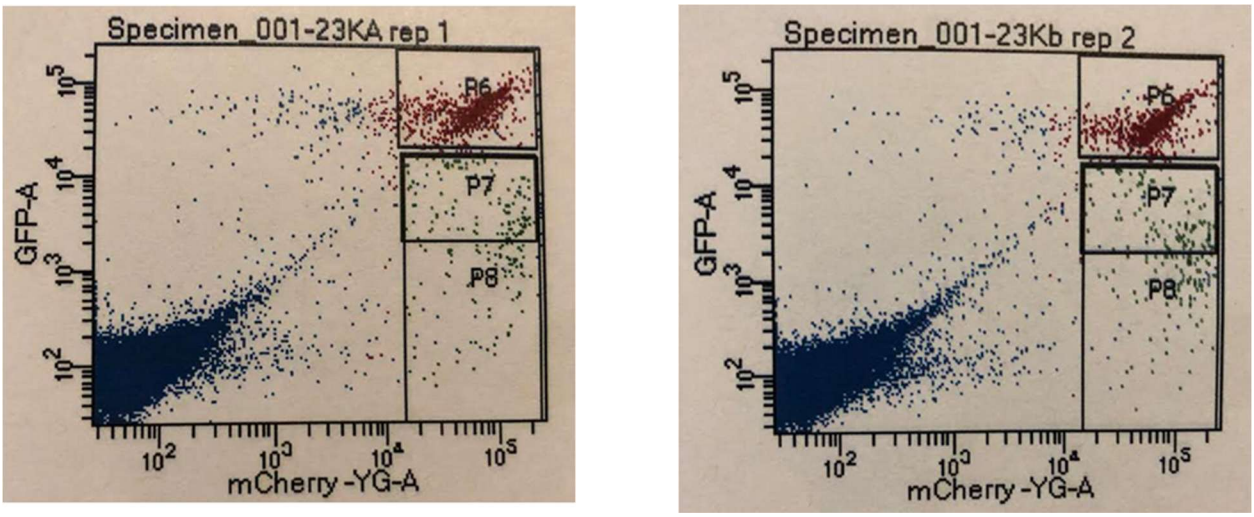
SUPPLEMENTARY FIGURES

Supplementary Figure 5.1: ~5% successful transfection efficiency of GPS vector in HEK293T landing pad cell lines



Supplementary Figure 5.1: ~5% successful transfection efficiency of GPS vector in HEK293T landing pad cell lines. FACS VAMP-Seq abundance assay of HEK293T single cells sorted against Pacific Blue (high : unintegrated cells) and mCherry (high: integration of GPS vector and high expression) for wildtype TPMT and the TPMT*3A haplotype. Box over integrated cell population across 100,000's single cells to capture 25,000 within the gated regain at labeled percentage of total number of sorted cells.

Supplementary Figure 5.2: Peptide barcodes generally do not alter stability.



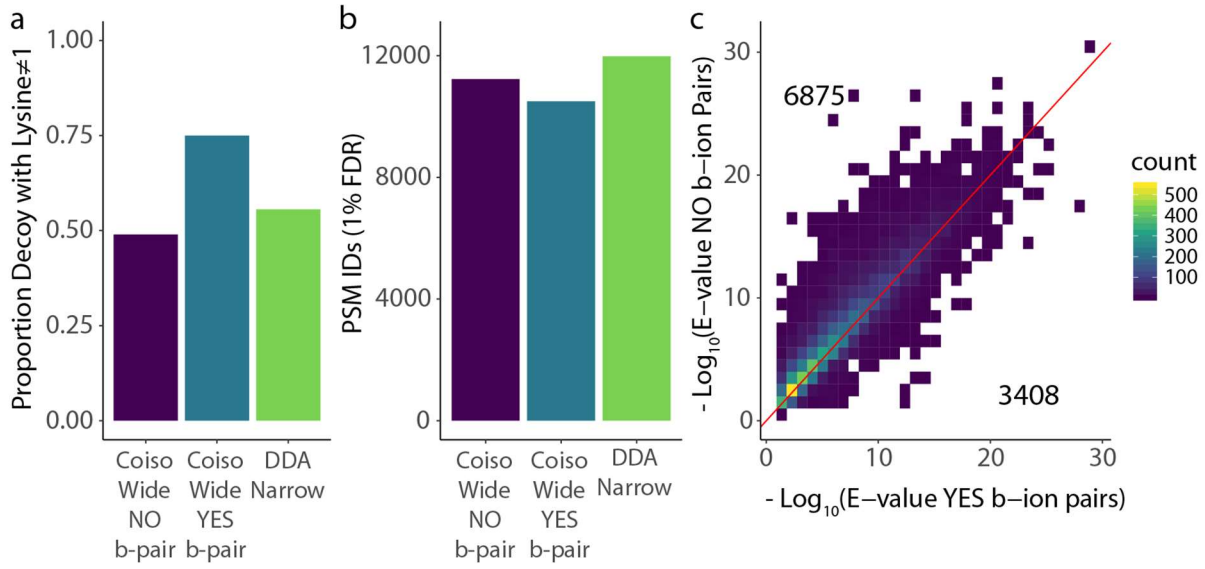
Supplementary Figure 5.2: Peptide barcodes generally do not alter stability. a-b) 25,000 HEK293T landing pad cells expressing wildtype TPMT tagged with a library of 23,000 peptide barcodes sorted using VAMP-Seq FACS assay for stability/abundance (GFP) and expression control (mCherry) across two replicates (*left* = rep1, *right* = rep2).

APPENDIX E

Appendix E: Supplementary information for Chapter 6: Coisolation of peptide pairs for identification and MS/MS-based quantification

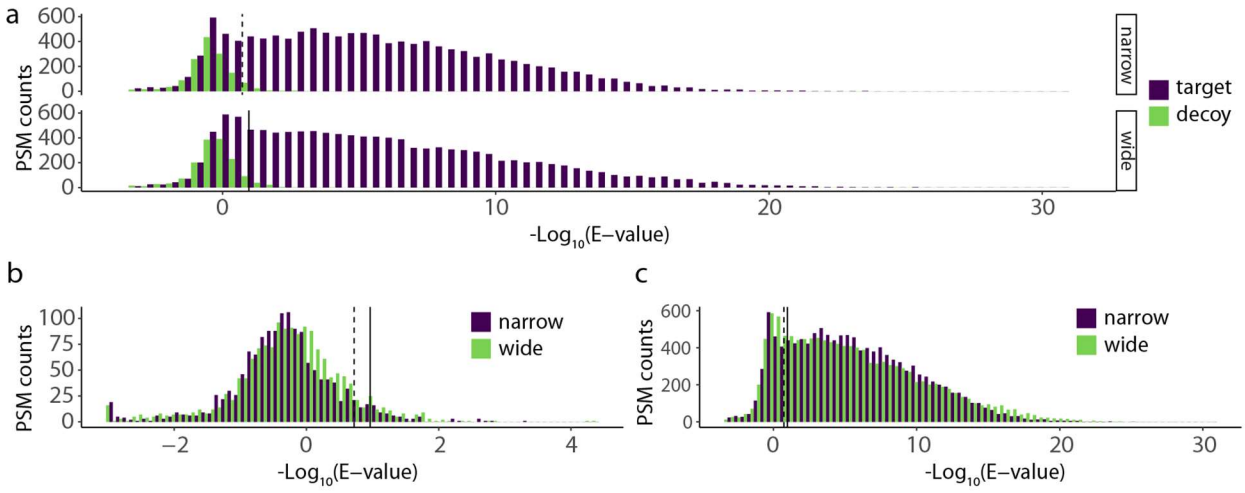
SUPPLEMENTARY FIGURES

Supplementary Figure 6.1: Inclusion or exclusion of paired b-ions for peptide-spectral matching



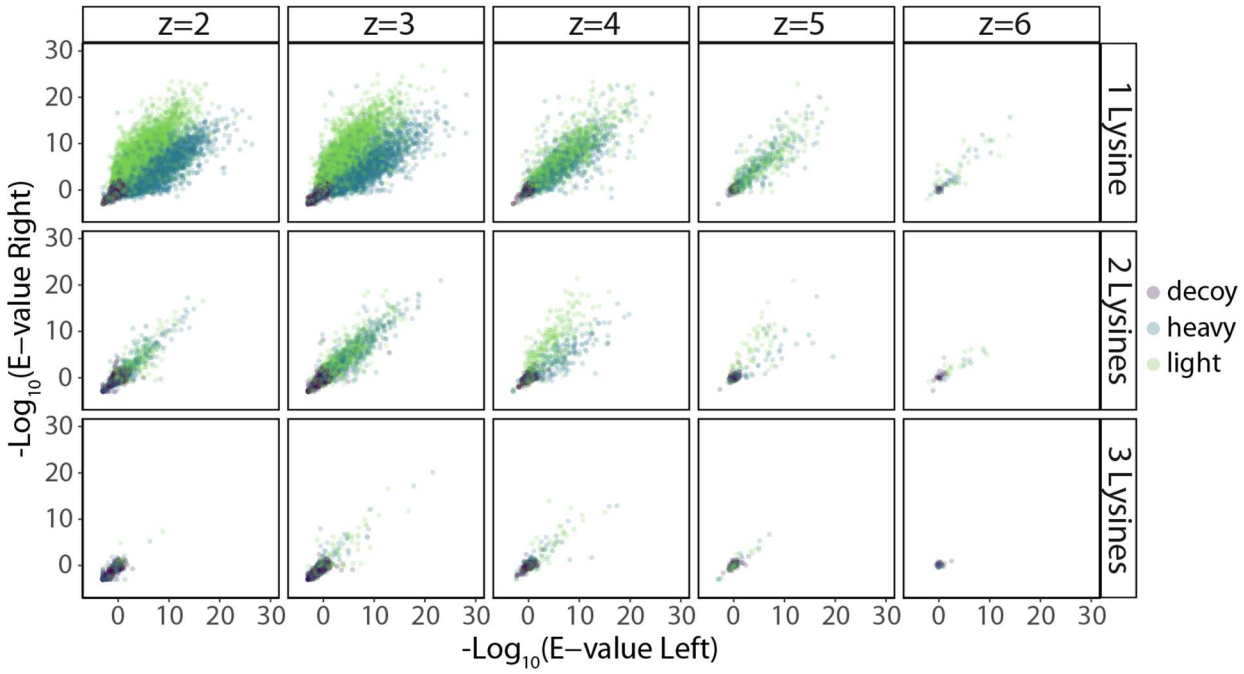
Supplementary Figure 6.1: Coiso SILAC Comet searches excluding b-ion pairs removes search bias. a) For *S. cerevisiae* 1:1 SILAC ratio proteome mixture, bar plot of the proportion of decoy hits that have lysine not equal to 1 for Coiso Wide search including or excluding b-ion paired fragments and the traditional DDA search with narrow isolation considering the same targeted precursor m/z. **b)** Bar plot of PSMs at 1% FDR based on Comet E-value for the same acquisition search combinations as in (**a**). **c)** Density-binned scatter plot comparing Comet standard E-value including (x-axis) and excluding (y-axis) b-ion paired fragments for matching PSM identification and triggered precursor m/z (filtered in both searches at 1% FDR based on E-value).

Supplementary Figure 6.2: E-values for DDA and Coiso SILAC analysis



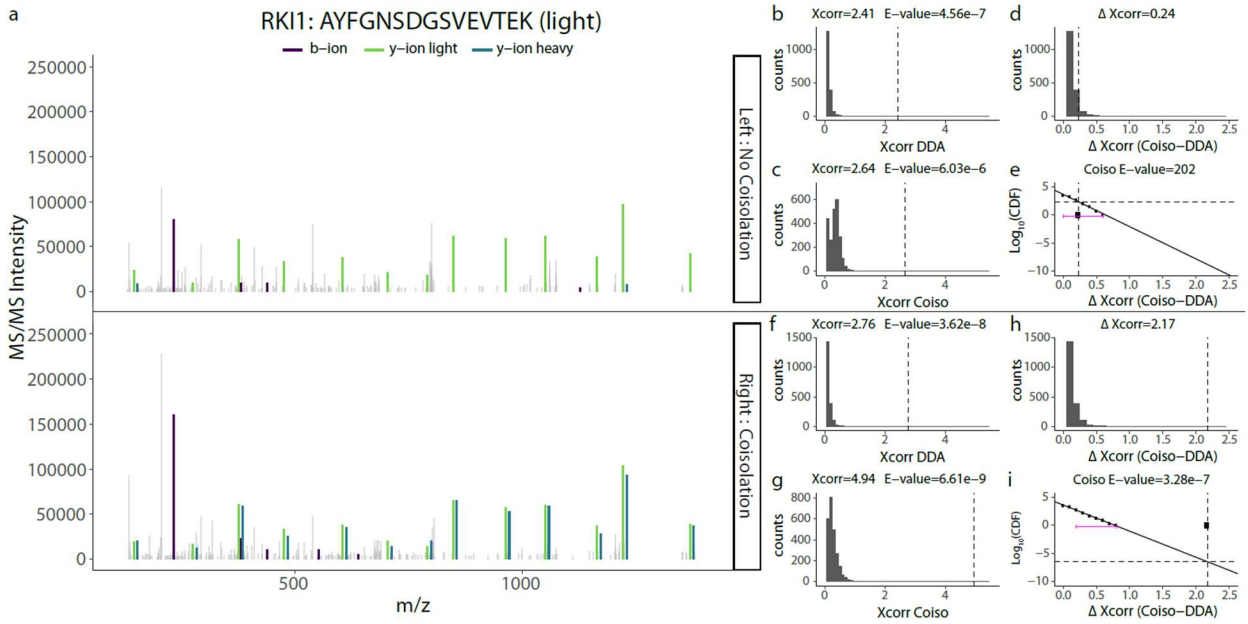
Supplementary Figure 6.2: Coiso SILAC E-values shifted for targets and decoys. a) For *S. cerevisiae* 1:1 SILAC ratio proteome mixture, histograms for PSM counts for targets (purple) and decoys (green) for narrow window scans searched with traditional DDA Comet parameters (top) and for wide window offset scans searched with Coiso Comet parameters (bottom). E-value 1% PSM FDR cut-offs are designated by vertical lines (dashed : DDA Narrow; solid: Coiso Wide). **b)** Histograms for PSM decoys of (**a**) for both MS acquisition-search combinations overlaid. Wide window scans with Coiso search in green and narrow window scans with traditional DDA search in purple, with respective 1% PSM FDR designations as in (**a**). **c)** Histograms for PSM targets as in (**a**) for both MS acquisition-search combinations overlaid. Same color designations as in (**b**) and 1% PSM FDR cut-offs as in (**a**).

Supplementary Figure 6.3: E-value comparing Coiso scans between offsets



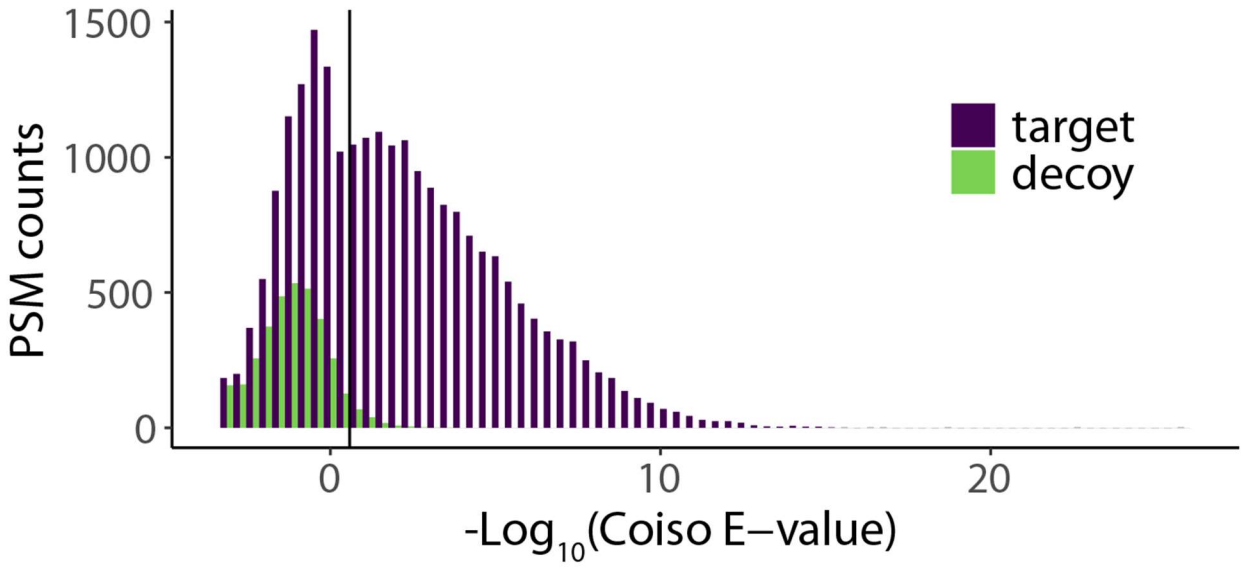
Supplementary Figure 6.3: Comet E-value can marginally distinguish between successful non-successful SILAC peptide pair coisolation. Scatterplot for Comet standard E-value (from Coiso search parameters) of the 1:1 *S. cerevisiae* SILAC proteome mixture faceted by PSM charge state and number of lysines for matching left and right offset Coiso scans. PSM assignment either heavy (blue), light (green), or decoy (purple) based on correctly Coiso scans PSM sequence.

Supplementary Figure 6.4: Coiso E-value calculation



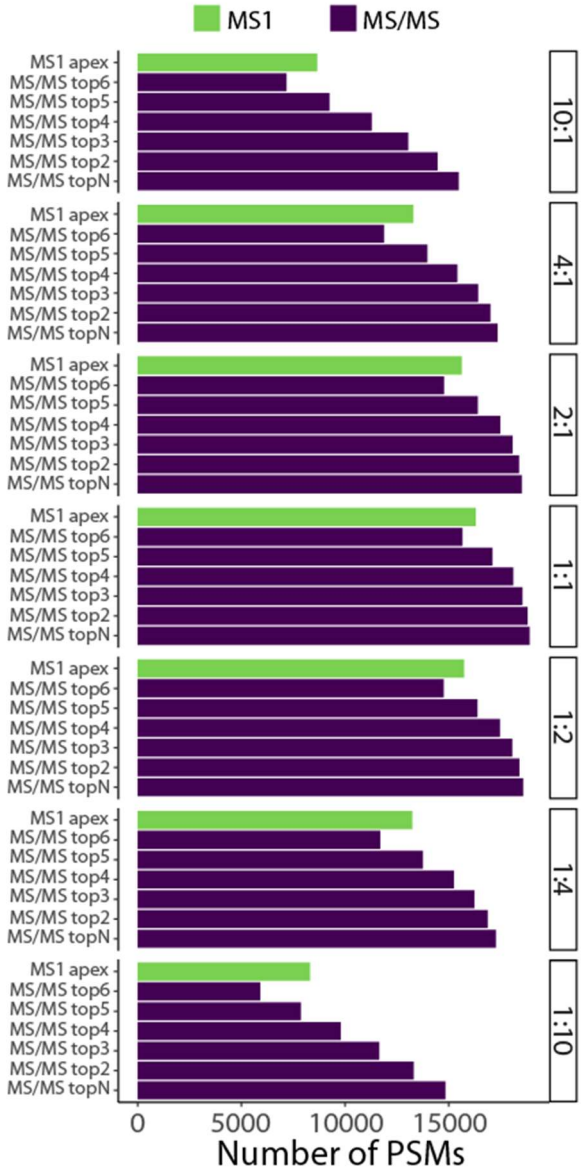
Supplementary Figure 6.4: Coiso E-value calculation with a left and right Coiso scan example. a) Left (top panel) and right (bottom panel) offset wide window Coiso scans for the triggered RKI1 light peptide (zoomed in over fragment ions; dropping dominant non-PSM peaks) with annotated PSM b-ions (purple), light y-ions (green), and heavy y-ions (blue). **b)** Histogram of PSM candidates' Comet Xcorrs when searched with traditional DDA search parameters (only light b- and y-ions) for left offset Coiso scan from (**a**: top panel) with dashed vertical line designating the Xcorr for the PSM sequence identification in (**a**). PSM's Xcorr and E-value for DDA search are noted. **c)** Same as (**b**) for the same PSM candidates when spectra are searched with Coiso SILAC Comet parameters. PSM's Xcorr and E-value for Cosio search are noted. **d)** Same as (**b**) for the same PSM candidates for delta Xcorr (Coiso Xcorr - DDA Xcorr). Coiso Xcorr from (**c**) minus DDA Xcorr from (**b**) for matching candidates. PSM's delta Xcorr is noted. **e)** Scatter plot of the logarithmic transform of the cumulative distribution function (CDF) based on the histogram in (**c**) binned by 0.1 delta Xcorr. Linear regression trendline (solid line) fit on the right tail of delta Xcorr $\log_{10}$(CDF) with the range defined by magenta brackets for E-value calculation. The horizontal dashed line defines the point of intersection between the PSM's delta Xcorr (vertical dashed line and point on the x-axis) and the linear regression trendline which is used to extrapolate the Coiso E-value noted on the plot. **f)** Same as (**b**) for the right offset Coiso scan. **g)** Same as (**c**) for the right offset Coiso scan. **h)** Same as (**d**) for the right offset Coiso scan. **i)** Same as (**e**) for the right offset Coiso scan.

Supplementary Figure 6.5: Coiso E-value enables FDR-controlled quantification



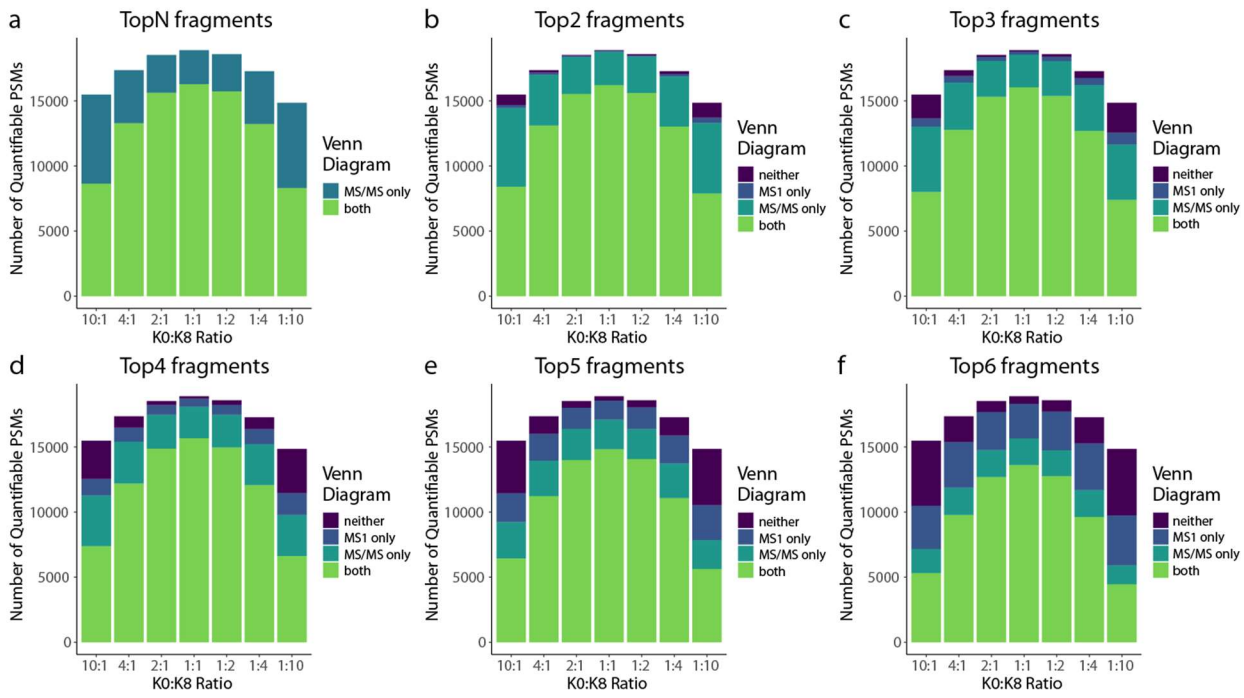
Supplementary Figure 6.5: Coiso E-values enables FDR-controlled quantification. For *S. cerevisiae* 1:1 SILAC ratio proteome mixture, histograms for PSM counts for targets (purple) and decoys (green) for wide window offset scans searched with Coiso Comet parameters. Coiso E-value 1% PSM FDR cut-offs are designated by the solid vertical line.

Supplementary Figure 6.6: Number of quantifiable MS1 vs MS/MS features



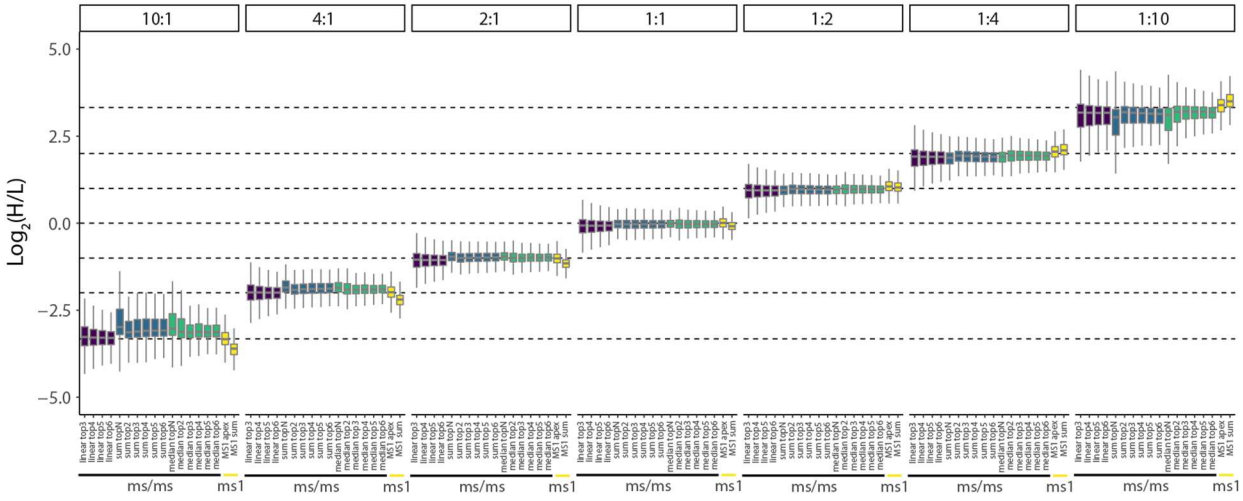
Supplementary Figure 6.6: Coiso SILAC method captures more quantifiable peptide-spectral matches. Bar plot of quantifiable peptide-spectral matches for summed or apex-based MS1 features (green) from Dinosaur or Coiso MS/MS-based quantifications (purple) for top2-top6 and topN (all matched) fragment ions. Plots are faceted by sample with defined SILAC ratio (K0:K8 respectively) of a SILAC-labeled *S. cerevisiae* proteome mixture.

Supplementary Figure 6.7: Venn Diagram of quantifiable MS1 and Coiso MS/MS features



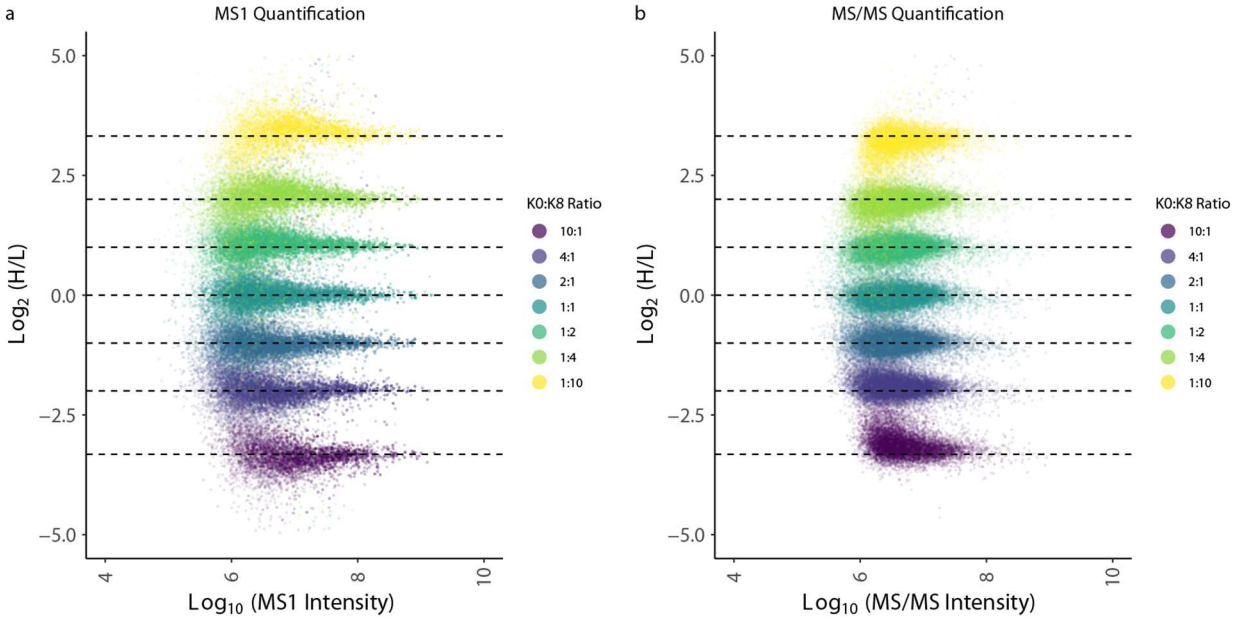
Supplementary Figure 6.7: Coiso SILAC method overlap with MS1 quantifications varies based on number of quantifiable fragment filters. (a-f) Stacked bar plots of the number of quantifiable PSMs colored by Venn diagram designation (both : quantifiable MS1 and MS/MS (green); neither: not quantifiable in MS1 and MS/MS (purple), MS1 only (dark blue), MS/MS only (light blue)). Each panel refers to the number of quantifiable fragments required to ensure an MS/MS quantifiable PSM (**a**:topN; **b**:top2; **c**:top3; **d**:top4; **e**:top5; **f**:top6). Plots are generated based on SILAC-labeled *S. cerevisiae* proteome mixtures with defined SILAC ratios of K0:K8 respectively.

Supplementary Figure 6.8: MS1 and MS/MS SILAC ratio quantification distributions



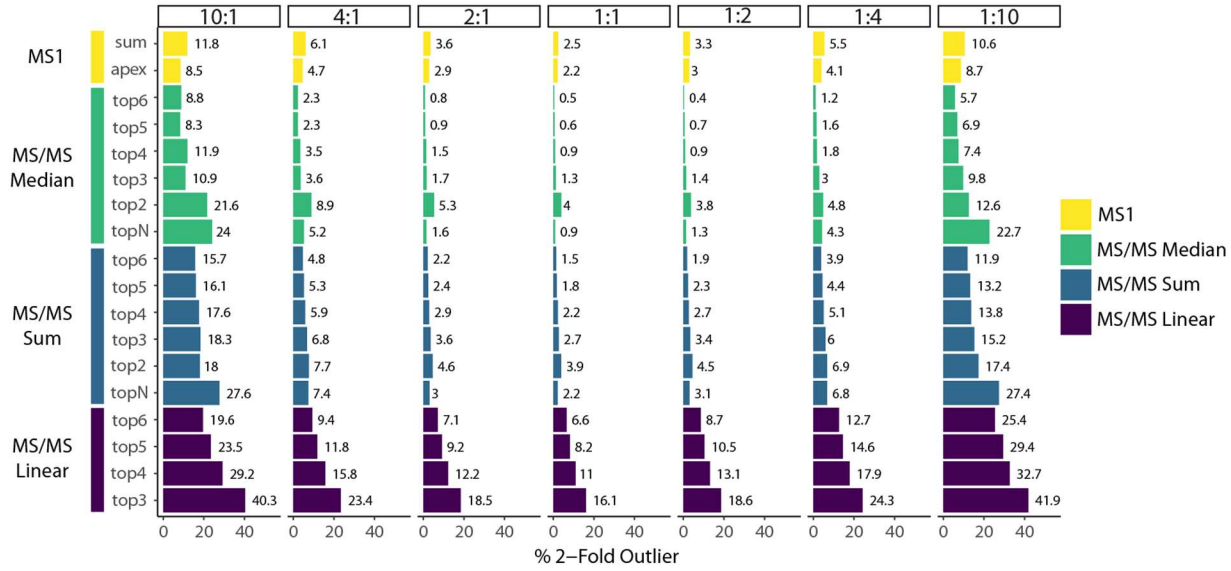
Supplementary Figure 6.8: Coiso SILAC quantification methods and filters vary in precision and accuracy. Box plots for peptide-spectral matches across SILAC *S.cerevisiae* proteome mixtures (K0:K8 respectively). Three MS/MS-based quantification methods were used: median-based (green), sum-based (blue), and linear regression-based (purple) filtered based on the topN or top2-6 quantifiable paired y-ion fragments. MS1-matched SILAC features from Dinosaur (yellow) using either apex or sum-based quantification. Box plots represent the distribution from all quantifiable PSMs from Supplementary Figure 6. In the box plot, the horizontal line represents the median, box designates the IQR, and the whiskers indicate 1.5 x IQR from the box ends.

Supplementary Figure 6.9: MS1 vs MS/MS quantifications across dynamic range



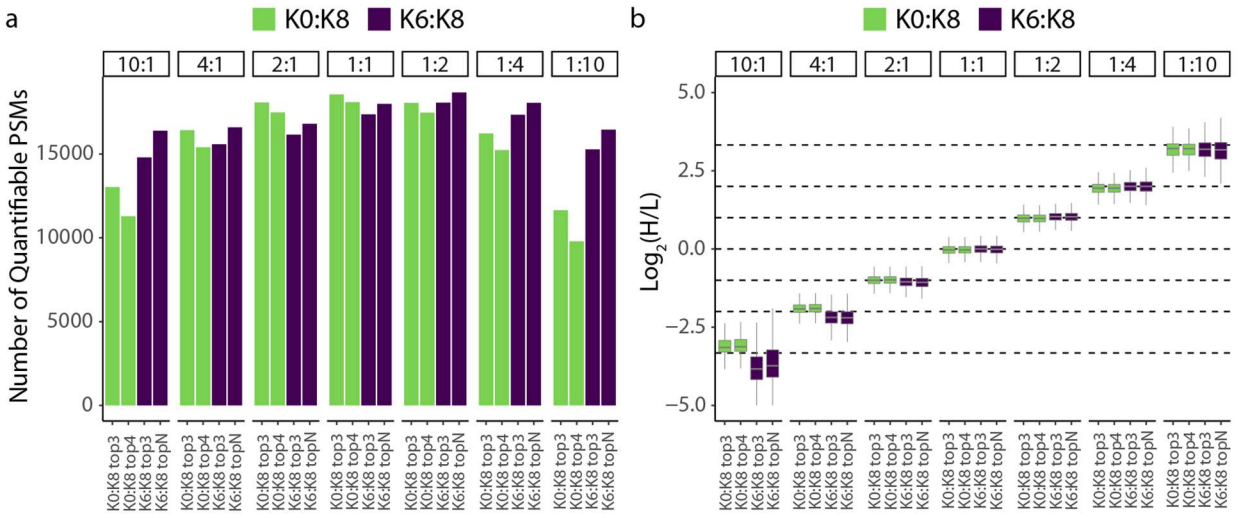
Supplementary Figure 6.9: Coiso SILAC MS/MS and MS1 quantifications map to expected ratios across dynamic range. Scatter plot of all quantifiable PSMs colored by SILAC sample for SILAC S. cerevisiae proteome mixtures (K0:K8 ratio respectively). MS1 intensities are calculated via the sum of heavy and light peptide feature's intensities. MS/MS intensities are derived from the sum of all the peptides' (heavy and light) fragment ion signals from the single MS/MS of the peptide-spectral match.

Supplementary Figure 6.10: Percentage of 2-fold outliers for all MS1 and MS/MS quantification strategies



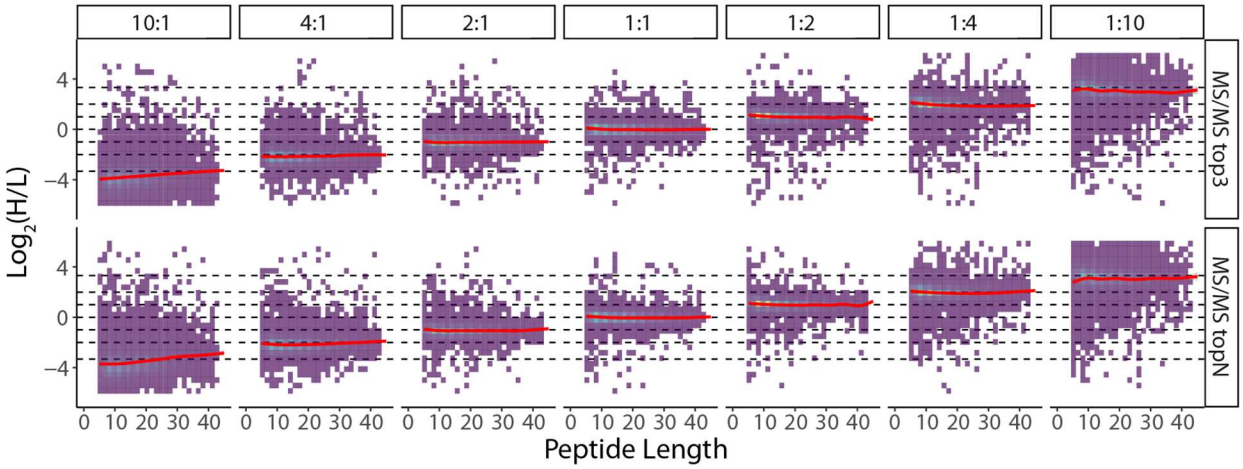
Supplementary Figure 6.10: Coiso SILAC quantification methods and filters vary in 2-fold outliers. Bar plots of percentage peptide-spectral matches that are two-fold outliers from the expected value across SILAC *S.cerevisiae* proteome mixtures (K0:K8 respectively). MS/MS quantification methods (median-based:green, sum-based:blue, linear-based:purple) filtered based on the number of quantifiable paired y-ion fragments and Dinosaur MS1-derived quantifications (yellow) are represented here.

Supplementary Figure 6.11: Comparing K0:K8 and K6:K8 SILAC mixture's quantifiable identifications and quantification precision and accuracy.



Supplementary Figure 6.11: K0:K8 vs. K6:K8 Coiso SILAC MS/MS quantification comparison of quantifiable PSMs and quantification distributions. a) Bar plot of quantifiable peptide-spectral matches for SILAC-labeled *S. cerevisiae* K0:K8 (green) and K6:K8 (purple) proteome mixtures using Coiso MS/MS-based qualifications with respective fragment ion filters. Plots are faceted by sample with defined SILAC ratio (matching K0:K8 and K6:K8 respectively) of a SILAC-labeled *S. cerevisiae* proteome mixture. **b)** Box plots of SILAC ratio for the distribution of all quantifiable PSMs from(**a**). MS/MS-based quantification methods from K0:K8 (green) and K6:K8 (purple) filtered by respective quantifiable paired y-ion fragments. In the box plot, the horizontal line represents the median, box designates the IQR, and the whiskers indicate 1.5 x IQR from the box ends.

Supplementary Figure 6.12: K6:K8 deconvolution algorithm overcorrects light contribution in 10:1 sample when theoretical heavy contribution is smaller



Supplementary Figure 6.12: K6:K8 Coiso SILAC MS/MS quantification changes associated with peptide size for the 10:1 K6:K8 sample. Density-binned scatter plot of Coiso SILAC MS/MS deconvolution quantifications across peptide length for median top3 and topN fragment ion filters. Dashed lines designated expected theoretical SILAC ratios. Red trendline represents the data fit to a general additive model (GAM).

SUPPLEMENTARY DISCUSSION

Inclusion or exclusion of b-ion fragments for Coiso SILAC Comet database search

Unique to coisolation searches, missed cleavages can result in not only paired y-ions, but also paired b-ions. Particularly of note, when performing a Comet internal decoy search, candidate peptides ending in two lysines will generate decoys when the sequence is pseudo reversed[190] with double the *in silico* b-ion fragments possible for calculating an Xcorr, thus providing a bias to missed cleaved decoys and additionally all peptide candidates that have a missed cleavage.

For generating a coisolation search algorithm, we tested generating theoretical SILAC paired spectra for peptide spectral matching that contained paired y-ions and paired b-ions if possible (silac_pair_fragments=1) and a strategy that only considers paired y-ions and only light or heavy b-ions (for light and heavy triggered peptides respectively; silac_pair_fragments=2). To test this hypothesis, we generated a 1-to-1 SILAC (lysine+0 : $^{12}C_6$,$^{14}N_2$-lysine+8) *Saccharomyces cerevisiae* proteome mixture that was digested with Lys-C protease. Using our modified DDA coisolation schema, we triggered three scans on the same precursor: isolated 4 Da offset to the left when heavy (1) and to the right when light (2) with 6.5 m/z wide window scans and our DDA control being an 1.6 m/z narrow window scan (3) targeted the single light or heavy precursor (Figure 6.1b), giving the resultant MS/MS spectra (Figure 6.1c).

The correctly coisolated wide window scan (left or right) was coisolation searched with Comet considering (silac_pair_fragments=1) or excluding (silac_pair_fragments=2) possible paired b-ions. Both coisolation searches were compared to the narrow window scan searched spectra with the traditional non-paired DDA Comet parameters for each targeted precursor. The

proportion of decoy SILAC pairs that did not have exactly one lysine were over-represented in the wide window-coisolation search considering b-ion pairs as compared to the similar proportion observed in wide window-coisolation search excluding b-ion pairs and narrow window-DDA search (Appendix E Supplementary Figure 6.1a). This observation suggests a bias toward missed cleaved candidate peptides (target or decoy) when including paired b-ion fragments. Additionally, the missed cleavage PSM bias is reflected in lower identifications at a PSM 1% FDR based on E-values (Appendix E Supplementary Figure 6.1b). Considering E-values are strongly influenced by the tail of the cumulative distribution function of the candidate Xcorr distribution for each scan and the missed cleavage b-ion bias will likely influence a subset of candidates towards this tail broadening the Xcorr distribution, one could expect that the coisolation search including possible paired b-ions will increase target PSMs E-values (lower better). This notion was supported by the observation that the E-values from the coisolation search excluding paired b-ion outperformed the coisolation search including possible b-ion pairs greater than 2-fold of the time comparing the same spectra with the same matched PSM (Appendix E Supplementary Figure 6.1c). Therefore, we utilized the coisolation search algorithm that excludes the possible paired b-ions for it standardizes the possible theoretical fragments based on peptide length across all possible candidates, generates more PSMs, and improves E-values.

VITA

Ian R. Smith was born in Mission Viejo, CA on December 12, 1991. He grew up in Medford, NJ where he graduated from Shawnee High School in 2010. Following graduation, he pursued a bachelor's degree at Muhlenberg College in Allentown, PA, where he was first introduced to scientific research in the laboratory of Dr. Keri Colabroy. During his time at Muhlenberg, he studied the kinetic mechanism of the enzyme l-DOPA-2,3-dioxygenase in *Streptomyces lincolnensis*. Upon graduation, Ian received a B.S. in Biochemistry and a Minor in Mathematics. For two years following undergraduate, he worked as a research technician in the Children's Hospital of Philadelphia Proteomics Core Facility, where he was introduced to the exciting field of mass spectrometry. In 2016, Ian began pursuing a PhD in the Department of Genome Sciences at the University of Washington. His graduate work was conducted in the laboratory of Dr. Judit Villén, where he developed novel proteomic methods to understand the functional role of proteoforms at scale. Ian defended his thesis on March 11, 2022, and he plans to continue his training in academia as postdoctoral fellow.